CONTINUAL LEARNING FOR INFINITE HIERARCHICAL CHANGE-POINT DETECTION

Pablo Moreno-Muñoz, David Ramírez and Antonio Artés-Rodríguez

Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
Gregorio Marañón Health Research Institute, Spain

ABSTRACT

Change-point detection (CPD) aims to locate abrupt transitions in the generative model of a sequence of observations. When Bayesian methods are considered, the standard practice is to infer the posterior distribution of the change-point locations. However, for complex models (high-dimensional or heterogeneous), it is not possible to perform reliable detection. To circumvent this problem, we propose to use a hierarchical model, which yields observations that belong to a lower-dimensional manifold. Concretely, we consider a latent-class model with an unbounded number of categories, which is based on the chinese-restaurant process (CRP). For this model we derive a continual learning mechanism that is based on the sequential construction of the CRP and the expectation-maximization (EM) algorithm with a stochastic maximization step. Our numerical results show that the proposed method is able to recursively infer the number of underlying latent classes and perform CPD in a reliable manner.

Index Terms— Bayesian inference, continual learning, change-point detection (CPD), chinese-restaurant process (CRP), expectation-maximization (EM) algorithm.

1. INTRODUCTION

Change-point detection (CPD), which consists of locating abrupt transitions in the generative model of the observations, is a problem with a plethora of applications. For instance, CPD is widely used in finance [1, 2], the analysis of social networks [3, 4], or cognitive radio [5,6]. The main focus of CPD methods has been traditionally on batch settings, where the entire sequence of observations is available and has to be segmented. However, CPD is most useful in online scenarios, where change points must be detected as new incoming samples are observed. Online CPD methods have two intertwined tasks to solve: i) segmentation of sequential data into partitions (or segments) and ii) estimation of the generative model parameters for the given partitions.

Since each partition has a different generative distribution, the identifiability of change points is related to the difference between such distributions. In this context, Bayesian inference is useful for inferring the distributions given a prior distribution in a reliable manner. The Bayesian online change-point detection (BOCPD) approach [7] used this idea for recursively performing density estimation, which yields a more robust detection process as the propagation of uncertainty is considered. However, it can be observed that, for complex likelihood models, which have a number of parameters much higher than the number of observations between two consecutive change points, reliable CPD becomes unfeasible. This can be the case of, although is not restricted to, high-dimensional and/or heterogeneous observations (mixture of continuous and discrete variables), which usually have a prohibitive number of parameters.

To address the aforementioned issue, in [8] we presented a hierarchical probabilistic model based on latent classes, i.e., a mixture model. The CPD problem can be carried out directly on the lower-dimensional manifold, where the discrete latent variables lie. Hence, this method requires less evidence than the observational counterpart since the number of parameters is reduced, which yields faster and more reliable detections. However, [8] requires that the number of classes is fixed *a priori*.

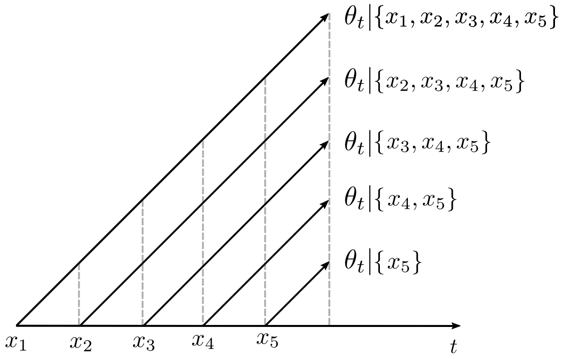
The main contribution of this paper is to introduce a novel approach, based on continual learning [9–11], to recursively infer the underlying sequence of latent classes, its distributions, and the change points. The key idea of the proposed model is to allow for an unbounded order on the latent model, that is, the number of classes is not fixed and could even become infinite. In particular, we use the Chinese-restaurant process (CRP) [12], which is a well-known Bayesian non-parametrics method, to model the latent variables with an unbounded number of classes. That is, the CRP may increase the number of classes as new observations come in. Moreover, as with any mixture model, the expectation-maximization (EM) algorithm [13] is used, but in this work the maximization step (M-step) is substituted by a stochastic M-step [14]. Finally, the experimental results on real data show how both the latent-class inference process and the change-point detection perform reliably.

