Conjugate Bayesian Two-step Change Point Detection for Hawkes Process

Zeyue Zhang[1,2], **Xiaoling Lu**[1,2], **Feng Zhou**[1,3*]
[1]Center for Applied Statistics and School of Statistics, Renmin University of China
[2]Innovation Platform, Renmin University of China
[3]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing
{zhangzeyue, xiaolinglu, feng.zhou}@ruc.edu.cn

Abstract

The Bayesian two-step change point detection method is popular for the Hawkes process due to its simplicity and intuitiveness. However, the non-conjugacy between the point process likelihood and the prior requires most existing Bayesian two-step change point detection methods to rely on non-conjugate inference methods. These methods lack analytical expressions, leading to low computational efficiency and impeding timely change point detection. To address this issue, this work employs data augmentation to propose a conjugate Bayesian two-step change point detection method for the Hawkes process, which proves to be more accurate and efficient. Extensive experiments on both synthetic and real data demonstrate the superior effectiveness and efficiency of our method compared to baseline methods. Additionally, we conduct ablation studies to explore the robustness of our method concerning various hyperparameters. Our code is publicly available at https://github.com/Aurora2050/CoBay-CPD.

1 Introduction

Point process data, characterized by a series of discrete events occurring over time, finds extensive applications in various domains, including finance [1], neuroscience [29], and social networks [20]. Hawkes processes [8], a subclass of point processes, have gained attention due to their ability to model self-exciting and clustering behaviors. However, traditional modeling of Hawkes processes relies on the assumption that the parameters of the process, i.e., the distribution of the process, remain invariant over time. In practice, this assumption often fails to hold, as the underlying dynamics of point process data can change over time [23; 27].

The violation of the time-invariant parameter assumption poses a significant challenge in accurately modeling Hawkes process data. Real-world processes are inherently dynamic and subject to various influences, leading to fluctuations in process distribution. For instance, in a social media platform, the intensity of user interactions may change due to a sudden surge in activity during an emergency event or a significant drop in user engagement during a service outage. To address this issue, the change point detection (CPD) has emerged as a pivotal task in analyzing point process data. CPD aims to identify the locations where the parameters (distribution) of the process undergo a significant change.

In this work, our focus is on CPD within Hawkes process. Previous studies have tackled this issue, offering diverse methodologies for this purpose [2; 4; 23; 27]. Among these, the two-step estimation-prediction method [4] has gained widespread application due to its simplicity and intuitiveness. This method involves using historical data to estimate model parameters and then utilizing these estimated parameters to estimate the distribution of the next event point. If the observed point aligns closely

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

with the prediction, model parameters are assumed unchanged. But if there's a notable deviation, it suggests parameter change, indicating a current change point.

An issue with the estimation-prediction method lies in the inaccurate estimation of model parameters. This issue stems from the constraints within the CPD system: on one hand, for efficient detection, we cannot use a large amount of historical data as it would lead to a heavy computational burden; on the other hand, once a change point is identified, the reliance on samples post-change point for estimating subsequent model parameters restricts the availability of adequate data for initial post-change point estimations. This often results in inaccurate parameter estimations due to our reliance on a limited historical dataset, subsequently affecting the prediction of the next event point. As a consequence, the algorithm may wrongly identify many non-change points as change points (false positives) and vice versa (false negatives).

To alleviate the aforementioned issues, many studies propose utilizing Bayesian approaches. Compared to frequentist methods, Bayesian methods exhibit better robustness as they not only consider samples but also incorporate prior knowledge. When data is limited, prior knowledge acts as regularization, effectively preventing overfitting. Specifically, the Bayesian estimation-prediction method estimates the posterior distribution of model parameters based on historical data and then leverages this posterior distribution to estimate the predictive distribution of the next event point. This predictive distribution considers all possible model specifications, which differs from the frequentist method. Similarly, if the observed point closely matches the prediction, it indicates no change point, suggesting unchanged model parameters. Conversely, a significant deviation suggests a change point[2].

Past studies have investigated the Bayesian Hawkes process [10; 21]. Due to the non-conjugacy between point process likelihoods and any priors, inferring the posterior of Hawkes process parameters presents significant challenges. Currently, the majority of work utilizes methods such as Markov chain Monte Carlo (MCMC) [18] or variational inference [3] to infer the posterior in non-conjugate scenarios. Inference methods derived in such non-conjugate scenarios often lack analytical expressions, leading to low computational efficiency. Therefore, they are not well-suited for the CPD system, which requires timeliness.

To address this challenge, this paper employs a data augmentation strategy recently proposed in the Bayesian point process field [5; 15; 26; 29; 30]. This strategy augments the Hawkes process likelihood with auxiliary latent variables, enabling the augmented Hawkes process likelihood conditionally conjugate to the prior. Leveraging the conditionally conjugate model, we can derive an analytical Gibbs sampler that enables closed-form iterative sampling. The effectiveness and efficiency of our proposed method are demonstrated through experiments using both synthetic and real data.

