Confirmatory Bayesian Online Change Point Detection in the Covariance Structure of Gaussian Processes

Jiyeon Han[*] , **Kyowoon Lee**[*] , **Anh Tong** and **Jaesik Choi**[†]

Ulsan National Institute of Science and Technology

{jiyeon, leekwoon, anhth, jaesik}@unist.ac.kr

Abstract

In the analysis of sequential data, the detection of abrupt changes is important in predicting future events. In this paper, we propose statistical hypothesis tests for detecting covariance structure changes in locally smooth time series modeled by Gaussian Processes (GPs). We provide theoretically justified thresholds for the tests, and use them to improve Bayesian Online Change Point Detection (BOCPD) by confirming statistically significant changes and non-changes. Our Confirmatory BOCPD (CBOCPD) algorithm finds multiple structural breaks in GPs even when hyperparameters are not tuned precisely. We also provide conditions under which CBOCPD provides the lower prediction error compared to BOCPD. Experimental results on synthetic and real-world datasets show that our proposed algorithm outperforms existing methods for the prediction of nonstationarity in terms of both regression error and log-likelihood.

1 Introduction

In sequential data, the *change point detection* (CPD) problem, i.e., analysis of the data with the aim of detecting abrupt changes, is an important component in improving the prediction of future events. When the underlying distribution does not change (stationary), predicting future events becomes tractable. However, the stationarity assumption does not hold in practice.

A *change point* (CP) is a specific sequential position at which the underlying distribution changes. CPs play critical roles in numerous real-world applications, including climate modeling [Manogaran and Lopez, 2018], speech recognition [Panda and Nayak, 2016], image analysis [Tewkesbury *et al.*, 2015], and human activity recognition [Cleland *et al.*, 2014]. In econometrics, *structural breaks*, which essentially apply CPD to regression models, have been studied for decades in an attempt to identify structural stability in forecasting time series [Chow, 1960; Ewing and Malik, 2016]. Trend filtering determines CPs by assuming piecewise linearity in

the sequential data [Kim *et al.*, 2009]. CPD also plays an important role in the domain adaptation problem known as a covariate shift [Sugiyama *et al.*, 2008], as the distribution of the test data changes from that of the training data.

Most existing CPD methods are based on either statistical (Bayesian) inference or hypothesis tests. Statistical inference methods compute the probability of the occurrence of a change. Bayesian CPD algorithms [Barry and Hartigan, 1993; Xuan and Murphy, 2007] identify CPs using Bayesian framework. BOCPD algorithms [Adams and MacKay, 2007; Garnett *et al.*, 2010; Saatçi *et al.*, 2010; Kim and Choi, 2015] detect CPs sequentially considering the correlated interval, the so-called *run length*, between CPs. Such probabilistic methods, however, do not provide a statistical error bound leading to a lack of reliability and are highly sensitive to selected hyperparameters.

Hypothesis-test-based approaches determine the existence of changes based on a statistical test, where the error probability is naturally determined during the computation. Representative hypothesis test techniques include kernel methods such as two-sample tests based on the maximum mean discrepancy [Gretton *et al.*, 2012; Li *et al.*, 2015], kernel Fisher discriminant ratio [Eric *et al.*, 2008], and likelihood ratio tests (LRTs) such as the cumulative sum (CUSUM) test [Chernoff and Zacks, 1964; Gombay *et al.*, 1996]. While the conventional GPs only deal with stationary (globally smooth) functions, GPs with CPs can mimic locally smooth functions [Saatçi *et al.*, 2010], allowing them to represent many real-world time series data. There have been a number of studies of CPD in GPs using hypothesis tests. For instance, CPD in GPs using p-value test on the likelihood given the null hypothesis of stationary time series has been investigated [Isupova, 2017]. If the null distribution is not legitimate, this approach cannot decisively say there is a change even if the null hypothesis is rejected. Another work has studied to detect the mean changes in GPs through LRTs [Keshavarz *et al.*, 2018].

In this paper, we propose new statistical LRTs that detect changes in the covariance structure of GPs and build a theoretically justified online detection algorithm, Confirmatory BOCPD, which detects CPs with a reasonable time delay. We also present sufficient conditions under which CBOCPD provides the lower prediction error compared to BOCPD. Moreover, our algorithm adjusts the parameter of BOCPD to avoid false alarms and missed detections when the results of hy-
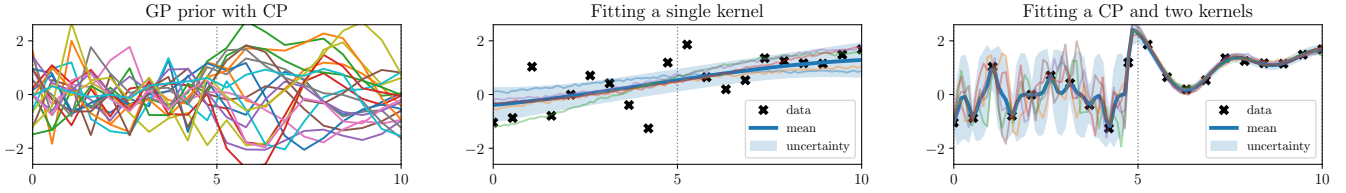
---

[*]Equally contributed

[†]Contact Author

Figure 1: Left: GP prior with a covariance structure change. Middle: GP posterior after fitting data with a single kernel. Right: GP posterior after fitting data with two kernels with a CP.

pothesis tests are sound. When the results are not sound, our algorithm takes advantages of Bayesian inference from BOCPD.

