Journal of Signal Processing Systems (2022) 94:215–227
https://doi.org/10.1007/s11265-021-01705-8

Multinomial Sampling of Latent Variables for Hierarchical Change-Point Detection

Lorena Romero-Medrano[1] · **Pablo Moreno-Muñoz**[2] · **Antonio Artés-Rodríguez**[1]

Received: 26 February 2021 / Revised: 2 September 2021 / Accepted: 9 September 2021 / Published online: 8 October 2021

Abstract
Bayesian change-point detection, with latent variable models, allows to perform segmentation of high-dimensional time-series with heterogeneous statistical nature. We assume that change-points lie on a lower-dimensional manifold where we aim to infer a discrete representation via subsets of latent variables. For this particular model, full inference is computationally unfeasible and pseudo-observations based on point-estimates of latent variables are used instead. However, if their estimation is not certain enough, change-point detection gets affected. To circumvent this problem, we propose a multinomial sampling methodology that improves the detection rate and reduces the delay while keeping complexity stable and inference analytically tractable. Our experiments show results that outperform the baseline method and we also provide an example oriented to a human behavioral study.

Keywords Bayesian inference · Change-point detection (CPD) · Latent variable models · Multinomial likelihoods

1 Introduction

Change-point detection (CPD) methods aim to identify abrupt transitions in sequences of observations, for both univariate and multivariate cases. Typically, a change-point (CP) is only considered if there is a noticeable difference between the generative parameters of the data before and after the change-point event. Two classical families of approaches can be found in signal processing and machine learning for this task. First, the main focus of early literature has been on *batch* settings [7, 14], where the entire dataset is available for analysis. Second, *online* CPD methods [1] avoid the previous assumption to fulfill two intertwined tasks: i) estimation of the generative parameters of the model as new observations come in and ii) segmentation of the data sequence into non-overlapping partitions based on the parameters obtained.

The identifiability of change-points (CP) is directly related to the discrepancy between the distributions governing each partition. In this context, the Bayesian framework provides a reliable solution to obtain uncertainty measures over both the parameters and the CP locations. The Bayesian online CPD algorithm (BOCPD) introduced in [1] uses this idea to derive a recursive exact inference method. However, when observations become high-dimensional and the number of parameters in the model grows exponentially, there is not enough evidence in the sequential data to obtain reliable estimates of the true generative parameters.

Latent variable models are particularly amenable to overcome the high-dimensionality issue. Under the assumption that change-points lie on a lower-dimensional manifold, one can extend the BOCPD algorithm to accept subsets of surrogate discrete latent variables. Each data point is therefore linked to a single assignment, as it is done in mixture models. The main drawback is that true latent class assignments are never observed but inferred, leading to introduce pseudo-observations [11]. For this purpose, there are two main strategies: i) use the posterior probability vector as a continuous multivariate datum, i.e. as a Dirichlet distributed variable or ii) observe single point-estimates of the discrete latent

Lorena Romero-Medrano and Pablo Moreno-Muñoz have equally contributed.

✉ Lorena Romero-Medrano
lromero@tsc.uc3m.es

✉ Pablo Moreno-Muñoz
pabmo@dtu.dk

Antonio Artés-Rodríguez
antonio@tsc.uc3m.es

[1] Dept. Signal Theory and Communications, Universidad Carlos III de Madrid and Evidence-Based Behavior (eB2), Leganés, Spain

[2] Section for Cognitive Systems, Technical University of Denmark (DTU), Kgs. Lyngby, Denmark

variable. Despite that the first idea was explored in previous works out of the CPD problem [12], it still requires expensive approximate methods due to non-tractability issues. The second idea allows reliable detection instead, particularly when posterior densities over the latent variables are certain enough.

In this paper, we consider the case of having poor inference of point-estimates over the latent variables that lead to catastrophic results on the CPD. Our contribution is to provide a novel extension for the hierarchical CP model that improves the detection rate and reduces delay even under extremely *flat* posterior distributions with high variance. The proposed solution considers latent variable samples as multivariate observations, that we model as multinomial distributed. This keeps the original analytic simplicity of the Bayesian CPD inference as well as the complexity cost remains significantly low. In the experiments, we prove the utility of the new inference method on synthetic data and we also provide insights to be applicable in real-world scenarios, such as change-point detection in a human behavior study.

2 Bayesian Change-Point Detection

Based on the work presented in [1], we assume that a sequence of observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_t$ may be partitioned into *non-overlapping* segments. We consider that each segment or partition ρ with $\rho = \{1, 2, \dots\}$ has an associated generative distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_\rho)$ where the parameters $\boldsymbol{\theta}_\rho$ are unknown and observations are assumed to be independent and identically distributed (i.i.d). The maximum number of partitions is also unknown and bounded by the total number of data points t at each time-step. Therefore, it may increase as new observations $\boldsymbol{x}_t$ come in.

We are concerned with discovering the true generative distributions $p(\boldsymbol{x}|\boldsymbol{\theta}_t)$ and hence, their parameters $\boldsymbol{\theta}_t$ at each time-step. To alleviate the combinatorial problem of estimating parameters based on every partition hypothesis and time-step t, we introduce an auxiliary random variable (r.v.) r_t, also called the *run-length* in the original version of [1]. The discrete variable counts the number of time-steps since the last change-point, that is

$$r_t = \begin{cases} 0, & \text{CP at time } t \\ r_{t-1} + 1, & \text{otherwise.} \end{cases} \tag{1}$$

The main idea behind the *run-length*, r_t, is that it converts the partition hypothesis problem into a Bayesian inference task as well as introduces a relatively simple CP indicator. This strategy augments the model, leading to a double inference mission: i) estimating the posterior distribution over r_t given the data, $p(r_t|\boldsymbol{x}_{1:t})$ and ii) obtaining reliable values of the generative parameters $\boldsymbol{\theta}_t$ conditioned to the partition hypothesis of r_t.

The general method is based on the marginalization of the model parameters $\boldsymbol{\theta}_t$ for the generative distribution given the data at each time-step,

$$p(r_t, \boldsymbol{x}_{1:t}) = \int p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) d\boldsymbol{\theta}_t, \tag{2}$$

and the factorization of the joint density $p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t)$. The discrete nature of the r_t counting r.v. also makes it appropriate for integration, being feasible to obtain the joint distribution in a recursive manner by marginalizing over r_{t-1},

$$p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) = \sum_{r_{t-1}} p(r_t, r_{t-1}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) \tag{3}$$

$$= \sum_{r_{t-1}} p(r_t|r_{t-1}) p(\boldsymbol{x}_t, \boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) p(r_{t-1}, \boldsymbol{x}_{1:t-1}) \tag{4}$$

$$= \sum_{r_{t-1}} p(r_t|r_{t-1}) p(\boldsymbol{x}_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) p(r_{t-1}, \boldsymbol{x}_{1:t-1}), \tag{5}$$

assuming that the prior probability of r_t depends only on the previous value r_{t-1}. Integrating in (3) as proposed in (2) leads to obtain,

$$p(r_t, \boldsymbol{x}_{1:t}) = \sum_{r_{t-1}} p(r_t|r_{t-1}) \Psi_t^{(r)} p(r_{t-1}, \boldsymbol{x}_{1:t-1}), \tag{6}$$

where we have previously defined

$$\Psi_t^{(r)} = \int p(x_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) d\boldsymbol{\theta}_t. \tag{7}$$

When the computation of $\Psi_t^{(r)}$ is not possible, for instance, due to the underlying generative model $p(\boldsymbol{x}|\boldsymbol{\theta}_t)$ is too expressive or complex, other ways for approximate inference must be considered [13, 15].