Specifically, we make the following contributions: **(1)** We propose the conjugate Bayesian two-step change point detection (CoBay-CPD) for Hawkes process, which leverages data augmentation to address the non-conjugate issue. This novel method allows for more accurate and efficient CPD in Hawkes process. **(2)** We develop an analytical Gibbs sampler tailored for the proposed model, enabling closed-form iterative sampling of the model parameters. This streamlines the inference process and alleviates the computational burden associated with non-conjugate scenarios. **(3)** The experiments demonstrate that our method achieves accurate and timely detection of change points in Hawkes process compared to baseline models, which highlights its practical applicability across various dynamic event modeling scenarios.

2 Related Works

In this section, we delve into the existing literature concerning CPD and Bayesian inference for Hawkes processes.

2.1 Change Point Detection for Hawkes Process

CPD in Hawkes processes has remained an exceptionally challenging endeavor. Within the frequentist framework, several methods have been proposed. For instance, techniques based on second-order statistics have been introduced [27], as well as approaches leveraging the cumulative sum (CUSUM)

[2]Here, "Bayesian" refers to the Bayesian treatment of model parameters; for the change point, it remains a point estimation.

method [23] and sequential testing strategies [2]. However, Bayesian approaches have remained underexplored. A pioneering contribution by [4] introduced a Bayesian approach based on Stein variational inference [13] tailored for Hawkes processes. However, this method derived in non-conjugate scenarios lacks analytical expressions, leading to low computational efficiency. Subsequent experiments demonstrate that our CoBay-CPD significantly enhances the computational efficiency.

2.2 Bayesian Inference for Hawkes Process

The Bayesian Hawkes process has been a hot topic in research, broadly categorized into two classes: parametric and non-parametric methods. Parametric approaches [12; 21] involve applying priors to parameters in Hawkes process and inferring the posterior. Non-parametric methods [24; 25; 28], offering greater flexibility than parametric ones, model the background rate and influence function (refer to Eq. (1)) as flexible functions, apply priors (commonly Gaussian processes) on these functions, and aim to infer the posterior of these functions. As the point process likelihood is not conjugate to any priors, both parametric and non-parametric approaches face challenges in posterior inference. Most studies rely on methods such as MCMC, variational inference, or Laplace approximation [14].

2.3 Data Augmentation for Hawkes Process

In recent years, the Bayesian point process field has introduced a novel data augmentation technique to address non-conjugate inference challenges. This method introduces auxiliary latent variables into point process likelihood, transforming non-conjugate problems into conditionally conjugate ones, and thus enabling the derivation of fully analytical inference algorithms. Parametric studies are referenced in [29; 30], while non-parametric research is detailed in [16; 22; 26]. With analytical expressions, the inference algorithms based on data augmentation exhibit higher computational efficiency than the non-analytical ones derived in non-conjugate scenarios. Our study adopts the data augmentation, opting for a computationally more efficient parametric approach to ensure the efficiency of CPD.

3 Methodology

In this section, we present our proposed CoBay-CPD method for Hawkes process. We outline its key derivation steps while providing further details in Appendix A.

3.1 Hawkes Process with Inhibition

The mathematical foundation of Hawkes process is defined by the conditional intensity function that represents the instantaneous rate of event occurrences at time t given the history up to but not including t. It is defined as follows:

$$\lambda^*(t) = \lambda(t|\mathcal{H}_{t^-}) = \mu + \sum_{t_i<t} \phi(t - t_i), \tag{1}$$

where μ is the background rate, $\phi(\cdot)$ is the influence function representing the self-excitation effect from event occurring at t_i to t, the summation captures the influence of all past events, $\mathcal{H}_{t^-}$ is the historical information up to but not including t, and $*$ indicates the intensity depends on the history. The self-exciting property of Hawkes processes allows events to trigger additional events, leading to clustering and bursty behavior in event sequences.

Traditional Hawkes processes only employ positive influence functions to avoid negative intensity, limiting them to capturing excitatory interactions. To incorporate both excitatory and inhibitory effects, many studies [7; 11; 17] have proposed nonlinear Hawkes process that allows the influence functions to be negative. This study adopts the nonlinear Hawkes process proposed by [30] which is defined as:

$$\lambda^*(t) = \bar{\lambda}\sigma(h(t)), \quad h(t) = \mu + \sum_{t_i<t} \phi(t - t_i),$$

where $\bar{\lambda} > 0$ is the intensity upperbound, $\sigma(\cdot)$ denotes the sigmoid function, $\mu \in \mathbb{R}$ is the baseline activation and $\phi(\cdot) \in \mathbb{R}$ is the influence function. Due to the presence of the sigmoid function $\sigma(\cdot)$, both μ and $\phi(\cdot)$ can be negative, allowing it to capture inhibitory effects. We choose the sigmoid as the link function due to its compatibility with the subsequent data augmentation technique.