2 Background

2.1 GP Models and Assumptions

A GP is a random process produced by a collection of random variables such that any finite set of those variables follows a multivariate Gaussian distribution. A GP is completely specified by the mean function $\mu(\cdot) = \mathbb{E}[f(\cdot)]$ and the kernel function $K(\cdot, \cdot) = \text{Cov}(f(\cdot), f(\cdot))$. The kernel function describes distinctive characteristics of the sequential data, such as variance, length scales, and periodicity. For example, in the case of the Radial Basis Function kernel $K(t, t') = \sigma^2 \exp(-\frac{(t-t')^2}{2l^2})$, the length scale hyperparameter l controls the smoothness of the sequential data.

In this paper, we assume that observations are corrupted by the white Gaussian noise with the variance σ_{no}^2. In modeling sequential data using GPs, the index t is considered as the input, and the corresponding output is modeled as $x_t \sim N(f(t), \sigma_{no}^2)$. Given the GP hyperparameters θ_m and σ_{no}, the log marginal likelihood of the GP over the observed samples $\mathbf{x}$ is specified by the mean vector μ and the covariance matrix Σ for the multivariate Gaussian distribution.[1]

2.2 Optimal CPD of the Mean

Here, we briefly review some results on the detection of a single CP in the mean function of a GP [Keshavarz *et al.*, 2018]. The goal is to set a formal statistical test that can identify a change in the mean function and to quantify the confidence of a change.

We denote the n observed sequential data as $X = \{X_t\}_{t=1}^n$. Let $t \in \mathcal{C}_n \subseteq \{1, ..., n\}$ represents the point at which a sudden change occurs. Given sequential points, two hypotheses are used for the LRT. One is the null hypothesis, $\mathbb{H}_0$, which insists there is no CP. The other is the alternative hypothesis, $\mathbb{H}_1$, which declares there is at least one CP. The hypothesis testing problem is constructed as follows. Given the two hypotheses, the likelihood ratio (LR) is defined as $2\mathfrak{L} = 2(\sup_{\theta \in \Theta_1} \ell(\theta_1) - \sup_{\theta \in \Theta_0} \ell(\theta_0))$ where ℓ is the log-likelihood function and Θ_0 and Θ_1 are the parameter spaces of $\mathbb{H}_0$ and $\mathbb{H}_1$ respectively.

The generalized likelihood ratio test (GLRT) is formulated as $\mathfrak{T}_{GLRT} = \mathbb{I}(2\mathfrak{L} \geq \mathfrak{R}_{n,\delta})$ with a proper threshold $\mathfrak{R}_{n,\delta}$ and the indicator function $\mathbb{I}(\cdot)$. Here, δ is the upper bound of the

[1]$\log p(\mathbf{x}|\theta_m, \sigma_{no}) = -\frac{1}{2}(\mathbf{x}-\mu)^T(\Sigma + \sigma_{no}^2 I)^{-1}(\mathbf{x}-\mu) - \frac{1}{2}\log|\Sigma + \sigma_{no}^2 I| - \frac{n}{2}\log 2\pi$, where n is the number of observed samples.

corresponding conditional detection error probability $\varphi_n(\mathfrak{T})$, which is defined as

$$\varphi_n(\mathfrak{T}) = \mathbb{P}(\mathfrak{T} = 1|\mathbb{H}_0) + \max_{t \in \mathcal{C}_n} \mathbb{P}(\mathfrak{T} = 0|\mathbb{H}_{1,t}). \quad (1)$$

We reject $\mathbb{H}_0$ when $\mathfrak{T}_{GLRT}=1$, and we fail to reject $\mathbb{H}_0$, otherwise. In the mean change detection problem in a GP, we assume that the samples are generated by a GP with zero mean under the null hypothesis, which can be stated as $\mathbb{H}_0 : \mathbb{E}X = \mathbf{0}$. Under the associative alternative hypothesis at time t, we assume that there is a change of size b in the mean as $\mathbb{H}_{1,t} : \exists b \neq 0, \mathbb{E}X = \frac{b}{2}\zeta_t$, where $\zeta_t \in \mathbb{R}^n$ is given by $\zeta_t(k) := sign(k - t)$ for any $t \in \mathcal{C}_n$. Combining all possible cases, the alternative hypothesis can be written as $\mathbb{H}_1 : \bigcup_{t \in \mathcal{C}_n} \mathbb{H}_{1,t}$, which states that there exists at least one CP with jump size b. Further, we can write $2\mathfrak{L}$ as

$$X^T \Sigma^{-1} X - \min_{t \in \mathcal{C}_n} \min_{b \neq 0} \left[\left(X - \frac{b}{2}\zeta_t \right)^T \Sigma^{-1} \left(X - \frac{b}{2}\zeta_t \right) \right], \quad (2)$$

where Σ is the covariance matrix of X. Maximizing Equation (2) with respect to b and plugging in the test, we obtain the following test.

$$\mathfrak{T}_{GLRT} = \mathbb{I} \left(\max_{t \in \mathcal{C}_n} \left| \frac{\zeta_t^T \Sigma^{-1} X}{\sqrt{\zeta_t^T \Sigma^{-1} \zeta_t}} \right|^2 \geq \mathfrak{R}_{n,\delta} \right).$$

With a suitable threshold $\mathfrak{R}_{n,\delta}$, the error probability is bounded as $\varphi_n(\mathfrak{T}_{GLRT}) \leq \delta$ under the sufficient condition on b [Keshavarz *et al.*, 2018], and $\mathfrak{R}_{n,\delta}$ could be chosen to be

$$\mathfrak{R}_{n,\delta} = 1 + 2\left[log\left(\frac{2n}{\delta}\right) + \sqrt{log\left(\frac{2n}{\delta}\right)} \right].$$

3 CPD of the Covariance

This section presents our new hypothesis tests to detect CPs in the covariance structure of a GP.