As we can see in equation (7), the learning process of $\boldsymbol{\theta}_t$ is conditioned to the run-length r_t and hence, the partition hypothesis, carrying out a multiple-thread inference mechanism. Importantly, at each time-step t we can estimate the posterior $p(r_t|\boldsymbol{x}_{1:t})$, obtaining a probability measure of the last CP location, given by the value of r_t. Equivalently, we obtain probability measures for the location of the starting observation of the current partition. For example, having observed $\boldsymbol{x}_{1:5}$ at some time-step $t = 5$, we would compute posterior estimates $\boldsymbol{\theta}_t|r_t, \boldsymbol{x}_{1:t}$, one per each r_t value. As a consequence, given the hypothesis $r_t = 2$, the estimation would be analogous to see $\boldsymbol{\theta}_t|\{\boldsymbol{x}_4, \boldsymbol{x}_5\}$, e.g. a CP is located two observations back, under the previous notation. This example is also depicted in the graphical scheme of Fig. 1.

However, a key inconvenient of this model appears as the size of observations $\boldsymbol{x}_t$ rises, and the Bayesian method works in a potentially high-dimensional setting. In such cases, the complexity of the generative model increases accordingly to the dimensionality of $\boldsymbol{x}_t$. This leads to an extremely large number of parameters $\boldsymbol{\theta}_t$ to estimate. In fact, it makes *almost* impossible to perform CPD in a reliable manner as there is not sufficient statistical evidence given $\boldsymbol{x}_{1:t}$, to update our posterior distribution. In such case, CPs are typically confounded with noise drifts in the underlying parameters.

3 CPD and Latent Variable Models

Latent variable models are a powerful tool in unsupervised learning, with significant connections with Bayesian statistics. This family of approaches typically assumes that there exists a finite low-dimensional representation of the data that characterizes the generative properties of the observed objects. Particularly, it has become a popular solution in probabilistic modelling when the high-dimensionality problems rises. It allows to easily take decisions about the dimensionality of the latent manifold, its nature (i.e. continuous or discrete) and its conditioning to the rest of r.v. implied in the generative model of the data.

In our particular scenario for Bayesian CPD, we may assume that the sequential observations $\boldsymbol{x}_{1:t}$ belong to a lower-dimensional manifold, where the true CPs lie. The generative model is then expressed as

$$p(\boldsymbol{x}_t|\boldsymbol{\theta}_t) = \int p(\boldsymbol{x}_t|z_t)p(z_t|\boldsymbol{\theta}_t)dz_t, \tag{8}$$

where the conditional distribution $p(\boldsymbol{x}_t|z_t)$ is now assumed to be *fixed* and $p(z_t|\boldsymbol{\theta}_t)$ is the new likelihood distribution over the latent variable z_t, that can be either continuous or discrete. With this approach in mind, we can obtain the posterior distribution $p(z_t|x_t)$ at each time-step t, allowing us to work over the latent space instead of performing parameter estimation over the initial observational space. Similar ideas were previously explored in [2, 11] as extensions of the BOCPD method, where only discrete z_t variables were considered.

3.1 Hierarchical CPD

Based on the previous idea, we introduce the hierarchical model presented in [11], where the latent variables at instant t, z_t, are considered as categorical r.v. or *classes* [9], such that $z_t \in \{1, 2, \dots, K\}$. Hence, they work as the assignments of each observation object $\boldsymbol{x}_t$. In the CPD scenario, it can be understood as a segmentation problem of different latent class models. Moreover, we would like to perform CP detection over the true sequence of assignments $z_{1:t}$, but we cannot observe them. Instead, we only know the sequence of posterior distributions $p(z_{1:t}|x_{1:t})$ that has been previously inferred via, for instance, the online expectation-maximization (EM) algorithm [5] or other continual learning strategies [10]. As the chosen approach, we consider to use *maximum a-posteriori* (MAP) estimates of the latent variables z_t as our pseudo-observations. Thus, the point-estimates are obtained from

$$z_t^\star = \arg\max_{z_t} p(z_t|\boldsymbol{x}_t) \quad \forall t. \tag{9}$$

Using the strategy presented in [1], we can obtain $p(r_t, \boldsymbol{z}_{1:t}^\star)$ as in (2). This also translates the problem of CP detection over the sequence of observations $\boldsymbol{x}_{1:t}$ to perform CPD directly over the sequence of MAP estimates $\boldsymbol{z}_{1:t}^\star$. Importantly, to compute the posterior distribution of r_t, we are also based on the marginalization over the parameters of the joint distribution. We can build the same recursiveness in (3) over the r.h.s. term by marginalizing over r_{t-1} inside the integral, that is

$$p(r_t, \boldsymbol{z}_{1:t}^\star) = \sum_{r_{t-1}} p(r_t|r_{t-1})\Psi_t^{(r)} p(r_{t-1}, \boldsymbol{z}_{1:t-1}^\star), \tag{10}$$

where we also consider,

$$\Psi_t^{(r)} = \int p(z_t^\star|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{z}_{1:t-1}^\star)d\boldsymbol{\theta}_t. \tag{11}$$

The term $p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{z}_{1:t-1}^\star)$ that appears whitin $\Psi_t^{(r)}$ is obtained by the multiple posterior updates depicted in the diagram of Fig. 1 and $p(z_t^\star|\boldsymbol{\theta}_t)$ is a categorical likelihood density of the new observed assignments $z_{1:t}^\star$. The details of the developments and definitions can be found in [11].

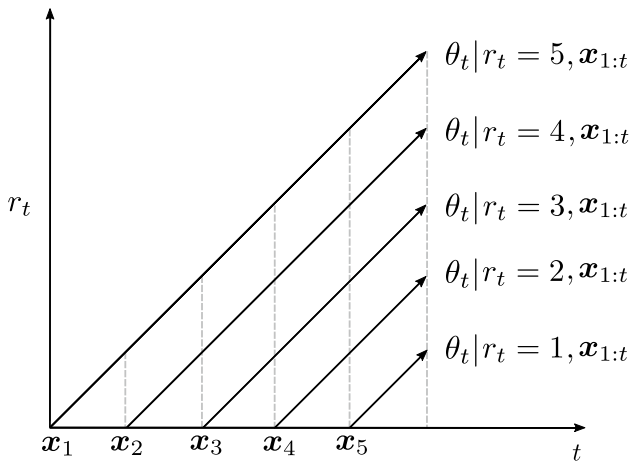

Fig. 1 Illustration of the parallel inference mechanism for the estimation of $\boldsymbol{\theta}_t$ conditioned on the run-length r_t given $\boldsymbol{x}_{1:t}$

3.1.1 The Problem of Flat Posterior Distributions

The hierarchical CP detector solves the problem of poor estimates of the likelihood parameters for high-dimensional observations. However, working with the sequence of MAP point-estimates over the latent space, may also lead to false-alarm or missing detection problems when the inferred posterior distribution $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$ is extremely *flat*, e.g. it is of high variance and there is still uncertainty over the true value of z_t. We have recursively seen this behavior when dealing for instance with a considerable amount of missing data over the observed sequence, or when discrimination of observations is hard. In these particular cases, the MAP estimation may not coincide with the true latent class assignment, introducing extra noise in the CPD with undesired results.

4 Multinomial Sampling

As presented in Section 3, we assume that the observations $\boldsymbol{x}_{1:t}$ belong to a lower-dimensional manifold where the true CPs lie. We consider that the generative model is fully expressed as in (8) and the latent space is of discrete nature, where we are concerned to perform the CP detection. Our goal now is to obtain a better characterization of the underlying posterior distribution, and hence z_t at each time step t, when the density is not well fitted. We have presented one option based on working with the sequence of MAP point-estimates obtained from the posterior $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$. However, if this distribution is extremely flat, we are introducing noise in the CP detection process.

We propose to generate a new type of pseudo observations of the latent variable by drawing S i.i.d. samples of the posterior density, such that $z_t^{(1)}, z_t^{(2)}, \dots, z_t^{(S)} \sim p(z_t|\boldsymbol{x}_t)\,\forall t$, rather than handling a single point-estimate $z_t^\star$. This allows us to work with more information about the unknown true assignment z_t of the observation at each time-step.

The new approach addresses the question of how to deal with multiple subsets of S samples instead of just one variable at each time step. A potential idea would be to introduce Monte-Carlo (MC) sampling methods, but it would lead to draw $S \cdot t$ samples at each time step, becoming unfeasible in the long term. Alternatively, we propose to assume that samples are multinomial distributed, which has the advantage of preserving the prior-conjugacy and is still consistent with the original recursiveness of the BOCPD algorithm [1].