For a flexible influence function, we model $\phi(\cdot)$ as a linear combination of multiple basis functions:

$$\phi(\cdot) = \sum_{b=1}^{B} w_b \tilde{\phi}_b(\cdot),$$

where $\tilde{\phi}_b(\cdot)$ is the b-th basis function and $w_b \in \mathbb{R}$ is the mixing weight. Following [30], we select the scaled and shifted beta densities with support $[0, T_\phi]$ as basis functions. We define the basis function with bounded support $[0, T_\phi]$ rather than unbounded support $[0, \infty]$ to assume that events occurring too early do not influence the current time. This choice ensures a more efficient computation of $\mathbf{\Phi}(t)$ afterward (details are provided in complexity analysis). Consequently, the formulation of $h(t)$ can be expressed in vector form:

$$h(t) = \mu + \sum_{t_i<t} \phi(t - t_i) = \mu + \sum_{t_i<t} \sum_{b=1}^{B} w_b \tilde{\phi}_b(t - t_i) = \mu + \sum_{b=1}^{B} w_b \sum_{t_i<t} \tilde{\phi}_b(t - t_i) = \mathbf{w}^\top \mathbf{\Phi}(t),$$

where $\mathbf{w} = [\mu, w_1, \ldots, w_B]^\top$, $\mathbf{\Phi}(t) = [1, \Phi_1(t), \ldots, \Phi_B(t)]^\top$ and $\Phi_b(t) = \sum_{t_i<t} \tilde{\phi}_b(t - t_i)$ represents the cumulative impact of past events on t through the b-th basis function. As a result, the probability density function of the proposed model is:

$$p(t_{1:N}|\mathbf{w}, \bar{\lambda}) = \prod_{i=1}^{N} \bar{\lambda}\sigma(h(t_i)) \exp\left(-\int_0^T \bar{\lambda}\sigma(h(t))dt\right),$$

where we assume $t_{1:N}$ are observed on $[0, T]$ and the model parameters are $\mathbf{w}$ and $\bar{\lambda}$.

3.2 Non-conjugate Bayesian CPD

The above section outlines the Hawkes process without change points. In this section, we introduce the Hawkes process with change points and how the Bayesian two-step CPD is designed to detect these change points. In Bayesian two-step CPD, our goal is to identify the change point where the underlying dynamics of point process shift. The two steps in this method involve *an estimation step* and *a prediction step*. Let $t_{1:m}$ represent the sequence of timestamps that is generated by a Hawkes process with parameters $\boldsymbol{\theta}$, and we consider $\boldsymbol{\theta}$ may undergo changes at certain timestamps. We define, for the timestamp t_m, the nearest change point's index as $\tau_m \in \{1, \ldots, m\}$ and assume that the timestamps before and after the change point are mutually independent. Following this assumption, during *the estimation step*, the Bayesian two-step CPD necessitates estimating the posterior of the model parameters based on $t_{\tau_m:m}$. To infer the posterior, we express the likelihood for the timestamps $t_{\tau_m:m}$ after the change point as:

$$p(t_{\tau_m:m}|\mathbf{w}, \bar{\lambda}) = \prod_{i=\tau_m}^{m} \bar{\lambda}\sigma(h(t_i)) \exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right). \tag{2}$$

According to Bayes' theorem, the posterior of model parameters is expressed as:

$$p(\mathbf{w}, \bar{\lambda}|t_{\tau_m:m}) \propto p(t_{\tau_m:m}|\mathbf{w}, \bar{\lambda})p(\mathbf{w})p(\bar{\lambda}), \tag{3}$$

where we choose the prior of $\mathbf{w}$ as Gaussian $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{K})$ and the prior of $\bar{\lambda}$ as an uninformative improper prior $p(\bar{\lambda}) \propto 1/\bar{\lambda}$. We use a Gaussian prior on $\mathbf{w}$ because it is equivalent to an L_2 regularizer, which stabilizes parameter estimation when there is insufficient observed data.

Then, in *the prediction step*, we leverage the posterior of model parameters to compute the predictive distribution of the next timestamp as:

$$p(t_{m+1}|t_{\tau_m:m}) = \iint p(t_{m+1}|t_{\tau_m:m}, \mathbf{w}, \bar{\lambda})p(\mathbf{w}, \bar{\lambda}|t_{\tau_m:m})d\mathbf{w}d\bar{\lambda}. \tag{4}$$

This formula calculates the distribution of the next timestamp t_{m+1} given the observed data points $t_{\tau_m:m}$. It is worth noting that this predictive distribution takes into account all possible model specifications, which is a key distinction between Bayesian and frequentist methods.

In implementation, solving Eqs. (3) and (4) is challenging. For Eq. (3), the non-conjugate nature between the point process likelihood and the prior prevents us from obtaining an analytical posterior.

For Eq. (4), evaluating the integral is also intractable. Therefore, we resort to sampling methods for approximation: **(1)** Use MCMC to obtain parameter samples from the posterior $\{\mathbf{w}^{(k)}, \bar{\lambda}^{(k)}\}_{k=1}^{K} \sim p(\mathbf{w}, \bar{\lambda}|t_{\tau_m:m})$. **(2)** Based on the sampled parameters, use the thinning algorithm [19] to sample the next timestamp $\{t_{m+1}^{(k)} \sim p(t_{m+1}|t_{\tau_m:m}, \mathbf{w}^{(k)}, \bar{\lambda}^{(k)})\}_{k=1}^{K}$. Create a confidence interval based on the samples of $\{t_{m+1}^{(k)}\}_{k=1}^{K}$. If the actual t_{m+1} falls within this interval, we conclude that no change point has occurred. Conversely, we infer the presence of a change point.