3.1 Motivational Examples

Figure 1 shows how CPD of a covariance structure could affect the quality of a GP regression. The left plot in Figure 1 shows samples from a GP with an intended CP in the middle. The middle plot shows samples from a GP model after the hyperparameters have been learnt using the whole datasets. The right plot shows samples from a GP model whose covariance structure breaks and the hyperparameters have been learnt separately. Figure 1 suggests that fitting nonstationary data to

a time-invariant GP results in an imprecise model. GP regression with a structural break in the covariance structure is more expressive and better suited to the analysis of nonstationary data.

3.2 Likelihood Ratio Test

To construct a test for detecting changes in the covariance structure, we define the null hypothesis as $\mathbb{H}_0 : \mathrm{Cov}(X_i, X_j) = K(i,j)$ and the alternative hypothesis as $\mathbb{H}_1 = \bigcup_{t \in \mathcal{C}_n} \mathbb{H}_{1,t}$, with

$$\mathbb{H}_{1,t} : \mathrm{Cov}(X_i, X_j) = \begin{cases} K(i,j), & i,j < t \\ K'(i,j), & i,j \geq t \\ K''(i,j), & \text{otherwise} \end{cases} \quad (3)$$

where K, K' and K'' are the kernel functions. Let Σ and Σ'_t denote the covariance matrices for $\mathbb{H}_0$ and $\mathbb{H}_{1,t}$, respectively. The LR $2\mathfrak{L}$ is written as

$$\max_{t \in \mathcal{C}_n} \left[X^T (\Sigma)^{-1} X - X^T (\Sigma'_t)^{-1} X + \ln \left(\frac{|\Sigma|}{|\Sigma'_t|} \right) \right]. \quad (4)$$

Theorem 3.1. *Let* $t^* = argmin_t \left[X^T (\Sigma'_t)^{-1} X + \ln |\Sigma'_t| \right]$. *Under the null hypothesis, i.e.,* $X \sim N(\mathbf{0}, \Sigma)$,

$$2\mathfrak{L} - \ln \left(\frac{|\Sigma|}{|\Sigma'_{t^*}|} \right) = \sum_{i=1}^{n} v_i - \sum_{i=1}^{n} \lambda_i u_i$$

where $\lambda_1, ..., \lambda_n$ *are the eigenvalues of* $\Sigma^{\frac{1}{2}} (\Sigma'_{t^*})^{-1} \Sigma^{\frac{1}{2}}$ *and* $u_i, v_i \sim \chi_1^2$, *a chi-squared distribution with degree 1.*

Theorem 3.2. *Let* $t^* = argmin_t \left[X^T (\Sigma'_t)^{-1} X + \ln |\Sigma'_t| \right]$. *Under the alternative hypothesis, i.e.,* $X \sim N(\mathbf{0}, \Sigma'_{t^*})$,

$$2\mathfrak{L} - \ln \left(\frac{|\Sigma|}{|\Sigma'_{t^*}|} \right) = \sum_{i=1}^{n} \lambda_i u_i - \sum_{i=1}^{n} v_i$$

where $\lambda_1, ..., \lambda_n$ *are the eigenvalues of* $\Sigma'^{\frac{1}{2}}_{t^*} \Sigma^{-1} \Sigma'^{\frac{1}{2}}_{t^*}$ *and* $u_i, v_i \sim \chi_1^2$, *a chi-squared distribution with degree 1.*

Theorems 3.1 and 3.2 show that the difference of two positive semi-definite quadratic terms can be expressed as a subtraction between a chi-square random variable with n degrees of freedom and a linear combination of independent chi-square random variables having one degree of freedom each. See Appendix[2] A for the proofs of Lemmas and Theorems in Sections 3 and 4.

3.3 Tests for the Covariance Structural Break

For the case where the covariance structure breaks into two different kernels, $\mathbb{H}_0$ and $\mathbb{H}_{1,t}$ are similarly defined as in Section 3.2 except that $K''(i,j) = 0$. The corresponding covariance matrices can then be written as

$$\Sigma = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}, \Sigma'_t = \begin{pmatrix} K_{aa} & 0 \\ 0 & K'_{bb} \end{pmatrix}.$$

Here, K_{rc} for $r, c \in \{a, b\}$ indicates the covariance matrix between X_r and X_c with kernel K where $X_a := X_{1:t}$ and $X_b := X_{t+1:n}$. We now define a LRT as $\mathfrak{T} = \mathbb{I}(2\mathfrak{L} \geq \mathfrak{R}_\delta)$. For further Lemmas and Theorems, we define constant C_t as follows.

---

[2] Appendix is available at https://arxiv.org/abs/1905.13168

Definition 1. *For the covariance matrices* Σ *and* Σ'_t, $C_t = \frac{1}{\lambda_{min}(\Sigma)} + \frac{1}{\lambda_{min}(\Sigma'_t)}$ *where* $\lambda_{min}(M)$ *represents the smallest eigenvalue of matrix* M.

Lemma 3.1. $C_t \leq C_0$ *for all* $t \in [1, n]$ *when* $C_0 = \frac{1}{\lambda_{min}(\Sigma)} + \frac{1}{\lambda_{min}(\Sigma) \wedge \lambda_{min}(\Sigma'_{t=n})}$ *where* $\wedge$ *stands for the minimum operator.*

Lemma 3.2. *Let* Q_t *be* $X^T((\Sigma)^{-1} - (\Sigma'_t)^{-1})X$. *When* X_t *is bounded with* $X_t \in [-V, V]$ *for all* $t \in [1, n]$, $Q_t - \mathbb{E}[Q_t]$ *is* $\frac{C_t V^2 n}{2}$-*subgaussian.*

Lemma 3.2 shows that, for bounded X_t, the tail probability of the centered quadratic term in the LR decays approximately as fast as the Gaussian with zero mean and the same variance as the centered quadratic term.