We recall that a multinomial distribution with natural parameters $\boldsymbol{\theta}_t \in \mathcal{S}^K$ and N, measures the probability that each class $k \in \{1, ..., K\}$ has been observed n_k times over N categorical independent trials with same probabilities $\boldsymbol{\theta}_t$. This model allows us to deal with an augmented number of pseudo-observations from z_t at each time t with just the cost of introducing one more parameter in the model: $N = S$, that is, the total number of samples drawn from the posterior.

To treat the samples as multinomial distributed, we have to transform them. Given the sampled vector $\boldsymbol{z}_t^\star = (z_t^{(1)}, z_t^{(2)}, \dots, z_t^{(S)}) \in \{1, ..., K\}^S$, we can define its associated counting vector $\boldsymbol{c}_t \in \mathbb{Z}_+^K$ where $c_t^k := \sum_{s=1}^S \mathbb{I}\{z_t^{(s)} = k\}$ $\forall k$. Each component c_t^k counts the times that class k has been drawn from the S trials, so that we have $\sum_{k=1}^K c_t^k = S$. Thus, at each time t, we can consider the counting vector $\boldsymbol{c}_t$ as an i.i.d. observation of a multinomial distribution with natural parameters $\boldsymbol{\theta}_t \in \mathcal{S}^K$ and $S \in \mathbb{N}$.

With the previous notation in mind, and also assuming the following prior distributions for the parameters

$$\boldsymbol{\theta}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{12}$$

$$\boldsymbol{c}_t \sim \text{Multinomial}(\boldsymbol{\theta}_t, S), \tag{13}$$

where $\boldsymbol{\alpha} \in \mathbb{R}_+^K$, the likelihood function expression of a new observation $\boldsymbol{c}_t$ is given by

$$p(c_t^1, ..., c_t^K|\boldsymbol{\theta}, S) = \frac{S!}{\prod_{k=1}^K c_t^k!} \prod_{k=1}^K \theta_k^{c_t^k}. \tag{14}$$

Using (12) and the conjugacy property, the posterior update rule of parameters has the following closed form $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{c}_t$, which introduces a direct recursion when a new sample is observed.

Algorithm 1 Multinomial CPD

Input: Observe $\boldsymbol{x}_t \rightarrow$ obtain $p(z_t|\boldsymbol{x}_t)$
Sample $z_t^{(1)}, z_t^{(2)}, \dots, z_t^{(S)} \sim p(z_t|\boldsymbol{x}_t)$
Count and build $\boldsymbol{c}_t$
for $r_t = 1$ **to** t **do**
  Evaluate $\Psi_t^{(r)}$ using (20)
  Calculate $p(r_t, \mathbf{c}_{1:t})$
  Obtain $p(\mathbf{c}_{1:t}) = \sum_{r_t} p(r_t, \mathbf{c}_{1:t})$
  Compute $p(r_t|\mathbf{c}_{1:t})$
  Update $\alpha_{t+1}^k = \alpha_t^k + \boldsymbol{c}_t^k \quad \forall k \in \{1, ..., K\}$
end for
Return: $r_t^\star = \arg\max p(r_t|\mathbf{c}_{1:t})$

Notice from the first term of (14) and the definition of $\boldsymbol{c}_t$ that by taking the proposed multinomial model, we are not explicitly working with distributions over the S-dimensional sampled vectors themselves anymore, but over their equivalence classes. Here, two sampled vectors are considered equivalent $\boldsymbol{z}_{S_1}^\star \sim \boldsymbol{z}_{S_2}^\star$ iff their associated counting vectors satisfy $\boldsymbol{c}_{S_1} = \boldsymbol{c}_{S_2}$. That is, if the vector $\boldsymbol{z}_{S_2}^\star$ is a permutation of the vector $\boldsymbol{z}_{S_1}^\star$.

We now aim to obtain the posterior distribution $p(r_t|\boldsymbol{c}_{1:t-1})$ by building up the recursiveness suggested by [1] for the joint distribution. This would allow us to consider a measure of uncertainty of the CP locations and

estimate the generative parameters of the multinomial distribution at each partition. Following the development of (6), we have

$$p(r_t, \boldsymbol{c}_{1:t}) = \int \sum_{r_{t-1}} p(r_t, r_{t-1}, \boldsymbol{c}_{1:t}, \boldsymbol{\theta}_t) d\boldsymbol{\theta}_t = \tag{15}$$

$$= \sum_{r_{t-1}} p(r_t|r_{t-1}) \Psi_t^{(r)} p(r_{t-1}, \boldsymbol{c}_{1:t-1}), \tag{16}$$

where again, we have defined

$$\Psi_t^{(r)} := \int p(\boldsymbol{c}_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)}) d\boldsymbol{\theta}_t. \tag{17}$$

As presented in [1, 11], we have assumed that there exists a term $p(r_t|r_{t-1})$ that only depends on the value of the run-length in the previous instant, r_{t-1}. This term acts like a conditional prior probability that modulates how likely is to detect a new CP, that is, $r_t = 0$ would be more or less likely conditioned to the previous value of run-length hypothesis r_{t-1}.

We also wish to infer the parameter vector $\boldsymbol{\theta}_t^{(r)}$ related to the current run-length r_t and its associated data samples, presented in (17), through the posterior $p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)})$. To carry out the inference method depicted in Fig. 1 we need to find $\Psi_t^{(r)}$, that in fact can be seen as the posterior predictive density of the new pseudo-observation $\boldsymbol{c}_t$ conditioned to the run length r_{t-1} and the previous data in the referred partition,

$$\Psi_t^{(r)} = p(\boldsymbol{c}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)}). \tag{18}$$

The predictive term $p(\boldsymbol{c}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)})$ has not closed form but it is still a function of the statistics of the model. Thus, its computation is straightforward

$$\Psi_t^{(r)} = \frac{\Gamma(S+1)\Gamma(S_\alpha) \prod_{k=1}^K \Gamma(c_t^k + \alpha_{t-1}^k)}{\prod_{k=1}^K \Gamma(c_t^k + 1) \prod_{k=1}^K \Gamma(\alpha_{t-1}^k) \Gamma(S + S_\alpha)},$$

where we have defined $S_\alpha := \sum_{k=1}^K \alpha_{t-1}^k$. Additionally, using both the definition of the binomial coefficient and the properties of the Gamma function, $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$, we transform the previous expression to the following one:

$$\Psi_t^{(r)} = \binom{S + S_\alpha - 1}{S}^{-1} \prod_{k=1}^K \binom{c_t^k + \alpha_{t-1}^k - 1}{c_t^k}. \tag{19}$$

The term S_α grows by S at each time-step, leading to numerical instabilities in the l.h.s term of (19) for high values of t. Therefore, we have considered the following expression that is numerically more stable and it is a result of the manipulations in the terms of (19). Then, it is

$$\Psi_t^{(r)} = \prod_{k=1}^K \prod_{j=0}^{c_t^k - 1} \frac{\alpha_{t-1}^k + j}{S_\alpha + S_c^{(k-1)} + j} \frac{S_c^{(k-1)} + j + 1}{j+1}, \tag{20}$$

with $S_c^{(k-1)} := \sum_{l=1}^{k-1} c_t^l \quad \forall k = 1 \dots K$. This predictive probability is the one that we introduce in (15) for the estimation of r_t, and hence the detection of CPs.

4.1 Change-Point Prior Distribution

The prior distribution of a change-point $p(r_t|r_{t-1})$ needs to be defined to carry out the recursiveness of equation (15) at each time-step t.