3.3 Conjugate Bayesian CPD

For non-conjugate Bayesian CPD, the MCMC algorithm in step 1 often lacks analytical expressions, significantly impacting the timeliness of CPD due to its low computational efficiency. To address this issue, our CoBay-CPD adopts the data augmentation strategy, which augments the Hawkes process likelihood with auxiliary latent variables, enabling the augmented likelihood conditionally conjugate to the prior. Based on the conditionally conjugate model, we derive an analytical Gibbs sampler to effectively obtain posterior samples of parameters. Specifically, we incorporate Pólya-Gamma variables and marked Poisson processes. Similar derivations have been presented in [30]; here we restate the key formulas for clarity.

3.3.1 Augmentation of Pólya-Gamma Variables

The sigmoid function $\sigma(\cdot)$ can be represented in the form of a Gaussian scale mixture:

$$\sigma(z) = \int_0^\infty e^{f(\omega,z)} p_{\text{PG}}(\omega|1,0) d\omega,$$

where $f(\omega, z) = z/2 - z^2\omega/2 - \log 2$ and $p_{\text{PG}}(\omega|1,0)$ denotes the Pólya-Gamma distribution with $\omega \in \mathbb{R}^+$. When substituting the above expression into the product term in Eq. (2), the parameter $\mathbf{w}$ within the model takes on a Gaussian form.

3.3.2 Augmentation of Marked Poisson Process

A marked Poisson process can be introduced to linearize the exponential integral term in Eq. (2):

$$\exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right) = \mathbb{E}_{p_\lambda}\left[\prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}\right],$$

where $\Pi = \{(\omega_r, t_r)\}_{r=1}^{R}$ denotes a realization of a marked Poisson process on the interval $[t_{\tau_m}, t_m]$, with its probability measure denoted as p_λ and having an intensity $\lambda(t,\omega) = \bar{\lambda} p_{\text{PG}}(\omega|1,0)$. Notably, the key difference between our proposed change point model and the prior work by [30] lies in the fact that here, we focus on the interval with change points, $[t_{\tau_m}, t_m]$, rather than the entire domain.

3.3.3 Augmented Joint Distribution

After introducing two sets of latent variables into Eq. (2), we obtain the augmented likelihood:

$$p(t_{\tau_m:m}, \boldsymbol{\omega}, \Pi|\mathbf{w}, \bar{\lambda}) = \prod_{i=\tau_m}^{m} [\lambda(t_i, \omega_i) e^{f(\omega_i, h(t_i))}] p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))},$$

where $\boldsymbol{\omega}$ is the vector of ω_i on each t_i in $t_{\tau_m:m}$, $\lambda(t_i, \omega_i) = \bar{\lambda} p_{\text{PG}}(\omega_i|1,0)$. The parameter $\mathbf{w}$ in the augmented likelihood takes on a Gaussian form, making it conditionally conjugate to the Gaussian prior. Combining the augmented likelihood with priors, we obtain the augmented joint distribution:

$$p(t_{\tau_m:m}, \boldsymbol{\omega}, \Pi, \mathbf{w}, \bar{\lambda}) = p(t_{\tau_m:m}, \boldsymbol{\omega}, \Pi|\mathbf{w}, \bar{\lambda}) p(\mathbf{w}) p(\bar{\lambda}).$$

3.3.4 Gibbs Sampler

Thanks to the conditional conjugacy of the augmented joint distribution, we can derive closed-form conditional densities for all variables, naturally leading to an analytical Gibbs sampler (derivation

provided in Appendix A):

$$p(\boldsymbol{\omega}|t_{\tau_m:m}, \mathbf{w}) = \prod_{i=\tau_m}^{m} p_{\text{PG}}(\omega_i|1, h(t_i)), \tag{5a}$$

$$\Lambda(t, \omega|t_{\tau_m:m}, \mathbf{w}, \bar{\lambda}) = \bar{\lambda}\sigma(-h(t))p_{\text{PG}}(\omega|1, h(t)), \tag{5b}$$

$$p(\bar{\lambda}|t_{\tau_m:m}, \Pi) = p_{\text{Ga}}(\bar{\lambda}|N_m + R, T_m), \tag{5c}$$

$$p(\mathbf{w}|t_{\tau_m:m}, \boldsymbol{\omega}, \Pi) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \boldsymbol{\Sigma}). \tag{5d}$$