Lemma 3.3. *The probability that the* $\mathfrak{T}$ *is correct under the null hypothesis (absence of CP) is at least* $1 - \delta/2$, *i.e.,* $\mathbb{P}(2\mathfrak{L} \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0}|\mathbb{H}_0) \leq \delta/2$, *for* $\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \max_t \left(n - Tr(\Sigma(\Sigma'_t)^{-1}) + \ln \left(\frac{|\Sigma|}{|\Sigma'_t|} \right) \right) + C_0 V^2 n \sqrt{0.5 \ln(2/\delta)}$.

Lemma 3.4. *The probability that the* $\mathfrak{T}$ *is correct under the alternative hypothesis (existence of a CP) is at least* $1 - \delta/2$, *i.e.,* $\mathbb{P}(2\mathfrak{L} \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}|\mathbb{H}_1) \leq \delta/2$, *for* $\mathfrak{R}_{n,\delta,\mathbb{H}_1} = \min_t \left(Tr(\Sigma'_t(\Sigma)^{-1}) - n + \ln \left(\frac{|\Sigma|}{|\Sigma'_t|} \right) \right) - C_0 V^2 n \sqrt{0.5 \ln(2/\delta)}$.

Using Lemma 3.2 and the concentration inequality, Lemmas 3.3 and 3.4 show that we can control the type I or type II errors to be bounded by $\delta/2$.

Theorem 3.3. *For* $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$, $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$ *in Lemmas 3.3 and 3.4, when* $\mathfrak{R}_{n,\delta,\mathbb{H}_1} \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0}$ *and* $\mathfrak{R}_{n,\delta,\mathbb{H}_0} \leq \mathfrak{R}_\delta \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, *the conditional detection error probability is bounded as*

$$\varphi_n(\mathfrak{T}) = \mathbb{P}(2\mathfrak{L} \geq \mathfrak{R}_\delta|\mathbb{H}_0) + \max_{t \in \mathcal{C}_n} \mathbb{P}(2\mathfrak{L} \leq \mathfrak{R}_\delta|\mathbb{H}_{1,t}) \leq \delta.$$

Proof. It follows directly from Lemmas 3.3 and 3.4. $\square$

Using Theorem 3.3, we guarantee that the LRT for a general covariance kernel change is statistically correct for any error bound δ under specified conditions. If we set the threshold to be greater than or equal to the upper epsilon bound of the null distribution, $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$, we can guarantee a bounded type I error. If we set the threshold to be less than or equal to the lower epsilon bound of the alternative distribution, $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$, we can guarantee a bounded type II error.

The inequalities for $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$ and $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$ can have three possible cases. If $\mathfrak{R}_{n,\delta,\mathbb{H}_0} > \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, there is no threshold guaranteeing both type I and type II errors. If $\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, there is only one threshold that can guarantee both type I and type II errors. If $\mathfrak{R}_{n,\delta,\mathbb{H}_0} < \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, the thresholds that can guarantee both type I and type II errors are indicated by the shaded area in Figure 3 in Appendix A.

4 Confirmatory Bayesian Online CPD

This section presents our algorithmic contribution which improves the performance of Bayesian online CPD (BOCPD) with statistical hypothesis tests.

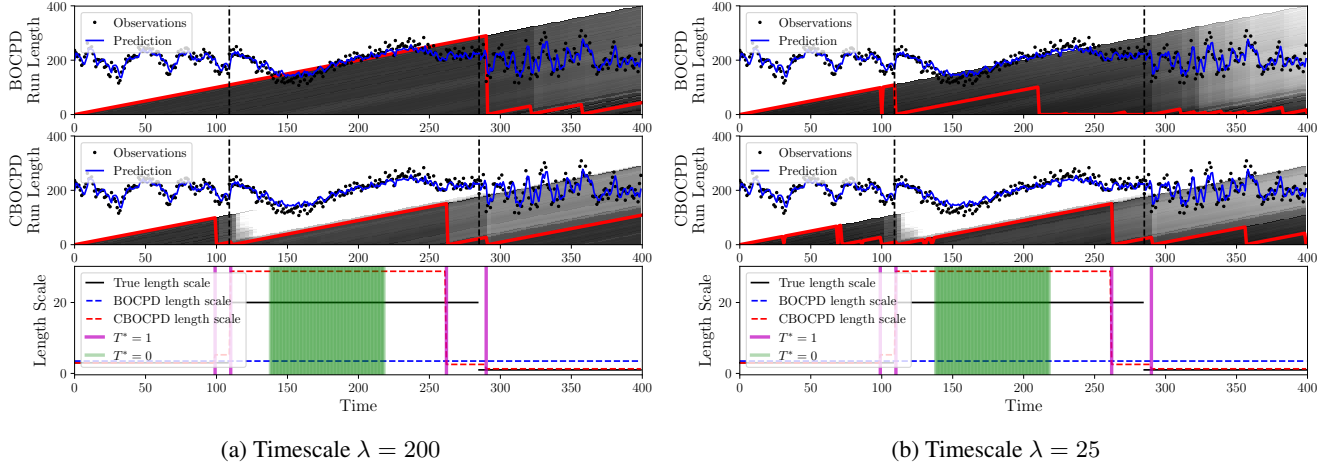(a) Timescale $\lambda = 200$

(b) Timescale $\lambda = 25$

Figure 2: Synthetic data with changes in hyperparameters and the derived run-length distribution computed from BOCPD (top) and CBOCPD (middle). The bottom figure shows the hyperparameter values learned from each algorithm with the true value of the hyperparameters used to generate the synthetic data. Dashed black lines indicate the true CPs.