2. BAYESIAN ONLINE CHANGE-POINT DETECTION

We start by considering a time series $\boldsymbol{x}_{1:t} = \{x_1, x_2, \dots, x_t\}$, which is divided into non-overlapping partitions, denoted by $\rho_i, i = 1, 2, \dots$ Each partition is separated from its neighbors by change points (CP). Based on [7], we assume that the data within each partition ρ_i is independent and identically distributed (i.i.d.) according to some generative probability distribution $p(x_t|\boldsymbol{\theta}_{\rho_i})$, where the parameter vector, $\boldsymbol{\theta}_{\rho_i}$, is unknown. Under this assumption, change points are determined by changes in the parameters:

$$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_{\rho_1}, & t < \mathrm{CP}_1, \\ \boldsymbol{\theta}_{\rho_2}, & \mathrm{CP}_1 \leq t \leq \mathrm{CP}_2, \\ \boldsymbol{\theta}_{\rho_3}, & \mathrm{CP}_2 \leq t \leq \mathrm{CP}_3, \\ & \vdots \end{cases} \tag{1}$$

This 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 P. Moreno-Muñoz has also been supported by FPI grant BES-2016-077626.



$\theta_t|\{x_1, x_2, x_3, x_4, x_5\}$
$\theta_t|\{x_2, x_3, x_4, x_5\}$
$\theta_t|\{x_3, x_4, x_5\}$
$\theta_t|\{x_4, x_5\}$
$\theta_t|\{x_5\}$
x_1
x_2
x_3
x_4
x_5
t

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

The main idea in [7] is the *run-length*, r_t, which is defined as a discrete random variable that counts the number of time-steps since the last CP, that is,

$$r_t = \begin{cases} 0, & \text{CP at time } t \\ r_t + 1, & \text{otherwise,} \end{cases} \tag{2}$$

and may be seen as a proxy for change points. The objective of the BOCPD technique is to compute the posterior distribution $p(r_t|\boldsymbol{x}_{1:t})$ recursively, from which we will identify a CP if the probability mass accumulates near $r_t = 0$.

The posterior distribution $p(r_t|\boldsymbol{x}_{1:t})$ is obtained by marginalizing the joint distribution $p(r_t, \boldsymbol{x}_{1:t})$ over all the r_t values seen so far, which, in turn, is computed by marginalizing the model parameters, $\boldsymbol{\theta}_t$. The learning of $\boldsymbol{\theta}_t$ given the partition, required for the computation of $p(r_t, \boldsymbol{x}_{1:t})$, is carried out using a multiple thread inference mechanism induced by the run-length. For instance, to learn $\boldsymbol{\theta}_3$ given $r_3 = 2$, only the observations $\{x_2, x_3\}$ are required. This parallel inference scheme is depicted in Figure 1, where we illustrate the aforementioned example using the notation $\boldsymbol{\theta}_3|\{x_2, x_3\}$.

The inference of $p(r_t|\boldsymbol{x}_{1:t})$ in [7] may become unfeasible when the complexity of the generative model increases, for instance, for high-dimensional and/or heterogenous observations. That is, if the likelihood $p(x_t|\boldsymbol{\theta}_{\rho_i})$ for the partition ρ_i depends on an extremely large number of parameters, it would not be possible to obtain sufficient statistical evidence to detect change points. This problem may yield the BOCPD method unusable in some problems.

3. CPD ON HIERARCHICAL MODELS

The aforementioned problem of the BOCPD for complex generative models can be overcome by introducing hierarchical models. We propose to use latent classes to obtain such hierarchical model. These latent classes, z_t, yield observations, x_t, that belong to a lower-dimensional manifold, and allow us to write the generative distribution of x_t as

$$p(x_t|\boldsymbol{\theta}_t) = \sum_{z_t=1}^{K} p(x_t|z_t)p(z_t|\boldsymbol{\theta}_t),$$

where z_t is a categorical random variable, with K being the maximum number of classes or categories, and $\boldsymbol{\theta}_t$ is the vector of parameters, i.e., the probability of each class. This form of latent-class model can be seen as a mixture model.