In Eq. (5c), $N_m = m - \tau_m + 1$, $R = |\Pi|$ is the number of points on the marked Poisson process, and $T_m = t_m - t_{\tau_m}$. In Eq. (5d), $\boldsymbol{\Sigma} = [\boldsymbol{\Phi}\mathbf{D}\boldsymbol{\Phi}^\top + \mathbf{K}^{-1}]^{-1}$ where $\mathbf{D}$ is a diagonal matrix with $\{\omega_i\}_{i=\tau_m}^{m}$ in the first $m - \tau_m + 1$ entries and $\{\omega_r\}_{r=1}^{R}$ in the following R entries, and $\boldsymbol{\Phi} = [\{\boldsymbol{\Phi}(t_i)\}_{i=\tau_m}^{m}, \{\boldsymbol{\Phi}(t_r)\}_{r=1}^{R}]$; $\mathbf{m} = \boldsymbol{\Sigma}\boldsymbol{\Phi}\mathbf{v}$, where the first $m - \tau_m + 1$ entries of $\mathbf{v}$ are $1/2$, and the following R entries are $-1/2$. Through iterative sampling using Eq. (5), we obtain a series of samples from the model parameter posterior. The pseudocode is provided in Appendix B.1.

3.3.5 Algorithm, Hyperparameters and Complexity

By employing the proposed Gibbs sampler for analytical posterior sampling of model parameters and subsequently using the thinning algorithm for prediction, we establish our two-step CoBay-CPD tailored for Hawkes process. The detailed procedure is outlined in Appendix B.2.

CoBay-CPD's hyperparameters, including covariance $\mathbf{K}$ in the Gaussian prior, confidence intervals, and basis functions, impact its performance. In experiments, we assume $\mathbf{K} = \sigma^2\mathbf{I}$. Oversized σ^2 weakens the prior, causing unstable parameter estimation and oversensitive change point detection, while undersized values result in sluggish detection. Similarly, narrow confidence intervals oversensitize, and wide intervals slow detection. Balancing accuracy and efficiency, more basis functions enhance prediction accuracy but challenge computational efficiency. In experiments, we select all hyperparameters through cross-validation.

Assuming the length of the entire sequence is N, the average length of $t_{\tau_m:m}$ is M, the average length of the latent marked Poisson process is R, the average number of points within the interval of T_ϕ is N_ϕ, and the number of Gibbs iterations is L, the computation complexity of CoBay-CPD is $\mathcal{O}(N(MN_\phi B + LRN_\phi B + LC_{\text{TH}} + L(M + R)(B + 1)^2 + L(B + 1)^3))$, where C_{TH} represents the complexity of the thinning algorithm. The detailed analysis is provided in Appendix C.

4 Experiments

We evaluate the performance of CoBay-CPD on both synthetic and real-world datasets. For the synthetic data, our aim is to validate the capability of CoBay-CPD in accurately recovering the ground-truth parameters and change points. For the real-world data, we compare CoBay-CPD against several baseline methods to determine whether our approach exhibits superior CPD performance.

4.1 Baselines

We conduct a comparison between CoBay-CPD and several Bayesian change point detection (BCPD) methods which are designed to address the non-conjugate challenge for Hawkes process: (1) **SMCPD** [6] combines BCPD and sequential Monte Carlo (SMC). Similar to our approach, it is a sampling-based method to address the non-conjugate inference in the BCPD framework. (2) **SVCPD** [4] similarly combines BCPD and Stein variational inference to address the non-conjugate inference in the BCPD framework. Differently, this method infers the posterior of model parameters by the variational technique. (3) **SVCPD+Inhibition** is an extension of SVCPD that incorporates a nonlinear Hawkes process with inhibitory effects. This baseline is designed because the original SVCPD only considered a linear Hawkes process.

4.2 Metrics

We use four metrics to assess the performance of all methods. (1) **False Negative Rate (FNR)** quantifies the probability of a change point being incorrectly identified as not a change point,

Table 1: The FNR, FPR, MSE and RT of CoBay-CPD and other baselines on the synthetic dataset.

Model	FNR(↓)	FPR(% ↓)	MSE(↓)	RT(minute ↓)
SMCPD	0.38 ± 0.41	0.76 ± 0.26	0.07 ± 0.01	5.50 ± 0.31
SVCPD	0.50 ± 0.35	0.76 ± 0.26	0.06 ± 0.00	7.78 ± 0.01
SVCPD+Inhi	0.33 ± 0.24	0.60 ± 0.00	0.16 ± 0.01	23.09 ± 0.60
CoBay-CPD	$\mathbf{0.13 \pm 0.22}$	$\mathbf{0.46 \pm 0.26}$	$\mathbf{0.05 \pm 0.00}$	$\mathbf{4.62 \pm 0.10}$

calculated as $1 - \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$. False negative cases are critical in many applications, as they indicate that certain crucial changes fail to receive attention. (2) **False Positive Rate (FPR)** quantifies the probability of a stable point being incorrectly identified as a change point, calculated as $1 - \frac{\text{True Negatives}}{\text{False Positives+True Negatives}}$. Minimizing false positive cases is also important because frequent false alarms will waste resources. (3) **Mean Square Error (MSE)** measures the distance between the predicted next timestamp (the average of $t^{(k)}$) and the actual timestamp, calculated as $\frac{1}{n}\sum_{i=1}^{n}(\bar{t}_i^{(k)} - t_i)^2$. This metric evaluates how accurately the model predicts the next data point. (4) **Running Time (RT)** measures the efficiency of the method by its runtime.