In particular, we have considered the prior term to be time independent, as proposed in the original BOCPD version. The conditional distribution takes the form,

$$p(r_t|r_{t-1}) = \begin{cases} H(r_{t-1}+1), & r_t = 0 \\ 1 - H(r_{t-1}+1), & r_t = r_{t-1} + 1 \end{cases} \tag{21}$$

where $H(\tau)$ is the *hazard function* for the geometric distribution with timescale parameter λ, that results in a memoryless process where $H(\tau) = 1/\lambda$ is constant, as detailed in [6]. We consider λ as a model hyperparameter that we fix, but there are also existing works [16] where λ is learned in an online manner. However, this usually leads to extra complexity in the model, and in our case, falls out of the scope of this work.

4.2 Definition of Change-Points

The presented recursive method a
llows us to obtain $p(r_t|\mathbf{c}_{1:t})$ and therefore, the probability that the last CP occurred a number of r_t time-steps ago. This fact is given from the normalization of the joint distribution

$$p(r_t|\mathbf{c}_{1:t}) = \frac{p(r_t, \mathbf{c}_{1:t})}{\sum_{r_t} p(r_t, \mathbf{c}_{1:t})}. \tag{22}$$

MAP estimates-based CPs. Given the posterior density we can define the sequence of likely CP locations $\{r_{1:t}^\star\}$ through the MAP estimates of the posterior distribution of the run-length,

$$r_t^\star = \arg\max p(r_t|\mathbf{c}_{1:t}).$$

This estimation $r_t^\star$ of the CP hypothesis variables is the most used in the literature and the one that we use in our experiments of section 5, since it defines the most likely CPs along the time sequence.

Cumulative probability-based CPs. We propose a new alternative strategy to characterize the variable r_t from $p(r_t|\mathbf{c}_{1:t})$. For fixed t, we consider the probability that a CP occurred in the previous n days so, in this approach, we define the sequence $\{r_{1:t}^\star\}$ by

Table 1 Multinomial CPD vs. Hierarchical CPD metrics. All delay values (×10)

η	$S = 10$ CPD RATE	$S = 50$ CPD RATE	$S = 100$ CPD RATE	HIER. CPD RATE	$S = 10$ DELAY	$S = 50$ DELAY	$S = 100$ DELAY	HIER. DELAY
2.0	-	0.12	0.32	-	∞	5.33 ± 2.30	5.37 ± 1.59	∞
3.0	0.52	0.88	0.84	0.2	5.30 ± 2.09	5.68 ± 3.01	4.20 ± 2.17	10.0 ± 7.87
4.0	0.88	0.96	1.0	0.76	3.57 ± 2.15	3.28 ± 2.53	2.30 ± 0.96	$5.27 + 2.00$
10.0	0.96	1.0	1.0	0.96	2.06 ± 1.77	1.32 ± 0.39	1.31 ± 0.40	3.52 ± 2.00

$$r_t^\star = p(r_t \leq n|\mathbf{c}_{1:t}) = \sum_{i=1}^{n} p(r_t = i|\mathbf{c}_{1:t}).$$

Note that $r_t^\star = 1$ for $t \leq n$, leading to need at least n days of data to use this strategy for CP location definition. Example results are shown in the experiments section, considering different number of days for the cumulative probability computation.

4.3 Computational Cost

Algorithm 1 presents all steps that must be followed to obtain the sequence of CP location estimates $\{r_{1:t}^\star\}$. As the original method of [1], the time-complexity of the general model equation (6) per time-step is linear in the number of data-points so far observed. In the other hand, we see that the complexity related to the number of samples, S, needed for the Multinomial CPD performance, grows linearly with S per time-step. This follows from expression (20) and the fact that $\sum_{k=1}^{K} c_t^k = S$. The low computational cost is one of the most important contributions of this work.

5 Experimental Results

In this section we evaluate the performance of the proposed multinomial sampling extension for hierarchical CPD. First, we study the improvements of the method (named Multinomial CPD) over synthetic data, where we may increase or decrease the quality of inference over the latent variables to prove that detection is still reliable. In the second experiment, we evaluate the method using real-world data of a monitored user from an authorized human behavior study, analyzing how we are able to reduce the delay in the whole detection process. In the third experiment, we study the performance of the method for very large number of latent classes, that is, different values of K, the dimension of the latent variables. In the experiments, we consider that a change point is detected at time-step $t = t'$ if there is an abrupt decrease from $r_{t'-1}^\star$ to $r_{t'}^\star$, which means that the CP occurred at instant $t = t' - r_{t'}^\star$. We set $r_t^\star < r_{t-1}^\star - 20$ as the condition for detection. This can be also adapted if more precision is required.

5.1 Synthetic Data Evaluation

In our first experiment, the Multinomial CPD model has been applied to sequences of synthetic data and the results have been summarized in Fig. 2 and Table 1. Particularly, we want to evaluate the performance of the method for several sampling sizes S, drawn at each time-step and for different levels of *flatness* (uncertainty) of the generative posterior distribution.

We have fixed the number of CPs on the latent sequence to five, that is, six partitions, each one occurring every 100 time steps. Moreover, we have run the algorithm for $T = 600$. In the experiment, the posterior distributions $p(z_t|\boldsymbol{x}_t)$ of the latent variables are also simulated. This guarantees that the posterior densities are as ill-fitted as we want in every example, avoiding the intervention of inference's stochastic conditions in the results. For each partition ρ, we have generated a set of 100 K-dim vectors $\boldsymbol{\theta}_{\rho,t}$ from a Dirichlet distribution with parameters $\boldsymbol{\beta}_\rho$. At the same time, these 6 K-dim vectors $\boldsymbol{\beta}_\rho$ have been sampled from a Uniform distribution in the interval $(0, \eta)$. This parameter η defines the *flatness* of the synthetic posterior distribution, where a lower η implies a flatter generative distribution. The hyperparameter K has been fixed to 20 classes for the whole experiment. In the proposed model, each S-vector has been sampled from a Multinomial$(\boldsymbol{\theta}_{\rho,t}, S)$ with the vector $\boldsymbol{\theta}_{\rho,t}$ previously presented.

The prior probability of the run length r_t is a function of the hyperparameter λ^{-1}, which controls the prior probability of a change: the higher is λ, the less probable is a change. For the Multinomial CPD method (MCPD), we have defined it as a function of the number of samples $\lambda = 10^S$ to make both comparable the terms involved in (6) and also the results in the experiment for different number of samples. The intuition behind this choice is that, for high values of S, we want the prior probability of a change to be almost zero, so that the change-point occurrence is determined from the data. However, more accurate results may be found by tuning the λ parameter at each particular case. For the comparison with the Hierarchical CPD method (HCPD) we considered the same values except for the hyperparameter lambda, that has been fixed to 10^{20} independently of the *flatness* level of the simulated distributions.

In Fig. 2 we compare the MCPD (left column) for different number of samples $S = 10, 50, 100, 150, 200$ with