Even assuming a hierarchical model, we are still interested in $p(r_t|\boldsymbol{x}_{1:t})$, which would require the marginalization over $\boldsymbol{z}_{1:t}$ as follows

$$p(r_t, \boldsymbol{x}_{1:t}) = \sum_{\boldsymbol{z}_{1:t}} p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}). \tag{3}$$

However, for large values of t and K, the marginalization in (3) is computationally unfeasible due to the combinatorial sums. In [8], to avoid the marginalization, we assumed that we observe $\boldsymbol{z}_{1:t}$, instead of marginalizing them, by directly plugging in the values of the *maximum a posteriori* (MAP) estimates, which are given by

$$z_t^\star = \arg\max_{z_t} p(z_t|x_t). \tag{4}$$

Now, using the MAP estimates as observations and assuming that the joint distribution on the right hand side (r.h.s.) of (3) factorizes as

$$p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}^\star) = p(\boldsymbol{x}_{1:t}|\boldsymbol{z}_{1:t}^\star)p(r_t, \boldsymbol{z}_{1:t}^\star),$$

with

$$p(r_t, \boldsymbol{z}_{1:t}^\star) = \int p(r_t, \boldsymbol{z}_{1:t}^\star, \boldsymbol{\theta}_t)d\boldsymbol{\theta}_t,$$

we are effectively considering that the change points occurred on the sequence of latent classes. Using the extended recursion of [8], which is given by

$$p(r_t, \boldsymbol{z}_{1:t}^\star) = \sum_{r_{t-1}} p(r_t|r_{t-1})\Psi_t^{(r)}p(r_{t-1}, \boldsymbol{z}_{1:t-1}^\star), \tag{5}$$

where $p(r_t|r_{t-1})$ is the conditional prior and

$$\begin{aligned}\Psi_t^{(r)} &= p(z_t^\star|r_{t-1}, \boldsymbol{z}_{1:t-1}^\star) \\ &= \int p(z_t^\star|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{z}_{1:t-1}^\star)d\boldsymbol{\theta}_t, \end{aligned} \tag{6}$$

is the predictive distribution of the present latent variable conditioned on previous data and the run-length, we have all the ingredients to compute

$$p(r_t|\boldsymbol{z}_{1:t}^\star) = \frac{p(r_t, \boldsymbol{z}_{1:t}^\star)}{\sum_{r_t} p(r_t, \boldsymbol{z}_{1:t}^\star)}, \tag{7}$$

which determines the location of the change points.

3.1. Infinite-dimensional Hierarchical BOCPD

The problem of the hierarchical BOCPD algorithm presented above is that the number of classes, K, must be known and fixed *a priori*. That is, K is not allowed to vary over time, which can be a stringent condition in some scenarios. In this section, we consider the more interesting case that K is unknown and can be time-varying, i.e., new classes may appear as $t \to \infty$. Then, we cannot select the order of the latent-class model in advance. A naive idea would be to fix an upper bound on K and proceed as in the previous section. However, this upper bound could not be available and, even if it is, the performance can be poor, as we will see in Section 5. In the following, we will present a method for unbounded and time-varying K, that is, K is incremented when an unseen type of observations appears, which translates into a hierarchical BOCPD with unbounded K.