4.1 Bayesian Online CPD

Many real-world applications require CPs to be detected in an online manner. BOCPD [Adams and MacKay, 2007] uses Bayesian inference to update the distribution of future data based on the distribution of the run length. We assume that, if we divide data with CPs, the data in different partitions will be independent whereas the data in the same partition will be autocorrelated.

The formulation of BOCPD is as follows. Let r_t be the run length at time t and x_t be the data observation at time t. We denote a set of consecutive discrete observations between time a and b inclusive as $x_{a:b}$. Let $x_t^{(r_t)}$ denote the data observations since the most recent CP. We simply use $x_t^{(r)}$ instead of $x_t^{(r_t)}$ when there is no ambiguity. Then, we calculate the distribution of the next data x_{t+1} given the data up to time t. Through marginalization of r_t, this can be written as

$$\mathbb{P}(x_{t+1}|x_{1:t}) = \sum_{r_t} \mathbb{P}(x_{t+1}|x_t^{(r)})\mathbb{P}(r_t|x_{1:t}). \quad (5)$$

The joint distribution of the run length and data $\mathbb{P}(r_t, x_{1:t})$ is written as,

$$\sum_{r_{t-1}} \mathbb{P}(r_t|r_{t-1}, x_{t-1}^{(r)})\mathbb{P}(x_t|r_{t-1}, x_{t-1}^{(r)})\mathbb{P}(r_{t-1}, x_{1:t-1}). \quad (6)$$

We can see that $\mathbb{P}(r_t, x_{1:t})$ has a recursive form with respect to time. The first term in Equation (6) can be substituted by $\mathbb{P}(r_t|r_{t-1})$ under the assumption that the current run length only depends on the previous run length. Here, the conditional prior[3] of r_t is given by

$$\mathbb{P}(r_t|r_{t-1}) = \begin{cases} H(r_{t-1}+1) & r_t = 0 \\ 1 - H(r_{t-1}+1) & r_t = r_{t-1}+1 \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

$\mathbb{P}_{gap}(g)$ is the a priori probability distribution over the interval between CPs. When $\mathbb{P}_{gap}(g)$ is a geometric distribution

---

[3] $H(\tau)$ is the hazard function with $H(\tau) = \frac{\mathbb{P}_{gap}(g=\tau)}{\sum_{t=\tau}^{\infty}\mathbb{P}_{gap}(g=t)}$.

with timescale λ, the hazard function becomes constant as $H(\tau) = 1/\lambda$. The second term in Equation (6) can be calculated using a GP. Thus, we can calculate the overall distribution $\mathbb{P}(x_{t+1}|x_{1:t})$ by recursive message passing scheme with $\mathbb{P}(r_t, x_{1:t})$.

The BOCPD framework efficiently determines changes in GP modeling. The change is modeled by considering all the possible run lengths. One of the main assumptions of this framework is that parameters such as the kernel parameters or the hazard function do not change. However, this online CPD framework is vulnerable to the parameters, because inappropriate parameters degrade predictive performance on non-stationary data leading imprecise run length distribution as shown in Figure 2.

4.2 Confirmatory BOCPD

Algorithm 1 presents a theoretically justified online change detection algorithm, CBOCPD. The main idea of CBOCPD is to overcome the limitations of the assumption that the run length is independent of the data. However, we claim that the first term in Equation (6) can be directly calculated by the LRT in Section 3.3. The first two lines in Algorithm 1 initialize the parameters. In lines 3–13, Equation (7) is altered as

$$\mathbb{P}(r_t = 0|r_{t-1}, x_{t-1}^{(r)}) = \begin{cases} 1-\delta, & \tau^* = t \text{ and } \mathfrak{T}^* = 1 \\ \delta, & \mathfrak{T}^* = 0 \\ H_{const}, & \text{otherwise.} \end{cases}$$

Here, we have two LRTs, $\mathfrak{T}^0 = \mathbb{I}(2\mathfrak{L} \geq \mathfrak{R}_{\delta,\mathbb{H}_0})$ and $\mathfrak{T}^1 = \mathbb{I}(2\mathfrak{L} \geq \mathfrak{R}_{\delta,\mathbb{H}_1})$, with thresholds in line 4. Empirical thresholds are used since the theoretically calculated thresholds in Lemmas 3.3 and 3.4 are not tight enough to use in practice. We further define $\mathfrak{T}^* = 1$ if $\mathfrak{T}^1 = 1$, and $\mathfrak{T}^0 = 1$ and $\mathfrak{T}^* = 0$ if $\mathfrak{T}^1 = 0$ and $\mathfrak{T}^0 = 0$. The LRTs are applied to the window around t with $W = x_{t-m:t+m}$. τ^* is the time point in the window that maximizes the LR, $\tau^* = \text{argmax}_{\tau \in \mathcal{C}_W} 2\mathfrak{L}_\tau$. Here $\mathcal{C}_W \subseteq \{t-m,...,t+m\}$ is a set of change point candidates for the window. If both LRTs at time τ^* pass