4.3 Synthetic Data

We validate the efficacy of CoBay-CPD using a synthetic dataset. Our goal is to verify whether CoBay-CPD can accurately recover the ground-truth parameters and change points.

Datasets The synthetic data is created by concatenating three segments of Hawkes process data, each characterized by different parameters. Specifically, all three segments of Hawkes processes adhere to the model configuration outlined in Section 3.1, with their influence functions assumed to be a mixture of multiple beta densities. Three segments employ identical basis functions and mixing weights. However, they have different intensity upperbounds, namely $\bar{\lambda}_1 = 5$, $\bar{\lambda}_2 = 10$ and $\bar{\lambda}_3 = 3$. We utilize the thinning algorithm to simulate data for three Hawkes processes, and concatenate them to form the synthetic data. The two change points are located at the 43-rd and 136-th points, indicated by grey lines in Fig. 1a in Appendix D.2. Further details can be found in Appendix D.

Results We evaluate the performance of change point detection using CoBay-CPD and other baseline models on the synthetic dataset. For CoBay-CPD, we adopt a prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = 0.5\mathbf{I}$. The detection outcomes are presented in Fig. 1a in Appendix D.2. The change points identified by CoBay-CPD are $44, 136$, while SMCPD detects $96, 136$, SVCPD detects $44, 96$ and SVCPD+Inhibition detects $51, 136$. This discrepancy suggests that CoBay-CPD achieves more accurate change point detection. Furthermore, the estimated parameter $\bar{\lambda}$ from CoBay-CPD for the synthetic data is depicted in Fig. 1b in Appendix D.2. The estimated parameter $\bar{\lambda}$ closely aligns with the ground truth, demonstrating the accuracy of parameter estimation by CoBay-CPD. Notably, there are prominent changes around the change points in Fig. 1b. This phenomenon arises due to the initiation of a new Hawkes process with distinct parameters at the occurrence of a change point. The challenge of accurately estimating parameters with limited data in such scenarios is alleviated by the Bayesian framework. In similar situations, frequentist methods tend to perform poorly.

We also compare CoBay-CPD against baseline methods in terms of FNR, FPR, MSE and RT. The results are presented in Table 1. As anticipated, CoBay-CPD outperforms the alternatives. This superiority can be attributed to CoBay-CPD's utilization of a nonlinear Hawkes process model, which encompasses both excitation and inhibition effects. In contrast, SMCPD and SVCPD employ a simpler linear Hawkes process model, constraining their expressive power. Additionally, CoBay-CPD employs Gibbs sampler to accurately characterize the parameter posterior, whereas both SVCPD and SVCPD+Inhibition utilize variational-based methods to approximate the parameter posterior. As a result, their change point detection accuracy is compromised.

4.4 Real-world Data

In this section, we conduct a comparison between CoBay-CPD and baselines on two real datasets.

Table 2: The FNR, FPR, MSE and RT of CoBay-CPD and other baselines on real-world datasets.

Model	WannaCry				NYC Vehicle Collisions			
	FNR(↓)	FPR(↓)	MSE($\times 10^2$ ↓)	RT(minute ↓)	FNR(↓)	FPR(% ↓)	MSE(↓)	RT(minute ↓)
SMCPD	0.38 ± 0.06	0.02 ± 0.01	3.59 ± 0.08	11.65 ± 0.07	0.56 ± 0.16	2.46 ± 0.55	0.02 ± 0.00	24.67 ± 0.26
SVCPD	0.34 ± 0.12	0.01 ± 0.01	3.47 ± 0.06	9.72 ± 0.06	0.58 ± 0.36	1.00 ± 0.43	0.02 ± 0.00	19.30 ± 0.09
SVCPD+Inhi	0.54 ± 0.09	$\mathbf{0.00 \pm 0.00}$	3.54 ± 0.06	29.76 ± 2.54	0.22 ± 0.16	1.55 ± 0.36	0.17 ± 0.01	64.47 ± 1.36
CoBay-CPD	$\mathbf{0.21 \pm 0.04}$	0.05 ± 0.02	$\mathbf{3.42 \pm 0.00}$	$\mathbf{6.24 \pm 0.49}$	$\mathbf{0.13 \pm 0.16}$	$\mathbf{0.89 \pm 0.16}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{8.70 \pm 0.26}$

Table 3: Ablation study. The FNR, FPR, MSE and RT of CoBay-CPD with different hyperparameters.