Using an unbounded number of classes results in the following problem when integrating over $\boldsymbol{\theta}_t$ to compute $\Psi_t^{(r)}$. Assuming a Dirichlet distribution for $\boldsymbol{\theta}_t$, which is the conjugate prior for categorical distributions and therefore yields a tractable integral in (6), the evidence $p(z_t) \to 0$ as K grows. To overcome this issue, we can consider an exchangeable distribution of the form

$p([\boldsymbol{z}_t]) = \sum_{\boldsymbol{z}_{1:t}\in[\boldsymbol{z}_{1:t}]} p(\boldsymbol{z}_{1:t})$, where $[\boldsymbol{z}_{1:t}]$ is a given division of classes, which is independent of the temporal assignments, i.e., $\boldsymbol{z}_{1:3} = \{1, 2, 2\}$ corresponds to the same division of objects as $\boldsymbol{z}_{1:3} = \{2, 1, 1\}$. This is often known as the *exchangeability* property [12,15] and is a safe assumption in our setup as we are interested in changes in the probabilities of $\mathbf{z}_t$, not in the particular sequences $\boldsymbol{z}_{1:t}$.

The latent-class model with an unbounded dimension can be addressed using the CRP [12], which is a Bayesian non-parametrics method [16]. The CRP is based on a metaphor where clients (observations x_t) are assigned to different tables (latent classes z_t) in a sequential manner. The assignment of classes to objects in the CRP is determined by the predictive posterior distribution, which is given by

$$p(z_t = k|z_1, \ldots, z_{t-1}) = \begin{cases} \frac{m_{k,t-1}}{t-1+\alpha}, & k \leq K_{t-1}, \\ \frac{\alpha}{t-1+\alpha}, & k = K_{t-1}+1, \end{cases} \tag{8}$$

where $m_{k,t-1}$ counts the number of assignments to class k up to time $t-1$, K_{t-1} is the number of classes associated with $m_{k,t-1} > 1$ and α is a hyperparameter, which corresponds to the natural parameter of a symmetric Dirichlet prior distribution, and controls how likely is the appearance of a new class.

Exploiting the aforementioned CRP construction, the computation of $\Psi_t^{(r)}$ in (5) is straightforward, and is given by

$$\Psi_t^{(r)} = p(z_t^\star = k|r_{t-1}, \boldsymbol{z}^\star_{1:t-1}), \tag{9}$$

where we now count the number of MAP estimates, $z_t^\star$, equal to k up to time $t-1$. Notice that this expression is analogous to (8) for a given run-length, i.e., for each parallel thread in Fig. 1. Then, we may proceed to compute the posterior $p(r_t|\boldsymbol{z}^\star_{1:t})$.

One final comment is in order. So far, we have derived a tractable recursive way to introduce latent-class models into Bayesian CPD methods with an unbounded number of classes. However, nothing has been said on how to compute the MAP estimates in a continual learning fashion, which are required in (7). This task is explored in Section 4.

4. CONTINUAL LEARNING OF THE CRP

In this section, we compute the MAP estimates of z_t in an online and recursive fashion. This task also involves the estimation of $\{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}$, which are the parameters of the mapping between observations and latent variables, that is, $p(x_t|z_t = k) = p(x_t|z_t = k, \boldsymbol{\varphi}_k)$. Here, the number of classes K_t increases if when sampling from the CRP predictive distribution the result is $K_{t-1}+1$. That is, at the beginning of each iteration we create a new class with an emission probability given by (8), which is only kept if the MAP estimate is $z_t^\star = K_{t-1}+1$.

Mixture models do not usually have closed-form solutions for the estimates of the parameters and the class assignments. Therefore, it is necessary to resort to the expectation-maximization (EM) algorithm [13], for which we need the log-likelihood of the complete data, which is given by

$$\mathcal{L}_{\boldsymbol{\varphi}}(\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}) = \log p(\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}|\{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}) = \log p(\boldsymbol{z}_{1:t}) + \sum_{\tau=1}^{t} \log p(x_\tau|z_\tau, \{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}), \tag{10}$$

where the prior distribution $p(\boldsymbol{z}_{1:t})$ factorizes as

$$p(\boldsymbol{z}_{1:t}) = p(z_t|\boldsymbol{z}_{1:t-1})p(z_{t-1}|\boldsymbol{z}_{1:t-2}) \cdots p(z_1).$$