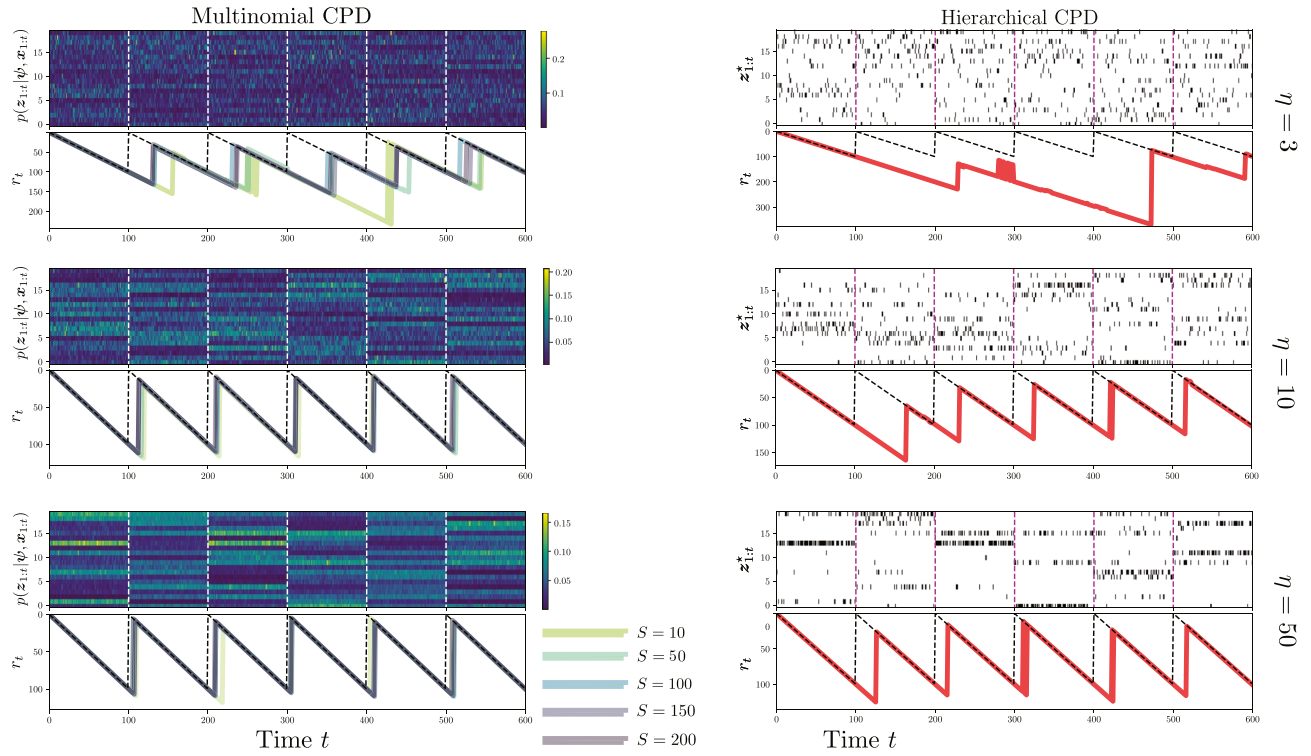

Fig. 2 Comparison between the multinomial CPD, based on sampling from the latent class posterior, and the baseline CPD method. The resulting CPs (bottom figures) are considered as jumps over the MAP estimates (solid lines) of the run-length $r_t \, \forall t$. Dashed lines indicate the true change points. **Left Column:** Each row represents an example with a more or less flat posterior distribution (upper figures) indicated by η. Colors of the r_t lines indicate the number of samples S used. **Right Column:** Results for CPD from different point-estimate pseudo-observations $z^{\star}_{1:t}$ (upper figures)

the HCPD (right column) and different levels of *flatness* $\eta = 3.0, 10.0, 50.0$ (each row). In the upper figures we can see the distributions of the latent variables or the MAP assignments at each time step, respectively. In the bottom figures the MAP estimates of the run-length r_t are jointly shown with dashed lines indicating the true change-points. We have also summarized the results of running the method five times for each pair of values (S, η) in Table 1. There, we also show the total *precision* rate, defined as the ratio of change points detected for each pair, and the mean and standard deviation of the *delay*, defined as the time points between the instant of the detection $t = t'$ and the real instant $t = t' - r^{\star}_{t'}$ in which the CP occurred. For example, if a CP is detected at $t = 150$ and $r^{\star}_{150} = 30$, this means that a change occurred at $t = 120$, and the delay of the detection would be 30 steps. Looking at the results included in Fig. 2, with the MCPD we detect the five change points for every value of η and many of the values of S considered. In the table we confirm that the precision increase as S grows, detecting less change points when the distribution is highly flat for lower values of S and in particular for the HCPD, that is similar to the limit case in which $S = 1$. For $\eta = 2.0$ no change points are detected using the HCPD method. However, with the MCPD, even if the distribution is that flat we are able to find the change points by increasing the number of samples, obtaining a precision of 88% for 50 samples at $\eta = 3.0$ versus the 20% in the HCPD case, or even a 100% of precision already at $\eta = 4.0$ when $S = 100$.

For higher values of η, we can see both in the Fig. 2 and the Table 1 that the performance is good enough for both methods in terms of CPD precision. However, the delay of the detections is always notably lower in the proposed MCPD. In comparison to the HCPD, we can see in the table that the average of the delay in the detections is reduced by more than a half when 100 samples are considered, independently of the *flatness* of the distribution, with just 23.08 time steps of average delay when $\eta = 4.0$ or 13.1 when $\eta = 10.0$.

5.2 Computational Cost Analysis

We analyze the precision rate and average delay in the detection for several sampling sizes and their associated computational cost. In Fig. 4 we show these measurements from $S = 10$ to $S = 200$ evaluated each 10 samples using the MCPD proposed method. The first point of every line ($\approx S = 1$) has been obtained using the HCPD method for every row plot. We consider the same synthetic data generated for subsection 5.1 with $K = 20$ latent classes, $\lambda = 10^S$

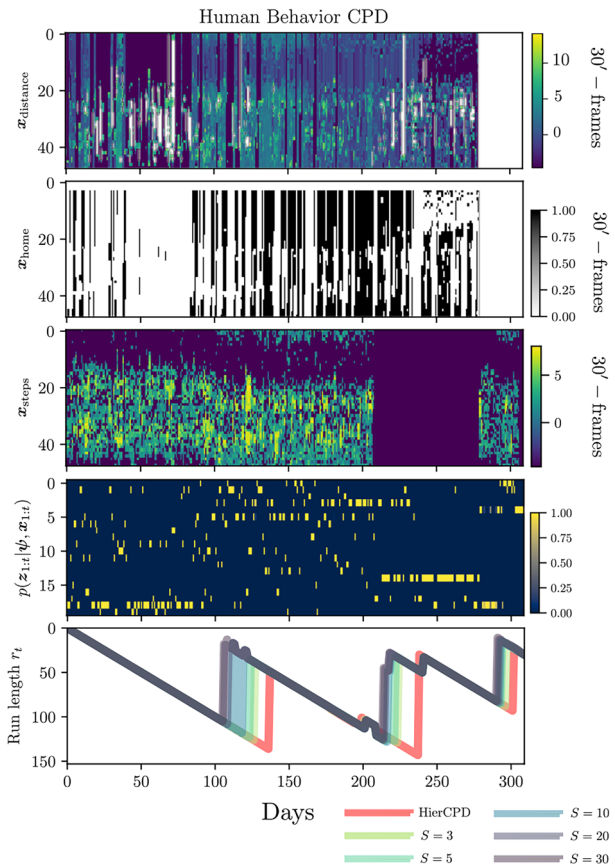

Fig. 3 Human behavior CPD with heterogeneous daily mobility metrics from a user. **Three upper rows.** Respectively, 310 days of distance wandered, presence at home and number of steps every 30 minutes. **Fourth row.** Posterior expectations over the $K = 20$ latent class indicator z_t. **Fifth row.** Hierarchical CPD for several multinomial-sampling cases

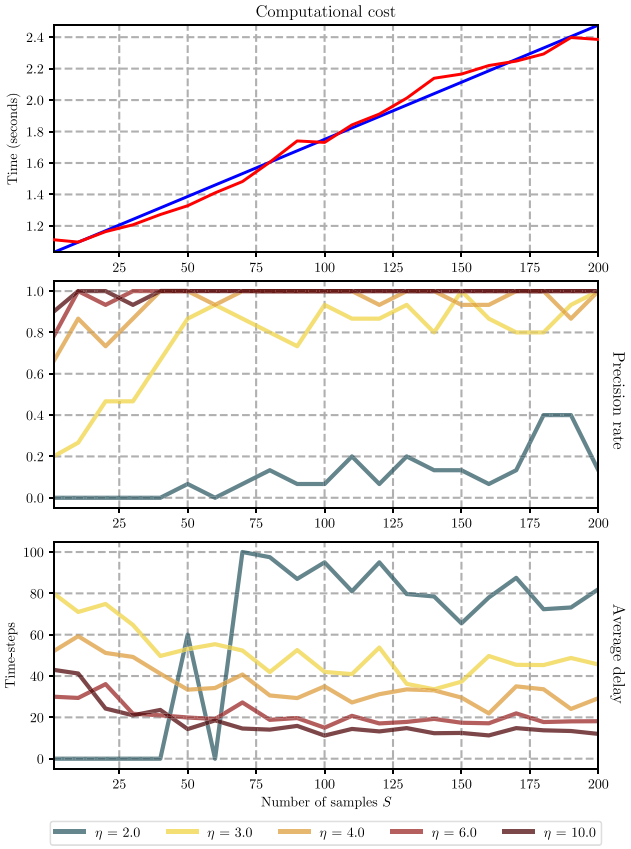

Fig. 4 Computational cost analysis from 10 to 200 samples (MCPD). The first point of the lines ($\approx S = 1$) in every plot corresponds to the HCPD result. **First row.** Computational cost (red line) evolution and adjusted linear regression (blue line). **Second row.** Total precision rate (range: 0.0-1.0) of CP detection for 3 runs and different levels of flatness. **Third row.** Average delay (range: 0-100 time-steps) of CP detection for 3 runs and different levels of flatness

