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 [Cléland 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; Saatçî 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 [Saatçî 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-
which declares there is at least one CP. The hypothesis testing
A GP is a random process produced by a collection of ran-
tom variables such that any finite set of those variables follows
a multivariate Gaussian distribution. A GP is com-
pletely 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 hyperparam-
eter \( l \) controls the smoothness of the sequential data.

In this paper, we assume that observations are corrupted
by the white Gaussian noise with the variance \( \sigma^2_{no} \). In mod-
eling sequential data using GPs, the index \( t \) is considered as
the input, and the corresponding output is modeled as
\( x_t \sim N(f(t), \sigma^2_{no}) \). Given the GP hyperparameters \( \theta_m \) and
\( \sigma_{no} \), the log marginal likelihood of the GP over the observed
samples \( x \) is specified by the mean vector \( \mu \) and the covariance
matrix \( \Sigma \) for the multivariate Gaussian distribution.\(^1\)

2 Background

2.1 GP Models and Assumptions

A GP is a random process produced by a collection of ran-

dom variables such that any finite set of those variables follows

a multivariate Gaussian distribution. A GP is com-
pletely 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 hyperparam-
eter \( l \) controls the smoothness of the sequential data.

In this paper, we assume that observations are corrupted
by the white Gaussian noise with the variance \( \sigma^2_{no} \). In mod-
eling sequential data using GPs, the index \( t \) is considered as
the input, and the corresponding output is modeled as
\( x_t \sim N(f(t), \sigma^2_{no}) \). Given the GP hyperparameters \( \theta_m \) and
\( \sigma_{no} \), the log marginal likelihood of the GP over the observed
samples \( x \) is specified by the mean vector \( \mu \) and the covariance
matrix \( \Sigma \) 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 \( t \in \mathcal{C}_n \subseteq \{1, \ldots, n\} \) represents the point at which a sudden
change occurs. Given sequential points, two hypotheses are
used for the LRT. One is the null hypothesis, \( H_0 \), which insists
there is no CP. The other is the alternative hypothesis, \( H_1 \),
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
\( 2\mathcal{L} = 2(\sup_{\Theta_0} \ell(\theta)) - (\sup_{\Theta_0} \ell(\theta_0)) \) where \( \ell \) is the log-likelihood function and \( \Theta_0 \)
and \( \Theta_1 \) are the parameter spaces of \( H_0 \) and \( H_1 \) respectively.

The generalized likelihood ratio test (GLRT) is formulated as
\( \mathcal{S}_{GLRT} = \# \left( 2\mathcal{L} \geq \mathbb{R}_{n, \delta} \right) \) with a proper threshold \( \mathbb{R}_{n, \delta} \) and the indicator function \( \#(\cdot) \). Here, \( \delta \) is the upper bound of the

\(^1\)log \( p(x|\theta_m, \sigma_{no}) = -\frac{1}{2}(x-\mu)^T(\Sigma + \sigma^2_{no}I)^{-1}(x-\mu) - \frac{1}{2} \log |\Sigma + \sigma^2_{no}I| - \frac{1}{2} \log 2\pi, \) where \( n \) is the number of observed samples.
To construct a test for detecting changes in the covariance structure, we define the null hypothesis as \( \mathbb{H}_0 : \text{Cov}(X_t, X_j) = K(i, j) \) and the alternative hypothesis as \( \mathbb{H}_1 = \bigcup_{t \in C_n} \mathbb{H}_{1,t} \), with

\[
\mathbb{H}_{1,t} : \text{Cov}(X_t, X_j) = \begin{cases} 
K(i, j), & i, j < t \\
K'(i, j), & i, j \geq t \\
K''(i, j), & \text{otherwise} 
\end{cases}
\]

where \( K, K' \) and \( K'' \) are the kernel functions. Let \( \Sigma \) and \( \Sigma' \) denote the covariance matrices for \( \mathbb{H}_0 \) and \( \mathbb{H}_1 \), respectively. The LR 2\( \mathcal{L} \) is written as

\[
\max_{t \in C_n} \left[ X^T (\Sigma)^{-1} X - X^T (\Sigma')^{-1} X + \ln \left( \frac{|\Sigma|}{|\Sigma'|} \right) \right].
\]