Algorithm 1 Infinite-dimensional Hierarchical BOCPD

```
Input: Observe $x_t$ and initialize $\hat{\boldsymbol{\varphi}}_{K_{t-1}}$.
Sample $z_t \sim p(z_t|\boldsymbol{z}^\star_{1:t-1})$
if $z_t = K_{t-1}+1$ then
    Initialize $\hat{\boldsymbol{\varphi}}_{K_{t-1}+1}$
end if
Compute $p(z_t = k|\boldsymbol{z}^\star_{1:t-1}), \forall k \leq K_{t-1}+1$
Compute $\mathbb{E}[\mathbb{I}\{z_t = k\}|\boldsymbol{z}^\star_{1:t-1}, x_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}], \forall k \leq K_{t-1}+1$
Update parameters $\{\hat{\boldsymbol{\varphi}}_k\}_{k=1}^{K_{t-1}+1}$ using (11)
Calculate $z_t^\star = \arg\max(p(z_t|\boldsymbol{z}^\star_{1:t-1}, x_t, \{\hat{\boldsymbol{\varphi}}_k^{(t)}\}_{k=1}^{K_{t-1}+1})$
if $z_t^\star = K_{t-1}+1$ then
    $K_t = K_{t-1}+1$
end if
for $r_t = 1$ to $t$ do
    Evaluate $\Psi_t^{(r)}$ using (9)
    Calculate $p(r_t, \boldsymbol{z}^\star_{1:t})$
    Obtain $p(\boldsymbol{z}^\star_{1:t}) = \sum_{r_t} p(r_t, \boldsymbol{z}^\star_{1:t})$
    Compute $p(r_t|\boldsymbol{z}^\star_{1:t})$
    Update $m_{k,t}^{(r)} \leftarrow m_{k,t-1}^{(r)} + \mathbb{I}\{z_t^\star = k\}$
end for
Return: $r_t^\star = \arg\max p(r_t|\boldsymbol{z}^\star_{1:t})$
```

This factorization is possible due to the chain-rule and the CRP construction described in Section 3.1. Once the complete data log-likelihood is available, we may apply the expectation step (E-step) and the maximization step (M-step) of the EM algorithm. In this work, we have slightly modified the M-step to accept the proposed continual learning framework. Concretely, the estimation of the parameter at each step is simply performed by taking one iterate of a steepest descent method, yielding a stochastic M-step [14]. The E-step amounts to

$$\mathbb{E}[\mathbb{I}\{z_t = k\}|\boldsymbol{z}^\star_{1:t-1}, x_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}] = p(z_t = k|\boldsymbol{z}^\star_{1:t-1}, x_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}) \propto p(x_t|z_t = k, \hat{\boldsymbol{\varphi}}_k^{(t-1)})p(z_t = k|\boldsymbol{z}^\star_{1:t-1}),$$

where $\mathbb{E}[\cdot]$ is the expectation operator, $\hat{\boldsymbol{\varphi}}_k^{(t)}$ is the estimate of $\boldsymbol{\varphi}_k$ at time t, and we have exploited (8). In the M-step, the estimate of the parameters $\{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}$ is updated based on the gradient:

$$\hat{\boldsymbol{\varphi}}_k^{(t)} \leftarrow \hat{\boldsymbol{\varphi}}_k^{(t-1)} + \eta_{k,t}\nabla_{\boldsymbol{\varphi}_k}\mathbb{E}[\mathcal{L}_{\boldsymbol{\varphi}}(\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t})], \tag{11}$$

where $\eta_{k,t}$ is the (adaptive) learning rate for the kth class at time t. In this expression, we have assumed that the same initial learning rate is chosen for the parameters of a given class, but it is possible to select multiple learning rates per class. Once we have the E- and M-steps, we can compute the posterior of z_t and maximize it to obtain $z_t^\star$ as in (4). Finally, Algorithm 1 presents all the necessary computations of the proposed recursive method at each time instant t and the Python implementation can be found in `https://github.com/pmorenoz/continual_ihcpd` for reproducibility purposes.