for the MCPD experiments and $\lambda = 10^4$ for the HCPD results. In the first row of Fig. 4, we show the computational cost evolution in seconds (red line) as the number of samples increases, jointly with an adjusted linear regression (blue line). As commented in subsection 4.3, we see that the computational cost increases linearly with the number of samples. We have fixed $\eta = 3.0$ for this experiment, but notice that the computational cost is constant for any level of flatness. The method has been run 3 times and, in the second and third row of the figure, we show the total precision rate (range 0.0-1.0) and average delay (range 0-100 time-steps) in the detection for the same range of sample sizes but different levels of flatness $\eta = 2.0, \eta = 3.0, \eta = 4.0, \eta = 6.0$ and $\eta = 10.0$. Recall that higher η corresponds to lower flatness. In this experiment, the delay has been computed just for the detected CPs, leading to some cases of lower delay values for higher flatness.

We see that both the precision rate and the delay tend to the maximum (minimum for the delay) the method can get for each flatness level as the number of samples increases. The precision rate reaches this value around 50 samples for $\eta \geq 3.0$, that is 1.0 for $\eta \geq 4.0$. The delay presents faster stabilization, due to the computation just over the detected CPs. With respect to the associated computational cost, it increases just 0.3 seconds from the use of the HCPD method (1.1 sec) to the use of MCPD (1.4 sec) for $S = 50$ samples. Looking at these results and taking into account that the maximum accuracy of the method is superseded to the quality of the original data, we can conclude that a size of $S = 50$ samples could give enough good results in most of the cases. Anyway, choosing a higher sampling size like 100 or even 200 would increase the computational cost in just 1.0 second, ensuring to achieve the maximum precision rate and lower delay independently of the data quality level.

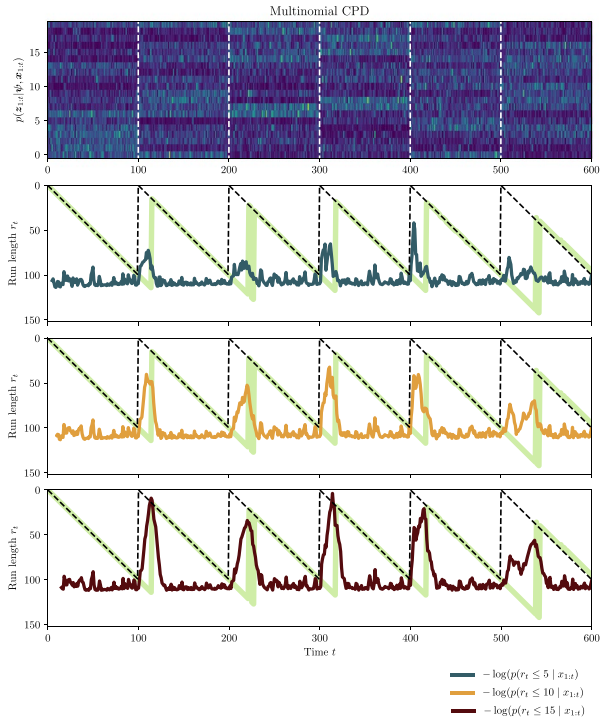

Fig. 5 Cumulative probability measurements. **First row.** Posterior distributions over the $K = 20$ latent class indicator z_t. **Second, third and fourth row.** Results of the MCPD method for $S = 50$ samples: MAP estimates (green line) of the run-length r_t $\forall t$ and negative log of the cumulative probability for 5 (blue line), 10 (orange line) and 15 (brown line) days, respectively. Dashed lines indicate the true change points

5.3 CPD for Large Number of Latent Classes

In this work we have discussed an approach to lead with high-dimensionality data by working over the latent representation of the observations. In this experiment we study the performance of the method in the sense of precision and delay in the detection when the number of latent classes, K, is large, for different levels of *flatness*, η, of the posterior distribution over the latent variables. The number of CPs has been fixed to 5 on the latent sequence, that is, six partitions, each one occurring every 100 time steps. Moreover, we have run the algorithm for $T = 600$ and $S = 100$ for every pair (η, K). The hyperparameter λ has been fixed to 10^S as before. The details in the generation of the data are the same as explained in subsection 5.1. For metrics computation, we compare with the ground truth, but assuming that a CP is considered as not detected if the delay is higher than 100 time-steps. The delay of a not detected CP computes as 100 for the total metric of a trial.

In Tables 2 and 3 we compare the precision and delay respectively for the Multinomial CPD and Hierarchical CPD considering different number of latent classes $K = 10, 20, 40, 50, 100, 200$ (columns) and different levels of *flatness* $\eta = 3.0, 4.0, 5.0, 10.0, 20.0$ (rows). The precision is presented as the total ratio of detected points over the trials and the delay is shown as the mean and standard deviation of the delay of every CP and every trial.

In the results we see that, in general, the higher is K, the lower is the detection rate and the higher is the average delay. In terms of precision, the Multinomial CPD is able to detect more than 92% of the CPs when the flatness is higher than 4.0. Even though, the detection rate for $\eta = 3.0$ is always over 84% when K is lower than 100. For these values of flatness, the HCPD precision is lower than 64% when we have more than 40 classes, as expected.

The delay of the MCPD is less than a half compared to the HCPD results of Table 3 for every pair (η, K). In fact, the average delay in the detection when the flatness is higher enough is not larger than 18.2 for the MCPD even for 100 and 200 classes. In the case of the HCPD, the average delay is always over 80 for $K = 200$ as expected due to the low detection rate and the value considered in the not-detected cases to compute this metric.

5.4 Cumulative Probability Metric

In subsection 4.2 we have proposed an alternative characterization of the variable r_t for CP definition from the posterior $p(r_t \mid \mathbf{c}_{1:t})$: to consider the cumulative probability that a CP happened in the previous n days as a CP location indicator. We present an example of this metric using the synthetic data generated in subsection 5.1 for $\eta = 4.0$.

In Fig. 5 we show the output of the detection over the mentioned dataset for the MCPD method with $S = 50$ samples, $K = 20$ latent classes and $\lambda = 10^{50}$. In the first row we have the posterior distribution of the latent class indicator z_t along time. In second, third and fourth rows we show the sequence of MAP estimates of the run length (green line) jointly with the negative logarithm of the cumulative probability,$-\log p(r_t \leq n \mid \mathbf{c}_{1:t})$, for the 5 (blue line), 10 (orange line) and 15 (brown line) previous days, respectively. The true change points are indicated with dashed lines and the cumulative probability for $t \leq n$ is not plotted because it is 1 by definition and therefore, not informative for our goal.

We see that the cumulative probability based-approach is consistent with the use of MAP estimates for CP definition. However, it could brings to reduce even more the delay in the detection by considering that a CP occurred if there is a sufficient growth of the cumulative probability at a particular time-step. These increases are presented in the images as peaks, that coincide or occur some time-steps before the MAP estimate peaks. Clear examples happen at $t = 400$ and $t = 500$ where the delay could be reduced to 0 for the first CP or to more than a third for the second CP with respect to the MAP strategy.