**Theorem 3.1.** Let \( t^* = \arg\min_{t} \left[ X^T (\Sigma')^{-1} X + \ln |\Sigma'| \right] \).

Under the null hypothesis, i.e., \( X \sim N(0, \Sigma) \),

\[
2\mathcal{L} - \ln \left( \frac{|\Sigma|}{|\Sigma'|} \right) = \sum_{i=1}^{n} v_i - \sum_{i=1}^{n} \lambda_i u_i
\]

where \( \lambda_1, ..., \lambda_n \) are the eigenvalues of \( \Sigma' \), \( \Sigma'^{-1} \) and \( u_i, v_i \sim \chi^2_1 \), a chi-squared distribution with degree 1.

**Theorem 3.2.** Let \( t^* = \arg\min_{t} \left[ X^T (\Sigma')^{-1} X + \ln |\Sigma'| \right] \).

Under the alternative hypothesis, i.e., \( X \sim N(0, \Sigma') \),

\[
2\mathcal{L} - \ln \left( \frac{|\Sigma|}{|\Sigma'|} \right) = \sum_{i=1}^{n} \lambda_i u_i - \sum_{i=1}^{n} v_i
\]

where \( \lambda_1, ..., \lambda_n \) are the eigenvalues of \( \Sigma'^{-1} \), \( \Sigma'^{-2} \) and \( u_i, v_i \sim \chi^2_1 \), 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 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

\[
\Sigma = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}, \quad \Sigma' = \begin{pmatrix} K_{aa} & 0 \\ 0 & K_{bb} \end{pmatrix}.
\]

Here, \( K_{rc} \) for \( r, c \in \{ a, b \} \) indicates the covariance matrix between \( X_r \) and \( X_c \) with kernel \( K \) where \( X_a := X_{2:2} \) and \( X_b := X_{1:1} \). We now define a LRT as \( \mathcal{T} = 1 \left( 2\mathcal{L} \geq C_{B} \right) \).

For further Lemmas and Theorems, we define constant \( C_{t} \) as follows.

**Definition 1.** For the covariance matrices \( \Sigma \) and \( \Sigma' \), \( C_t = \frac{1}{\lambda_{\min}(\Sigma)} + \frac{1}{\lambda_{\min}(\Sigma')}, \)

where \( \lambda_{\min}(\Sigma) \) represents the smallest eigenvalue of matrix \( \Sigma \).

**Lemma 3.1.** \( C_t \leq C_0 \) for all \( t \in \{1, n\} \) when \( C_0 = \frac{1}{\lambda_{\min}(\Sigma)} + \frac{1}{\lambda_{\min}(\Sigma')}, \)

where \( \wedge \) stands for the minimum operator.

**Lemma 3.2.** Let \( Q_t \) be \( X^T (\Sigma)^{-1} - (\Sigma')^{-1} X \). When \( X_t \) is bounded with \( X_t \in [-V, V] \) for all \( t \in \{1, n\} \), \( Q_t - E[Q_t] \) is \( C_t V^2 n \)-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 \( \mathcal{T} \) is correct under the null hypothesis (absence of CP) is at least \( 1 - \delta/2 \), i.e., \( \mathbb{P}(2\mathcal{L} \geq \mathbb{H}_0, \mathbb{H}_0) \leq \delta/2, \)

for \( \mathbb{H}_0, \mathbb{H}_0 = \max_{t} \left( n - \frac{1}{2} \ln \right) \left( \frac{|\Sigma|}{|\Sigma'|} \right) \) + \( C_0 V^2 n \ln(2/\delta) \).

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 \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \), in Lemmas 3.3 and 3.4, when \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \),

the conditional detection error probability is bounded as

\[
\mathbb{P}(2\mathcal{L} \geq \mathbb{H}_0, \mathbb{H}_0) + \max_{t \in C_n} \mathbb{P}(2\mathcal{L} \leq \mathbb{H}_1, \mathbb{H}_1) \leq \delta.
\]

**Proof.** It follows directly from Lemmas 3.3 and 3.4.

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

The inequalities for \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \), can have three possible cases. If \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \), there is no threshold guaranteeing both type I and type II errors. If \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \), there is only one threshold that can guarantee both type I and type II errors. If \( \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_1, \), 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.
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 \( x_t(\cdot) \) denote the data observations since the most recent CP. We simply use \( x_t(\cdot) \) instead of \( x_t(\cdot) \) 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

\[
\mathbb{P}(x_{t+1}|x_{1:t}) = \sum_{r_t} \mathbb{P}(x_{t+1}|x_t(r)) \mathbb{P}(r_t|x_{1:t}).
\]