5. EXPERIMENTS

In this section we evaluate the performance of the proposed method. We apply the infinite-dimensional hierarchical BOCPD algorithm to real-world data, and in particular, to a sequence of raw nuclear magnetic response measurements taken during a well-drilling process. This data consists of 4500 real-valued univariate observations taken

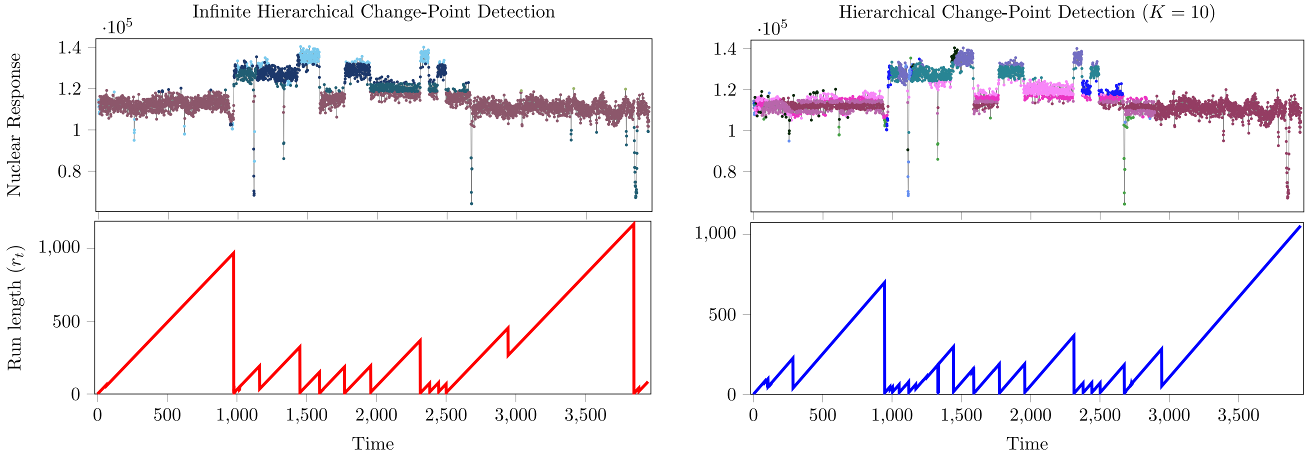


Fig. 2. Upper row plots show the well-drilling univariate signal for the unbounded latent variable model (left) and the hierarchical CPD method (right) with fixed K. The colors represent latent-class assignments. Bottom row plots show the MAP estimates of the run-length.

at a fixed sampling frequency. In the following, we assume that the time steps are ordered and discrete for simplicity.

To apply the proposed model, we choose $p(x_t|z_t = k, \boldsymbol{\varphi}_k)$ to be Gaussian distributed with unknown mean and variance, that is, $\boldsymbol{\varphi}_k = \{\mu_k, \sigma_k^2\}$. Moreover, the model has two hyperparameters that we need to select. The first one, which is related to the CPD method, is the parameter λ of the hazard function that is used as the conditional prior, $p(r_t|r_{t-1})$. In the experiments, we have selected $\lambda = 10^6$. The second one is the parameter α, which is involved in the CRP construction, and controls how likely is the appearance of a new unseen class. We set it to $\alpha = 1.0$. For the stochastic M-step, we use two different adaptive learning rates for the mean and variance whose initial values are given by $\eta_\mu = 1.0$ and $\eta_\sigma = 0.02$. Importantly, we made both learning rates decrease at a rate of 2% per time-step if $z_t = k$ was selected as the most likely latent class. This choice avoids adapting very old parameters with new incoming data.

Figure 2 shows the results obtained for $t = 4500$ iterations.[1] The unbounded model is compared with the hierarchical CPD approach with an upper bound on the number of classes $K = 10$. In the upper figures we can see the well-drilling signals, as well as the latent-class assignments in different colors for both approaches. As can be observed, the final number of classes inferred by the CRP was $K_{4500} = 7$. In the bottom figures we show the MAP estimates of the run-length, $r_t^\star$. These figures show that the MAP estimation of the run-length is well aligned with the signal transitions. Furthermore, it should be noted that the proposed method is more robust to outliers as can be seen for $t \approx 200$ and $t \approx 600$, where the outlier is captured by the latent class assignment but a CP is not declared. In fact, the MAP estimate of the run-length is noisier for the method with a fixed number of classes than for the unbounded model.