Table 2 MCPD vs. HCPD precision rate in the detection. The metric has been computed as the ratio of detected CPs over five trials per each pair (η, K) with $S = 100$. A CP is considered not detected if the delay in the detection is higher than 100 time-steps. Best results are underlined

Multinomial CPD						
	$K = 10$	$K = 20$	$K = 40$	$K = 50$	$K = 100$	$K = 200$
η	CPD RATE	CPD RATE	CPD RATE	CPD RATE	CPD RATE	CPD RATE
3.0	0.92	0.84	0.92	0.96	0.84	0.4
4.0	0.92	0.96	1.0	1.0	1.0	1.0
5.0	0.96	1.0	1.0	1.0	1.0	1.0
10.0	0.96	1.0	1.0	1.0	1.0	1.0
20.0	1.0	1.0	1.0	1.0	1.0	1.0
Hierarchical CPD						
	$K = 10$	$K = 20$	$K = 40$	$K = 50$	$K = 100$	$K = 200$
η	CPD RATE	CPD RATE	CPD RATE	CPD RATE	CPD RATE	CPD RATE
3.0	0.24	0.48	0.44	0.28	0.4	0.16
4.0	0.52	0.84	0.64	0.6	0.64	0.08
5.0	0.8	0.92	0.76	0.84	0.6	0.36
10.0	0.84	0.92	0.8	0.84	0.84	0.52
20.0	0.92	1.0	0.92	1.0	0.88	0.72

5.5 Human Behavior Dataset

The data are part of a human behavior study with daily measurements obtained by anonymized monitoring of users using their personal smartphones. The monitoring and preprocessing of data was performed by the Evidence-based Behavior (eB^2) *app* between April, 2019 and March, 2020 [3].

From monitored raw traces of latitude-longitude pairs, we calculate distance in kilometers between sequential locations and its global distance to the user starting point, i.e., his/her home. After splitting all data into 30-minutes frames per 24h, we obtained three multivariate heterogeneous observations per day: i) $\boldsymbol{x}_{\text{distance}} \in \mathbb{R}^{48}$, ii) $\boldsymbol{x}_{\text{home}} \in \{0, 1\}^{48}$, where 1 means staying at home and 0 otherwise, and iii) $\boldsymbol{x}_{\text{steps}} \in \mathbb{R}^{48}$, where the real-positive values where mapped to real-valued using the mapping $\log(1 + y)$. We introduced an *heterogeneous* mixture model given that each daily observation is $\boldsymbol{x}_t = \{\boldsymbol{x}_{\text{distance}}, \boldsymbol{x}_{\text{home}}, \boldsymbol{x}_{\text{steps}}\}$. We refer to *heterogeneous* as a mix of statistical data types. Additionally, we assume that there is a single latent class indicator z_t that indicates the behavioral profile that the user has followed on that particular day. The last step is to obtain the complete sequence of posterior estimates $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$ via the EM algorithm. The learning method of the mixture model can be adapted to the online nature of CPD using [5] or [10] if the number of classes K is unbounded. Results obtained are shown in Fig. 3 for different number of samples drawn by the posterior

Table 3 MCPD vs. HCPD delay in the detection. The metric has been computed as the mean ± standard deviation of the delay in the CPs detection over five trials per each pair (η, K) with $S = 100$. For the not-detected CPs the delay has been considered 100 to obtain a comparable metric. Best results are underlined

Multinomial CPD						
	$K = 10$	$K = 20$	$K = 40$	$K = 50$	$K = 100$	$K = 200$
η	DELAY	DELAY	DELAY	DELAY	DELAY	DELAY
3.0	44.2 ± 29.54	45.64 ± 27.51	57.42 ± 19.93	56.48 ± 19.14	71.04 ± 16.36	98.96 ± 2.96
4.0	28.62 ± 19.82	28.72 ± 11.58	27.36 ± 7.79	35.8 ± 16.09	38.16 ± 9.05	48.2 ± 8.4
5.0	18.76 ± 6.49	21.84 ± 8.61	22.72 ± 8.77	24.52 ± 6.35	27.96 ± 7.92	36.48 ± 6.29
10.0	14.84 ± 11.15	14.6 ± 5.71	13.88 ± 6.48	13.64 ± 4.29	14.96 ± 2.55	18.12 ± 4.19
20.0	10.0 ± 3.59	11.08 ± 4.36	10.2 ± 2.87	10.16 ± 3.25	10.04 ± 1.56	12.84 ± 2.22
Hierarchical CPD						
	$K = 10$	$K = 20$	$K = 40$	$K = 50$	$K = 100$	$K = 200$
η	DELAY	DELAY	DELAY	DELAY	DELAY	DELAY
3.0	91.28 ± 19.07	87.64 ± 19.38	94.72 ± 10.99	90.44 ± 19.28	96.28 ± 8.03	99.76 ± 1.18
4.0	64.48 ± 30.45	71.8 ± 25.31	78.92 ± 24.34	74.85 ± 20.91	85.36 ± 17.56	99.12 ± 3.58
5.0	43.27 ± 25.65	57.07 ± 27.65	61.96 ± 21.31	68.69 ± 28.14	83.64 ± 19.68	93.88 ± 14.16
10.0	31.4 ± 25.32	45.96 ± 27.23	44.35 ± 23.62	52.74 ± 24.07	61.04 ± 19.38	88.8 ± 17.99
20.0	25.01 ± 20.18	26.04 ± 19.18	38.31 ± 21.46	41.8 ± 20.31	58.59 ± 22.49	81.28 ± 20.18

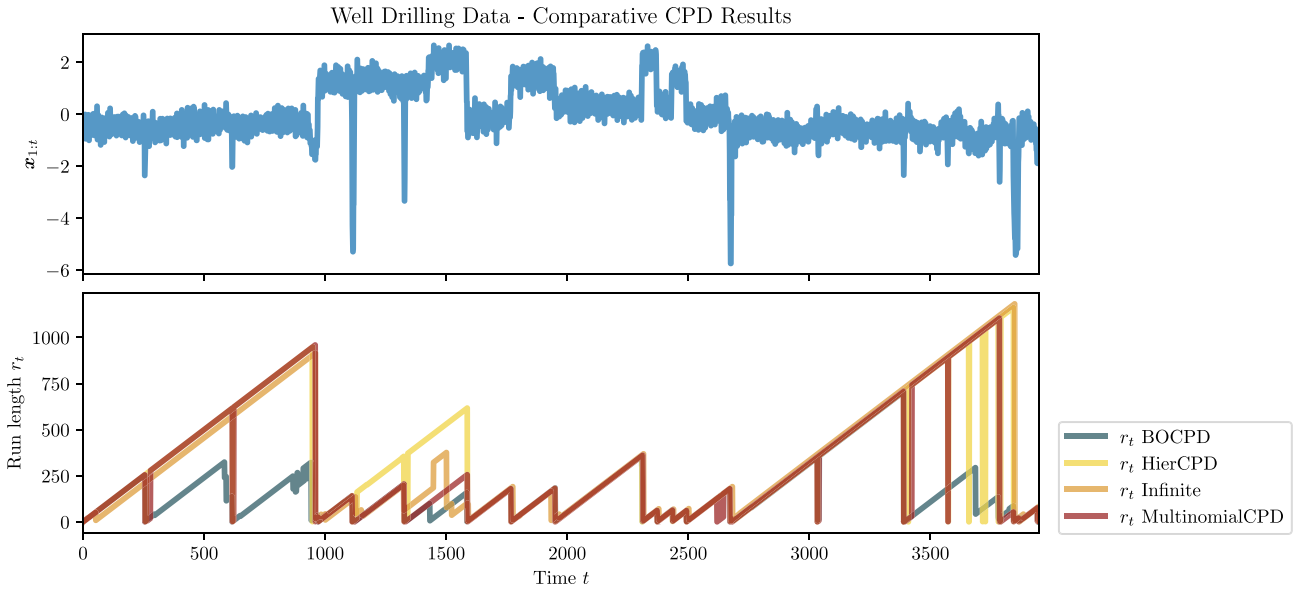

Fig. 6 Comparative CPD results for *four* different methods based on Bayesian inference. **First row:** Univariate data corresponds to the nuclear magnetic response obtained during the drilling of a well. **Second row:** All run-length curves are MAP estimates for r_t given each method. Infinite curve makes reference to the infinite-dimensional version of the Hierarchical CPD model

distribution over the latent variable. We can see that the method finds three change points around day 100, day 230 and day 290, clearly partitioning the time in four behavioral periods between the first and last day of monitoring. These changes have not been contrasted with external information of the user yet, but the results are consistent in terms of number of detections for every value of S considered, and seem to be coherent with the overview of the distributions in the third raw of the figure. Moreover, we can see that increasing the number of samples at each time step, we can reduce the delay in the detection almost 50 days w.r.t. the hierarchical CPD method.