Metric	Number of Basis Functions			Confidence Interval			Prior Covariance		
	1	2	3	95%	90%	85%	$\sigma^2 = 0.01$	$\sigma^2 = 0.5$	$\sigma^2 = 10$
FNR(↓)	0.38 ± 0.41	0.38 ± 0.22	$\mathbf{0.13 \pm 0.22}$	0.50 ± 0.00	$\mathbf{0.13 \pm 0.22}$	0.25 ± 0.25	$\mathbf{0.13 \pm 0.22}$	$\mathbf{0.13 \pm 0.22}$	0.50 ± 0.00
FPR(% ↓)	1.07 ± 0.50	0.91 ± 0.30	$\mathbf{0.61 \pm 0.00}$	$\mathbf{0.46 \pm 0.26}$	$\mathbf{0.46 \pm 0.26}$	1.83 ± 0.43	0.76 ± 0.26	$\mathbf{0.46 \pm 0.26}$	0.91 ± 0.30
MSE(↓)	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	0.05 ± 0.00	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	0.05 ± 0.00	0.05 ± 0.01
RT(minute ↓)	$\mathbf{1.57 \pm 0.03}$	2.61 ± 0.08	3.62 ± 0.10	5.03 ± 0.02	4.62 ± 0.10	$\mathbf{4.50 \pm 0.11}$	4.74 ± 0.02	4.62 ± 0.10	$\mathbf{4.41 \pm 0.10}$

Datasets We analyze two datasets from the domains of network security and transportation, with specific details provided below. More comprehensive information regarding data preprocessing can be found in Appendix E. (1) **WannaCry Cyber Attack**[3] [4]: The WannaCry virus infected more than 200,000 computers around the world in 2017 and received much attention. The WannaCry Cyber Attack data contains 208 traffic logs information observations. Each observation contains the relevant timestamp. (2) **NYC Vehicle Collisions**[4] [27]: The New York City vehicle collision dataset comprises approximately 1.05 million vehicle collision records, each containing information about the time and location of the collision. For our experiments, we select the records from Oct.14th, 2017.

The real-world data, unlike synthetic data, does not have ground-truth change points. Therefore, we use the points where timestamps surge as the ground-truth change points in the WannaCry dataset, and utilize the reported change points from [27] as the ground-truth change points in the NYC dataset.

Results Figures 3a to 3d in Appendix E.2 display the change point detection outcomes of different methods applied to WannaCry data. It is clear that CoBay-CPD exhibits the most favorable detection performance. The change points identified by our method are consistent with the actual change points. The SMCPD, SVCPD and SVCPD+Inhibition detect a relatively limited number of change points, resulting in missed change points. Table 2 presents various metrics of four methods for change point detection in the WannaCry data. Clearly, because SMCPD, SVCPD, and SVCPD+Inhibition detect too few change points, their FNR is high and FPR is low. In contrast, CoBay-CPD exhibits the lowest FNR, a reasonably balanced FPR, the smallest MSE, and requires the least runtime.

Figures 4a to 4d in Appendix E.2 show the change point detection outcomes of four methods for the NYC data. Notably, SVCPD detects fewer change points, while SMCPD identify an excessive number. The change points detected by CoBay-CPD are 43, 110, 160, 194, 284, 338, 398, corresponding to the times 2:30, 9:00, 12:00, 13:10, 16:00, 17:55, 20:00. These timestamps coincide with peak traffic hours on workdays. Table 2 presents various metrics of four methods for change point detection in the NYC data. Consistently, SMCPD exhibits high FNR and FPR. The high FPR is due to an excessive number of change points detected by SMCPD, while the high FNR is due to the inaccurate detection of numerous change points by SMCPD. Whereas SVCPD shows a high FNR and low FPR due to detecting too few change points. SVCPD+Inhibition achieves a relatively balanced FNR and FPR, indicating the beneficial impact of employing a nonlinear Hawkes process. CoBay-CPD demonstrates superior accuracy and efficiency, with the lowest FNR, FPR, MSE, and RT in change point detection compared to all baseline models.

4.5 Ablation Study

In this section, we conduct hyperparameter analysis and stress tests of CoBay-CPD on synthetic data.

[3] https://www.malware-traffic-analysis.net/2017/05/18/index2.html

[4] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Table 4: The results of stress tests. We conduct experiments to verify the performance change of CoBay-CPD w.r.t. the number of change points, the difference between adjacent $\bar{\lambda}$'s, and the closeness between two change points.

Metric	Number of Change Points			$\Delta\bar{\lambda}$			Δt		
	1	2	3	0.1	1	5	5	10	15
FNR(↓)	**0.00 ± 0.00**	0.13 ± 0.22	0.11 ± 0.14	1.00 ± 0.00	0.25 ± 0.43	**0.00 ± 0.00**	0.33 ± 0.24	**0.00 ± 0.00**	**0.00 ± 0.00**
FPR(% ↓)	0.43 ± 0.60	0.46 ± 0.26	**0.31 ± 0.50**	1.61 ± 0.57	**0.35 ± 0.60**	0.43 ± 0.60	1.00 ± 0.70	0.93 ± 0.65	0.31 ± 0.54
MSE(↓)	**0.04 ± 0.00**	0.05 ± 0.00	0.07 ± 0.01	**0.02 ± 0.00**	0.03 ± 0.00	0.04 ± 0.00	**0.01 ± 0.00**	0.02 ± 0.00	0.02 ± 0.00

Number of Basis Functions The number of basis functions impacts the expressiveness of the Hawkes process, influencing the model's detection performance. We assess the model's detection performance across varying numbers of basis functions, from 1 to 3, as shown in the Table 3. In experiment, we set the other hyperparameters as: 90% confidence interval and $\mathbf{K} = 0.5\mathbf{I}$. Observably, as the number of basis functions increases, FNR and FPR decrease, indicating enhanced detection accuracy, while RT increases, indicating increased computational burden.