(5)

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

\[
\sum_{r_t} \mathbb{P}(r_t|r_{t-1}, x_t(r)) \mathbb{P}(x_t|r_{t-1}, x_t(r)) \mathbb{P}(r_{t-1}, x_{1:t-1}).
\]

(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

\[
\mathbb{P}(r_t|r_{t-1}) = \begin{cases} 
H(r_{t-1} + 1) & r_t = 0 \\
1 - H(r_{t-1} + 1) & r_t = r_{t-1} + 1 \\
0 & \text{otherwise}
\end{cases}
\]

(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 with timescale \( \lambda \), 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

\[
\mathbb{P}(r_t = 0|r_{t-1}, x_t(r)) = \begin{cases} 
1 - \delta, & \tau^* = t \text{ and } \Xi^* = 1 \\
\delta, & \Xi^* = 0 \\
H_{const}, & \text{otherwise}
\end{cases}
\]

Here, we have two LRTs, \( \Xi^0 = \mathbb{I}(2\delta \geq \delta_{\text{th}}) \) and \( \Xi^1 = \mathbb{I}(2\delta \geq \delta_{\text{th}}) \), 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 \( \Xi^* = 1 \) if \( \Xi^1 = 1 \), and \( \Xi^* = 1 \) and \( \Xi^0 = 0 \) if \( \Xi^1 = 0 \) and \( \Xi^0 = 0 \). The LRTs are applied to the window around \( t \) with \( W = x_{t-m:t+m} \). \( \tau^* \) is the time point in the window that maximizes the LR. \( \tau^* = \text{argmax}_{\tau \in W} \Xi^0 \).

Here \( C_W \subseteq \{t-m, ..., t+m\} \) is a set of change point candidates for the window. If both LRTs at time \( \tau^* \) pass...
Algorithm 1 Confirmatory BOCPD

1: \( m, \delta \leftarrow \) half window size, error bound for the LRT
2: \( \mathbb{P}(x_0) \leftarrow \mathcal{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}}) \)
3: for \( t \in [1, T] \) do
4: \( \mathcal{H}_{t, \delta}, \mathcal{H}_{t, \delta_1} \leftarrow \) thresholds for \( \mathcal{S}_0, \mathcal{S}_1 \)
5: \( H \leftarrow H_{\text{const}} \)
6: if \( m < t < T - m \) then
7: set window \( W = [x_{t-m:t+m}] \)
8: \( 2\sigma_t \leftarrow \) the LR between \( \mathbb{H}_0 \) and \( \mathbb{H}_{1, \tau} \) with data \( W \)
9: \( \tau^*, 2\sigma^* = \arg\max_{\tau \in \mathbb{C}_W} 2\sigma_T, \max_{\tau \in \mathbb{C}_W} 2\sigma_T \)
10: if \( \mathcal{S}_1 = 1 \) and \( \mathcal{S}_0 = 1 \) and \( \tau^* = t \) then
11: \( H \leftarrow 1 - \delta \)
12: else if \( \mathcal{S}_1 = 0 \) and \( \mathcal{S}_0 = 0 \) then
13: \( H \leftarrow \delta \)
14: \( \pi_t^{(r)} \leftarrow \mathbb{P}(x_t | x_{t-1}^{(r)}) \)
15: \( \mathbb{P}(r_t = r_{t-1} + 1, x_{t-1}^{(r)}) \leftarrow \mathbb{P}(r_{t-1}, x_{t-1}^{(r-1)})\pi_t^{(r)}(1-H) \)
16: \( \mathbb{P}(r_t = 0, x_{t-1}^{(r)}) \leftarrow \sum_{r_{t-1}} \mathbb{P}(r_{t-1}, x_{t-1}^{(r-1)})\pi_t^{(r)}H \)
17: \( \mathbb{P}(x_{t-1}) \leftarrow \sum_{r_{t-1}} \mathbb{P}(x_{t-1}^{(r-1)}) \)
18: \( \mathbb{P}(x_{t-1} | x_{t-1}^{(r)}) \leftarrow \frac{\mathbb{P}(x_{t-1}^{(r-1)})}{\mathbb{P}(x_{t-1}^{(r)})} \)
19: \( \mathbb{P}(x_{t+1} | x_t) \leftarrow \sum_{r_t} \mathbb{P}(x_{t+1} | x_t^{(r)}) \mathbb{P}(r_t | x_t) \)