5.6 Comparative CPD Results

In this section, we show comparative CPD results for four different Bayesian approaches and two additional methods from the classical literature. Particularly, we use data from the magnetic response obtained during the drilling of a well. This dataset has been previously used in the context of CPD methods by [1]. We remark that the *true* location of CPs is not provided.

The methods considered for the comparison are i) the Bayesian CPD algorithm [1], ii) Hierarchical CPD [11], iii) the infinite-dimensional method of [10] and iv) the Multinomial-based approach proposed in this work. The detection curves are shown in Fig. 6, where we observe that the better performance comes from the proposed approach using Multinomial samples. The Bayesian CPD algorithm does not include the latent variable hierarchy, which in cases with corrupted, missing or heterogeneous observations is necessary. Its performance is a bit more noisy than ours around the time-step $t = 1000$. The hierarchical CPD method is similar to ours but only uses one single MAP estimate from the underlying distribution $p(z_t|\boldsymbol{x}_t)$ and in this case, it fails when the characterization of CPs requires a higher precision, e.g. around $t = 1200$. This is understandable as the sampling methodology allows us to better characterize the latent variable distribution in that sections of the signal. Moreover, the infinite-dimensional approach of [10] which does not consider a fixed dimensionality for the latent space performs similar to our method. However, it does not detect small transitions in the short-length scale of the time-series, as we do. Examples of these CPs can be found in $t < 1000$ and $t > 3500$ from Fig. 6.

Additionally, we compared our method with other non-Bayesian CPD approaches, which are included in the `ruptures` library for *offline* CPD methods.[1] In particular, we considered optimal partitioning [8] and the binary segmentation [14] method. The detection results were similar for both methods, and they only marked CPs in the t range [1000, 2500]. This means that the short-length transitions at the beginning and end of the signal were not considered as the other Bayesian methods do. In this context, it is worthy to mention the work of [4], where a thorough evaluation of CPD methods is performed w.r.t. a large benchmark of

[1] Python library is publicly available at https://pypi.org/project/ruptures/.

datasets. The final results shed light on the advantage of considering Bayesian CPD methods in certain cases, similarly as we see in our experiments.

6 Conclusion

In this paper, we have presented a novel methodology for improving the Bayesian CPD algorithm of [1] with latent variable models. Under the assumption that CPs lie in a lower-dimensional manifold, inference is carried out with pseudo-observations based on posterior point-estimates of the latent variables given the data. We introduce a multinomial-sampling method that improves the detection rate and reduces the delay when we treat with high-dimensional sequences of observations. The analytical tractability in the inference is maintained as well as a low computational cost. The experimental results show significant improvements in the CPD as posterior estimates become less certain. Interestingly, even under a good inference performance, the multinomial sampling method reduces the delay of detection, what in practice is a key point for its application to real-world problems. We illustrate an example on a human behavioral study, that detects changes in the circadian patterns of a user. In future work, this could be integrated with other CPD methods that consider the dimensionality of the latent variables unbounded [10].

Funding This work was supported by the Ministerio de Ciencia, Innovación y Universidades under grant TEC2017-92552-EXP (aMBITION), by the Ministerio de Ciencia, Innovación y Universidades, jointly with the European Commission (ERDF), under grants TEC2017-86921-C2-2-R (CAIMAN) and RTI2018-099655-B-I00 (CLARA), and by the Comunidad de Madrid under grant Y2018/TCS-4705 (PRACTICO-CM). The work of PMM has been supported by FPI grant BES-2016-077626 and ERC funding under the EU's Horizon 2020 research and innovation programme (grant agreement nº 757360). LRM has been supported by grant IND2018/TIC-9649 from the Comunidad de Madrid.

References

1. Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection. preprint arXiv:0710.3742.
2. Agudelo-España, D., Gomez-Gonzalez, S., Bauer, S., Schölkopf, B., & Peters, J. (2020). Bayesian online prediction of change points. UAI.
3. Berrouiguet, S., Ramírez, D., Barrigón, M. L., Moreno-Muñoz, P., Carmona, R., Baca-García, E., & Artés-Rodríguez, A. (2018). Combining Continuous Smartphone Native Sensors Data Capture and Unsupervised Data Mining Techniques to Detect Changes in Behavior: A Case Series of the Evidence-Based Behavior (eB2) Study. JMIR MHealth and UHealth.
4. van den Burg, G. J., Williams, C. K. (2020). An evaluation of change point detection algorithms. arXiv preprint arXiv:2003.06222.
5. Cappé, O., & Moulines, E. (2009). Online expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(3), 593–613.
6. Evans, M., Hastings, N., & Peacock, B. (2020). Statistical Distributions. WileyInterscience.
7. Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing, 16*(2), 203–213.
8. Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., et al. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters, 12*(2), 105–108.
9. McLachlan, G. J., & Basford, K. E. (1988). Mixture models: Inference and applications to clustering, vol. 38. M. Dekker New York.
10. Moreno-Muñoz, P., Ramírez, D., & Artés-Rodríguez, A. (2020). Continual learning for infinite hierarchical change-point detection. ICASSP.
11. Moreno-Muñoz, P., Ramírez, D., & Artés-Rodríguez, A. (2021). Change-point detection in hierarchical circadian models. *Pattern Recognition, 113*.
12. Nazábal, A., Garcia-Moreno, P., Artes-Rodriguez, A., & Ghahramani, Z. (2015). Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics, 20*(5), 1342–1351.
13. Saatci, Y., Turner, R., Rasmussen, C. E. (2010). Gaussian process change point models. ICML pp. 927–934.
14. Scott, A. J., & Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. Biometrics pp. 507–512.
15. Turner, R., Bottone, S., & Stanek, C. (2013). Online variational approximations to non-exponential family change point models: With application to radar tracking. NIPS.
16. Wilson, R.C., Nassar, M.R., Gold, J.I.: Bayesian online learning of the hazard rate in change-point problems. Neural computation 22(9), 2452–2476 (2010).

Lorena Romero-Medrano obtained her Degree in Mathematics from Universidad de Zaragoza in 2015 and a Master in Mathematical Modeling (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 modeling of liver hemodynamics and continued with this project as research visitor at Universidad Complutense de Madrid. She is currently PhD 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 modeling, machine learning, probabilistic methods, and its applications to human behavior and biomedical sciences.

Pablo Moreno-Muñoz obtained his B.Sc. and M.Sc in Telecommunication Engineering from Universidad Carlos III de Madrid, Spain in 2014 and 2016, respectively. In 2015, he held a 6-month traineeship at European Space Agency for the investigation of probabilistic methods applied to distance calculations in astronomy. He is currently PhD student at the Dept. of Signal Theory and Communications also at the Universidad Carlos III de Madrid. During the last years, he has been research visitor at the University of Sheffield, UK and the Max Planck Institute for Intelligent Systems in Tübingen, Germany. His research interests include probabilistic machine learning methods for heterogeneous data, Gaussian processes, change-point detection, continual Bayesian inference, and its application to human behaviour modelling in medicine.

Antonio Artés-Rodríguez was born in Alhama de Almería, Spain, in 1963. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees, both from the Universidad Politécnica 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 Politécnica 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 90 journal papers and more than 130 international conference papers. His research interests include signal processing, machine learning, and information theory methods, and its application to health and sensor networks.