Algorithm 1 Confirmatory BOCPD

1: $m, \delta \leftarrow$ half window size, error bound for the LRT
2: $\mathbb{P}(x_0) \leftarrow \mathcal{N}(\mu_{prior}, \sigma^2_{prior})$
3: **for** $t \in [1, T]$ **do**
4:     $\mathfrak{R}_{\delta, \mathbb{H}_0}, \mathfrak{R}_{\delta, \mathbb{H}_1} \leftarrow$ thresholds for $\mathfrak{T}^0, \mathfrak{T}^1$
5:     $H \leftarrow H_{const}$
6:     **if** $m < t < T - m$ **then**
7:         set window $W = x_{t-m:t+m}$
8:         $2\mathcal{L}_\tau \leftarrow$ the LR between $\mathbb{H}_0$ and $\mathbb{H}_{1,\tau}$ with data W
9:         $\tau^*, 2\mathcal{L} = \text{argmax}_{\tau \in \mathcal{C}_W} 2\mathcal{L}_\tau, \max_{\tau \in \mathcal{C}_W} 2\mathcal{L}_\tau$
10:         **if** $\mathfrak{T}^1 = 1$ and $\mathfrak{T}^0 = 1$ and $\tau^* = t$ **then**
11:             $H \leftarrow 1 - \delta$
12:         **else if** $\mathfrak{T}^1 = 0$ and $\mathfrak{T}^0 = 0$ **then**
13:             $H \leftarrow \delta$
14:     $\pi_t^{(r)} \leftarrow \mathbb{P}(x_t | x_{t-1}^{(r)})$
15:     $\mathbb{P}(r_t = r_{t-1}+1, x_{1:t}) \leftarrow \mathbb{P}(r_{t-1}, x_{1:t-1})\pi_t^{(r)}(1-H)$
16:     $\mathbb{P}(r_t = 0, x_{1:t-1}) \leftarrow \sum_{r_{t-1}} \mathbb{P}(r_{t-1}, x_{1:t-1})\pi_t^{(r)}H$
17:     $\mathbb{P}(x_{1:t}) \leftarrow \sum_{r_t} \mathbb{P}(r_t, x_{1:t})$
18:     $\mathbb{P}(r_t | x_{1:t}) \leftarrow \mathbb{P}(r_t, x_{1:t}) / \mathbb{P}(x_{1:t})$
19:     $\mathbb{P}(x_{t+1} | x_{1:t}) \leftarrow \sum_{r_t} \mathbb{P}(x_{t+1} | x_t^{(r)})\mathbb{P}(r_t | x_{1:t})$

and τ^* coincides with t, we decide that t is a CP and set $\mathbb{P}(r_t = 0 | r_{t-1}, x_{t-1}^{(r)}) = 1 - \delta$, which enhances the probability of change in the BOCPD framework. In contrast, if neither test passes, we strongly believe there is no change and reduce the probability of change in the BOCPD framework. This is why we name this algorithm *Confirmatory* BOCPD. We add $\tau^* = t$ to avoid situations where the same time point maximizes the LR in several consecutive windows, resulting in duplicate CPs. Lines 14–19 follow the BOCPD framework [Adams and MacKay, 2007].

4.3 Theoretical Analysis of CBOCPD

We present sufficient conditions under which CBOCPD provides the lower prediction error compared to BOCPD. Here, the prediction error is defined as the expected absolute difference between the predictive mean given the true run length and the predictive mean under BOCPD and CBOCPD at a (detected) CP t, as a good prediction at a CP is a key factor in overall performance. For simplicity, we denote the expected value of x_t under BOCPD and CBOCPD as $\mathbb{E}_{BO}[x_t|x_{1:t-1}]$ and $\mathbb{E}_{CBO}[x_t|x_{1:t-1}]$, respectively. Further we define $\alpha_i = 1 - \mathbb{P}_{BO}(r_{t-1} = i|x_{1:t-1})$ under BOCPD and $\beta_i = 1 - \mathbb{P}_{CBO}(r_{t-1} = i|x_{1:t-1})$ under CBOCPD.

Non-stationary Case
We first investigate the case where a change exists.

Theorem 4.1. *Consider BOCPD (Sec. 4.1) and CBOCPD (Sec. 4.2) algorithms where CBOCPD uses two statistical tests $\mathfrak{T}^0$ and $\mathfrak{T}^1$ with type II error of δ_0^{II} and δ_1^{II} respectively. When there exists a CP at t with mean value of μ_1 which satisfies*

$$\forall i \in [0, t-1], \epsilon_L \leq |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x_{i:t-1}]| \leq \epsilon_U$$

where $\mathbb{E}[x_t|\emptyset]$ indicates the expected value of x_t with prior distribution, if

$$\frac{\epsilon_U}{\epsilon_L} \leq \alpha_0 \left(1 + \frac{(1 - \delta_0^{II})(1 - \delta_1^{II})}{\delta_0^{II}\delta_1^{II}}\right)$$

with α_0 indicating the probability of the run length not to be zero, then the absolute error of expected value of CBOCPD at t is less than or equal to the one of BOCPD as

$$\mathbb{E}[|\mu_1 - \mathbb{E}_{BO}[x_t|x_{1:t-1}]|] \geq \mathbb{E}[|\mu_1 - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|].$$

Stationary Case
Next, we investigate the case where a change does not exist.