and \( \tau^* \) coincides with \( t \), we decide that \( t \) is a CP and set \( \mathbb{P}(r_t = 0 | r_{t-1}, x_{t-1}^{(r-1)}) = 1 - \delta \), 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 CBOPCD

We present sufficient conditions under which CBOPCD 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 CBOPCD 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 CBOPCD as \( \mathbb{E}_{BO}[x_t | x_{t-1}] \) and \( \mathbb{E}_{CBO}[x_t | x_{t-1}] \), respectively. Further, we define \( \alpha_t = 1 - \mathbb{E}_{BO}[r_{t-1} = i | x_{t-1}] \) under BOCPD and \( \beta_t = 1 - \mathbb{E}_{CBO}[r_{t-1} = i | x_{t-1}] \) under CBOPCD.

Non-stationary Case

We first investigate the case where a change exists.

**Theorem 4.1.** Consider BOCPD (Sec. 4.1) and CBOPCD (Sec. 4.2) algorithms where CBOPCD uses two statistical tests \( \mathcal{S}_0 \) and \( \mathcal{S}_1 \) with type II error of \( \delta_0^u \) and \( \delta_1^u \) respectively. When there exists a CP at \( t \) with mean value of \( \mu_1 \) which satisfies

\[
\forall i \in [0, t-1], \epsilon_L \leq |\mathbb{E}[x_i] - \mathbb{E}[x_i | x_{i-1}^{(r)}]| \leq \epsilon_U
\]

where \( \mathbb{E}[x_i] \) indicates the expected value of \( x_i \) with prior distribution, if

\[
\frac{\epsilon_U}{\epsilon_L} \leq \alpha_0 \left( 1 + \frac{(1 - \delta_0^u)(1 - \delta_1^u)}{\delta_0^u \delta_1^u} \right)
\]

with \( \alpha_0 \) indicating the probability of the run length not to be zero, then the absolute error of expected value of CBOPCD at \( t \) is less than or equal to the one of BOCPD as

\[
\mathbb{E}[|\mu_1 - \mathbb{E}_{BO}[x_t | x_{t-1}^{(r)}]|] \geq \mathbb{E}[|\mu_1 - \mathbb{E}_{CBO}[x_t | x_{t-1}^{(r)}]|].
\]

**Stationary Case**

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

**Theorem 4.2.** Consider BOCPD (Sec. 4.1) and CBOPCD (Sec. 4.2) algorithms where CBOPCD uses two statistical tests \( \mathcal{S}_0 \) and \( \mathcal{S}_1 \) with type I error of \( \delta_0^l \) and \( \delta_1^l \) respectively. When there exists a confirmed non-CP at \( t \) with mean value of \( \mu_2 \) which satisfies

\[
\forall i \in [0, t-1], \epsilon_L \leq |\mathbb{E}[x_i] - \mathbb{E}[x_i | x_{i-1}^{(r)}]| \leq \epsilon_U,
\]

if

\[
\frac{\epsilon_U}{\epsilon_L} \leq \beta_{t-1} \left( 1 - \delta_0^l \right) \left( 1 - \delta_1^l \right) + \delta_0^l \delta_1^l
\]

with \( \alpha_{t-1} \) and \( \beta_{t-1} \) indicating the probability of the run length not to be \( t - 1 \) under BOCPD and CBOPCD respectively, then the absolute error of the expected value of CBOPCD at \( t \) is less than or equal to the one of BOCPD as

\[
\mathbb{E}[|\mu_2 - \mathbb{E}_{BO}[x_t | x_{t-1}^{(r)}]|] \geq \mathbb{E}[|\mu_2 - \mathbb{E}_{CBO}[x_t | x_{t-1}^{(r)}]|].
\]

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 \( \sigma^2_{\text{no}} = 0.1 \). 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 \( \sigma^2 \) switches from 1 to 4 and from 4 to 0.3 with the length scale \( l = 3 \). We compare the proposed CBOPCD with BOCPD and CS (CUSUM)-BOCPD. For BOCPD, we use a modified version of the stationary GP-BOPCD [Saatçi et al., 2010] by fixing the timescale \( \lambda \). CS-BOPCD 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 CBOPCD 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. CBOPCD 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.
Table 1: Comparison of BOCPD, CS-BOCPD, and CBOCPD with NLL and MSE on synthetic and Gazebo robot simulation datasets. 