In addition, the latent-class assignments look more consistent in the case of the infinite-dimensional hierarchical CPD algorithm, where both the initial and final samples of the well-drilling signal are assigned to the same latent-class. The main two advantages of the method can be observed from the empirical results. First, the method uses past learned parameters to infer assignments over very recent data, that is, assignments coincide along time. Second, the CRP is able to discover new unseen latent-classes without fixing the model complexity *a priori* and avoids the overlapping with previous discovered classes. For instance, if a new latent class k^* appears, it would not coincide with the previous learned ones, and neither their parameters.

Finally, it is important to note that, since the unbounded model creates new classes as they become necessary, its computational complexity is smaller than that of the hierarchical CPD approach, which needs to estimate the parameters of $K = 10$ classes at every time step.

6. DISCUSSION AND FUTURE WORK

This work has extended the Bayesian online change-point detection method to more complex scenarios by considering a hierarchical model, which is based on latent-class variables. To prevent the limitation of fixing the order of the hierarchical model *a priori*, we allow for an unbounded number of classes using the chinese restaurant process. Moreover, the inference of the class assignments is done with an expectation-maximization algorithm, where the M-step is carried out stochastically, that is, only one iteration of a steepest descent method is taken. Finally, the performance of the proposed method is validated empirically over real-world data. We show its robustness and utility for the aforementioned purposes. In future work, it would be interesting to extend it to multi-channel settings with multivariate generative models. Also, instead of introducing a latent-class model, we may consider a feature-based approach, e.g., the indian buffet process (IBP) construction as an alternative method that captures interpretable correlations in the hierarchical layer of the CPD method.

7. REFERENCES

[1] E. Andersson, D. Bock, and M. Frisén, “Some statistical aspects of methods for detection of turning points in business cycles,” *Journal of Applied Statistics*, vol. 33, no. 3, pp. 257–278, 2006.

[2] I. Berkes, E. Gombay, L. Horváth, and P. Kokoszka, “Sequential change-point detection in GARCH (p, q) models,” *Econometric Theory*, vol. 20, no. 6, pp. 1140–1167, 2004.

[3] M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia, “Sequential anomaly detection in the presence of noise and

[1] A video demonstrating the complete simulation of the algorithms is available at `https://www.youtube.com/watch?v=ymZPNURhtIc`.

limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5544–5562, 2012.

[4] V. Krishnamurthy, "Quickest detection POMDPs with social learning: Interaction of local and global decision makers," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5563–5587, 2012.

[5] M. Arts, A. Bollig, and R. Mathar, "Quickest eigenvalue-based spectrum sensing using random matrix theory," *arXiv preprint arXiv:1504.01628v1*, 2015.

[6] L. Du, C.-H. Liu, M. Laghate, and D. Cabric, "Sequential detection of number of primary users in cognitive radio networks," in *Asilomar Conf. Signals, Systems and Computers*, 2015.

[7] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[8] P. Moreno-Muñoz, D. Ramírez, and A. Artés-Rodríguez, "Change-point detection on hierarchical circadian models," *arXiv preprint arXiv:1809.04197*, 2018.

[9] M. B. Ring, "Continual learning in reinforcement environments," Ph.D. dissertation, University of Texas at Austin, 1994.

[10] J. Schmidhuber, "Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem," *Frontiers in Psychology*, vol. 4, p. 313, 2013.

[11] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," *International Conference on Learning Representations (ICLR)*, 2018.

[12] J. Pitman, "Combinatorial stochastic processes," Dept. Statistics, UC Berkeley., Tech. Rep., 2002.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[15] J. F. C. Kingman, "The coalescent," *Stochastic Processes and their Applications*, vol. 13, no. 3, pp. 235–248, 1982.

[16] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," *Encyclopedia of Machine Learning*, pp. 81–89, 2010.