Theorem 4.2. *Consider BOCPD (Sec. 4.1) and CBOCPD (Sec. 4.2) algorithms where CBOCPD uses two statistical tests $\mathfrak{T}^0$ and $\mathfrak{T}^1$ with type I error of δ_0^I and δ_1^I respectively. When there exists a confirmed non-CP at t with mean value of μ_2 which satisfies*

$$\forall i \in [0, t-1], \epsilon_L \leq |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}[x_t|x_{i:t-1}]| \leq \epsilon_U,$$

$$\text{if } \frac{\epsilon_U}{\epsilon_L} \leq \frac{\alpha_{t-1}((1 - \delta_0^I)(1 - \delta_1^I) + \delta_0^I\delta_1^I)}{\beta_{t-1}(1 - \delta_0^I)(1 - \delta_1^I) + \delta_0^I\delta_1^I}$$

with α_{t-1} and β_{t-1} indicating the probability of the run length not to be $t-1$ under BOCPD and CBOCPD respectively, then the absolute error of the expected value of CBOCPD at t is less than or equal to the one of BOCPD as

$$\mathbb{E}[|\mu_2 - \mathbb{E}_{BO}[x_t|x_{1:t-1}]|] \geq \mathbb{E}[|\mu_2 - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|].$$

5 Experimental Evaluations

5.1 Synthetic Data

In this experiment, we investigate two synthetic datasets generated by GPs with changes in the length scale and the variance of a Radial Basis Function kernel, respectively. Observations are obtained by adding the white Gaussian noise with the variance $\sigma^2_{no} = 0.1$. For both datasets, two CPs are drawn uniformly from time intervals $(75, 125)$ and $(275, 325)$ with end time $T = 400$.

For the first experiment, the length scale l switches from 3 to 20 and from 20 to 1 with the variance $\sigma^2 = 1$. For the second experiment, the variance σ^2 switches from 1 to 4 and from 4 to 0.3 with the length scale $l = 3$. We compare the proposed CBOCPD with BOCPD and CS (CUSUM)-BOCPD. For BOCPD, we use a modified version of the stationary GP-BOCPD [Saatçi *et al.*, 2010] by fixing the timescale λ. CS-BOCPD uses CUSUM instead of the our proposed LRT in Algorithm 1. For all algorithms, the timescale of $\lambda = 200$ is used.

Figure 2a and Figure 2b show that CBOCPD identifies the length scale change in the data with the help of statistical tests, whereas BOCPD captures the change too less or too many times. The first two columns in Table 1 present the quantitative results from 100 runs of each algorithm after training on the first 100 points and testing on the remaining 300 points. CBOCPD outperforms BOCPD in terms of both negative log-likelihood (NLL) and mean squared error (MSE). The differences in performances become larger when there is a stronger correlation among the samples.

	LEN-CHANGE		VAR-CHANGE		Gazebo:Env1		Gazebo:Env2		Gazebo:Env3	
Method	NLL	MSE	NLL	MSE	NLL	MSE	NLL	MSE	NLL	MSE
BOCPD	0.99±0.38	0.47±0.32	1.13±0.61	0.55±0.58	2.07±0.51	0.14±0.05	2.24±0.48	0.57±0.26	0.28±0.12	0.11±0.03
CS-BOCPD	0.95±0.32	0.43±0.30	1.05±0.42	0.48±0.40	-0.11±0.33	0.12±0.04	1.82±1.51	0.55±0.27	0.25±0.16	0.10±0.02
CBOCPD	**0.79±0.33**	**0.34±0.24**	**0.89±0.36**	**0.41±0.34**	**-0.31±0.34**	**0.11±0.04**	**0.69±0.36**	**0.45±0.19**	**-0.99±0.47**	0.10±0.04

Table 1: Comparison of BOCPD, CS-BOCPD, and CBOCPD with NLL and MSE on synthetic and Gazebo robot simulation datasets.

Method	NLL	p-value	MSE	p-value
Nile Data (200 training points, 463 test points)				
ARGP	1.07±0.64	<0.0001	5.06±0.86	0.0005
ARGP-BOCPD	0.78±0.72	<0.0001	4.94±0.87	0.0017
GPTS	0.86±0.64	<0.0001	4.78±0.81	0.0100
BOCPD	0.57±0.77	0.0014	4.73±0.82	0.0115
CBOCPD	**0.00±0.80**	N/A	**4.32±0.74**	N/A
Well Log Data (500 training points, 3050 test points)				
ARGP	7.20±0.60	<0.0001	17.3±2.6	<0.0001
ARGP-BOCPD	**0.00±0.30**	N/A	**4.68±0.46**	N/A
GPTS	3.73±0.42	<0.0001	8.27±0.61	<0.0001
BOCPD	4.35±0.31	<0.0001	19.2±1.3	<0.0001
CBOCPD	0.30±0.27	0.0010	4.92±0.44	0.2124
Snow Data (500 training points, 13380 test points)				
ARGP	17.48±0.82	<0.0001	14.82±0.57	<0.0001
ARGP-BOCPD	0.06±0.39	<0.0001	9.65±0.39	<0.0001
GPTS	16.60±0.22	<0.0001	8.76±0.36	<0.0001
BOCPD	**0.00±0.39**	N/A	9.43±0.38	N/A
CBOCPD	1.92±0.37	<0.0001	**6.34±0.27**	<0.0001

Table 2: Results of predictive performance on Nile data, Well Log Data and Snow Data. The results are provided with 95% of confidence interval and the p-value of the null hypothesis that a method is equivalent to the best performing method according to NLL, using a one sided t-test.

5.2 Gazebo Robot Simulation Data

We also conducted experiments on the Gazebo robot simulator to detect changes in the environment of a robot. We gathered data by moving the robot through a changing environment (i.e., varying ground properties). We used the Pioneer3AT robot with kinetic ROS. There are three environments. In the first environment (Env1), the ground changes at the midpoint from 'Plane ground' to 'Bumpy ground 1'. In the second environment (Env2), the robot moves from 'Bumpy ground 1' to 'Bumpy ground 2' where 'Bumpy ground 2' is more coarser than 'Bumpy ground 1'. Finally, in the third environment (Env3), the robot moves from 'Bumpy ground 2' back to the 'Plane ground'. For simplicity, we used only z-directional position data as the inputs. The last three columns in Table 1 show that CBOCPD outperforms other methods in all environments. Interestingly, the results show that CUSUM test does not help BOCPD much when the variance decreases as in Env3 but proposed LRT improves BOCPD in all the cases.