<table>
<thead>
<tr>
<th>Method</th>
<th>NLL</th>
<th>MSE</th>
<th>NLL</th>
<th>MSE</th>
<th>NLL</th>
<th>MSE</th>
<th>NLL</th>
<th>MSE</th>
<th>NLL</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOCPD</td>
<td>0.99±0.38</td>
<td>0.47±0.32</td>
<td>1.13±0.61</td>
<td>0.55±0.58</td>
<td>2.07±0.51</td>
<td>0.14±0.05</td>
<td>2.24±0.48</td>
<td>0.57±0.26</td>
<td>0.28±0.12</td>
<td>0.11±0.03</td>
</tr>
<tr>
<td>CS-BOCPD</td>
<td>0.95±0.32</td>
<td>0.43±0.30</td>
<td>1.05±0.42</td>
<td>0.48±0.40</td>
<td>-0.11±0.33</td>
<td>0.12±0.04</td>
<td>1.82±1.51</td>
<td>0.55±0.27</td>
<td>0.25±0.16</td>
<td>0.10±0.02</td>
</tr>
<tr>
<td>CBOCPD</td>
<td>0.79±0.33</td>
<td>0.34±0.24</td>
<td>0.89±0.36</td>
<td>0.41±0.34</td>
<td>-0.31±0.34</td>
<td>0.11±0.04</td>
<td>0.69±0.36</td>
<td>0.45±0.19</td>
<td>-0.99±0.47</td>
<td>0.10±0.04</td>
</tr>
</tbody>
</table>

Table 2: Results of predictive performance on Nile data, Well Log Data and Snow Data. The results are provided with 95% 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Nile Data (200 training points, 463 test points)</th>
<th>Well Log Data (500 training points, 3050 test points)</th>
<th>Snow Data (500 training points, 13380 test points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGP</td>
<td>1.07±0.64 &lt;0.0001  5.06±0.86 0.0005</td>
<td>7.20±0.60 &lt;0.0001  17.3±2.6  &lt;0.0001</td>
<td>17.48±0.82 &lt;0.0001  14.82±0.57  &lt;0.0001</td>
</tr>
<tr>
<td>ARGP-BOCPD</td>
<td>0.78±0.72 &lt;0.0001  4.94±0.87 0.0017</td>
<td>0.00±0.00 N/A 4.68±0.46 N/A</td>
<td>0.00±0.00 N/A 4.60±0.36 N/A</td>
</tr>
<tr>
<td>GPTS</td>
<td>0.86±0.64 &lt;0.0001  4.78±0.81 0.0100</td>
<td>3.73±0.42 &lt;0.0001  8.27±0.61  &lt;0.0001</td>
<td>3.73±0.33 &lt;0.0001  8.27±0.61  &lt;0.0001</td>
</tr>
<tr>
<td>BOCPD</td>
<td>0.57±0.77 0.0014  4.73±0.82 0.0115</td>
<td>4.35±0.31 &lt;0.0001  19.2±1.3  &lt;0.0001</td>
<td>4.35±0.31 &lt;0.0001  19.2±1.3  &lt;0.0001</td>
</tr>
<tr>
<td>CBOCPD</td>
<td>0.00±0.80 N/A 4.32±0.74 N/A</td>
<td>0.30±0.27 &lt;0.0010  4.92±0.44 0.2124</td>
<td>0.30±0.27 &lt;0.0010  4.92±0.44 0.2124</td>
</tr>
</tbody>
</table>

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.

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 an 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

This work was supported by the Industrial Convergence Core Technology Development Program funded by the Ministry of Trade, Industry & Energy (MOTIE), Korea (No. 10063172, Development of robot intelligence technology for mobility with learning capability toward robust and seamless indoor and outdoor autonomous navigation) and Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (MSIT), Korea (No. 2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence).
References