5.3 Real World Data

Further, we compare our proposed algorithm with autoregressive GP (ARGP), autoregressive GP with BOCPD (ARGP-BOCPD), GP time series with BOCPD (BOCPD) and GP time series model (GPTS). For baseline methods we used the publicly available source codes[4].

We first consider Nile data which records the lowest annual water levels on the Nile river during AD 622-1284 measured at the island of Roda. There is domain knowledge suggesting a CP in the year 715 due to the construction of a 'nilometer' [Popper, 1951; Balek, 1977]. Secondly, we used Well Log data which records 4050 nuclear magnetic resonance measurements while drilling a well. We also considered Snowfall data which records daily snowfall in Whistler, BC, Canada. In this dataset each method is evaluated by the ability to predict snowfall of next day using 35 years of test data.

Table 2 shows the predictive performance of our proposed algorithm compared to other GP based BOCPD methods. In Nile data, we see that combining BOCPD method with GP improves the performance. CBOCPD further improves the performance by leveraging the proposed statistical hypothesis tests and outperforms all other algorithms. In Well Log data, the nonlinear temporal correlations within each regime give a slight advantage to underlying predictive model of ARGP. However, CBOCPD still shows the competitive result. In Snowfall data, CBOCPD well detects the difference in noise levels and achieves the best performance in MSE.

6 Conclusion

This paper presents a novel framework for detecting changes in the covariance structure of the GP. We propose a online CPD algorithm, Confirmatory BOCPD, which is an improved version of BOCPD with statistical hypothesis tests. Experiments using synthetic and real-world datasets demonstrate that CBOCPD outperforms conventional BOCPDs.

Acknowledgments

---

[4]https://sites.google.com/site/wwwturnercomputingcom/software/ThesisCodeAndData.zip

References

[Adams and MacKay, 2007] Ryan Prescott Adams and David J.C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[Balek, 1977] Jaroslav Balek. *Hydrology and Water Resources in Tropical Africa*. Developments in water science. Elsevier Scientific Publishing Company, 1977.

[Barry and Hartigan, 1993] Daniel Barry and John A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, pages 309–319, 1993.

[Chernoff and Zacks, 1964] Herman Chernoff and Shelemyahu Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, pages 999–1018, 1964.

[Chow, 1960] Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, pages 591–605, 1960.

[Cleland et al., 2014] Ian Cleland, Manhyung Han, Chris Nugent, Hosung Lee, Sally McClean, Shuai Zhang, and Sungyoung Lee. Evaluation of prompted annotation of activity data recorded from a smart phone. *Sensors*, pages 15861–15879, 2014.

[Eric et al., 2008] Moulines Eric, Francis R. Bach, and Zaïd Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In *Proceedings of the Neural Information Processing Systems*, pages 609–616, 2008.

[Ewing and Malik, 2016] Bradley T. Ewing and Farooq Malik. Volatility spillovers between oil prices and the stock market under structural breaks. *Global Finance Journal*, pages 12–23, 2016.

[Garnett et al., 2010] Roman Garnett, Michael A. Osborne, Steven Reece, Alex Rogers, and Stephen J. Roberts. Sequential bayesian prediction in the presence of changepoints and faults. *The Computer Journal*, pages 1430–1446, 2010.

[Gombay et al., 1996] Edit Gombay, Lajos Horváth, and Marie Husková. Estimators and tests for change in variances. *Statistics and Risk Modeling*, pages 145–160, 1996.

[Gretton et al., 2012] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, pages 723–773, 2012.

[Isupova, 2017] Olga Isupova. *Machine learning methods for behaviour analysis and anomaly detection in video*. PhD thesis, University of Sheffield, 2017.

[Keshavarz et al., 2018] Hossein Keshavarz, Clayton Scott, and XuanLong Nguyen. Optimal change point detection in gaussian processes. *Journal of Statistical Planning and Inference*, pages 151–178, 2018.

[Kim and Choi, 2015] Taehoon Kim and Jaesik Choi. Reading documents for bayesian online change point detection. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 1610–1619, 2015.

[Kim et al., 2009] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. l_1 trend filtering. *SIAM review*, pages 339–360, 2009.

[Li et al., 2015] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In *Proceedings of the Neural Information Processing Systems*, pages 3366–3374, 2015.

[Manogaran and Lopez, 2018] Gunasekaran Manogaran and Daphne Lopez. Spatial cumulative sum algorithm with big data analytics for climate change detection. *Computers and Electrical Engineering*, pages 207–221, 2018.

[Panda and Nayak, 2016] Soumya Priyadarsini Panda and Ajit Kumar Nayak. Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology*, pages 9–18, 2016.

[Popper, 1951] William Popper. *The Cairo Nilometer: Studies in Ibn Taghrî Birdî's Chronicles of Egypt: I*. Publications in Semitic Philology. University of California Press, 1951.

[Saatçi et al., 2010] Yunus Saatçi, Ryan D. Turner, and Carl E. Rasmussen. Gaussian process change point models. In *Proceedings of the International Conference on Machine Learning*, pages 927–934, 2010.

[Sugiyama et al., 2008] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, pages 699–746, 2008.

[Tewkesbury et al., 2015] Andrew P. Tewkesbury, Alexis J. Comber, Nicholas J. Tate, Alistair Lamb, and Peter F. Fisher. A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*, pages 1–14, 2015.

[Xuan and Murphy, 2007] Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the International Conference on Machine Learning*, pages 1055–1062, 2007.