Confidence Interval We try 3 different confidence intervals: 95%, 90%, and 85% for the next timestamp, as shown in Table 3. In the experiment, we choose 4 basis functions and $\mathbf{K} = 0.5\mathbf{I}$. A wider confidence interval results in fewer detected change points, leading to a larger FNR and a smaller FPR. Conversely, a narrower confidence interval leads to the detection of more change points, resulting in numerous incorrect change points. Consequently, the FPR increases significantly, while the FNR also shows a slight rise. So the compromise, 90% confidence intervals, is the best.

Prior Covariance The Gaussian prior covariance $\mathbf{K} = \sigma^2\mathbf{I}$ also has a large impact on the detection results. In this experiment, we choose 4 basis functions and 90% confidence interval. If σ^2 is too large, FNR and FPR will increase. This is because the prior is too loose, causing the posterior samples of model parameters to spread excessively and fail to concentrate around the true values. On the contrary, when σ^2 is too small, the posterior samples of model parameters are too concentrated in a certain position that may be a wrong value, resulting in a larger FPR, as shown in Table 3.

Stress Tests The stress tests assess how well a model performs under difficult or extreme conditions. We conduct three stress tests experiments: one involving the number of change points (more indicating greater difficulty), another focusing on the difference between adjacent $\bar{\lambda}$'s, $\Delta\bar{\lambda}$ (smaller indicating greater difficulty), and the third examining the closeness between adjacent change points, Δt (smaller indicating greater difficulty). The results are shown in Table 4. (1) Experiments with different numbers of change points reveal consistent performance across varying numbers. (2) Regarding $\Delta\bar{\lambda}$, our model effectively detects change points even when $\Delta\bar{\lambda}$ is small (e.g., $\Delta\bar{\lambda} = 1$). However, excessively small $\Delta\bar{\lambda}$ values lead to decreased performance, as the parameters on both sides of the change point become too similar to distinguish. (3) Regarding Δt, the model maintains good even when two change points are close, although performance slightly declines. More experimental details can be found in Appendix F.

5 Limitations and Broader Impacts

Although our proposed CoBay-CPD method offers an efficient and accurate solution to the change point detection problem in Hawkes process, it still has some limitations. For instance, extending CoBay-CPD to multivariate Hawkes processes remains challenging because the current method requires change points to occur at specific event locations. However, in multivariate Hawkes processes, a change point in one dimension (an event location) does not necessarily correspond to event locations in other dimensions. This challenge needs to be addressed further in future research.

The introduction of CoBay-CPD for Hawkes process holds promise for both positive and negative social impacts. This method effectively addresses the non-conjugate inference challenge, improving the efficiency and accuracy of change point detection across various fields. However, as automated change point detection methods become more accurate and efficient, there is a risk of overreliance on these systems without proper validation or human oversight. This could lead to erroneous decisions or missed opportunities for critical interventions.

6 Conclusions

In summary, this work introduces a novel conjugate Bayesian two-step change point detection method for Hawkes process, which effectively addresses the non-conjugate inference challenge. Leveraging data augmentation, we transform the non-conjugate inference problem to a conditionally conjugate one, enabling the development of an analytical Gibbs sampler for efficient parameter posterior sampling. Our proposed approach surpasses existing methods, showcasing superior accuracy and efficiency in detecting change points. The contributions of this research hold great potential for advancing event-driven time series analysis and change point detection across various applications.

Acknowledgments and Disclosure of Funding

This work was supported by NSFC Projects (Nos. 62106121, 72171229), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China, the fundamental research funds for the central universities, and the research funds of Renmin University of China (24XNKJ13).

References

[1] Bacry, E., Mastromatteo, I., and Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[2] Bhaduri, M., Rangan, D., and Balaji, A. Change detection in non-stationary hawkes processes through sequential testing. In *ITM Web of Conferences*, volume 36, pp. 01005. EDP Sciences, 2021.

[3] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] Detommaso, G., Hoitzing, H., Cui, T., and Alamir, A. Stein variational online change-point detection with applications to hawkes processes and neural networks. *arXiv preprint arXiv:1901.07987*, 2019.

[5] Donner, C. and Opper, M. Efficient Bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19(1):2710–2743, 2018.

[6] Doucet, A., Johansen, A. M., et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[7] Gerhard, F., Deger, M., and Truccolo, W. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2):e1005390, 2017.

[8] Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[9] Kingman, J. F. C. *Poisson processes*, volume 3. Clarendon Press, 1992.

[10] Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pp. 914–922, 2017.

[11] Linderman, S. W. *Bayesian Methods for Discovering Structure in Neural Spike Trains*. PhD thesis, Harvard University, 2016.

[12] Linderman, S. W. and Adams, R. P. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

[13] Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[14] MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[15] Malem-Shinitski, N., Ojeda, C., and Opper, M. Flexible temporal point processes modeling with nonlinear hawkes processes with gaussian processes excitations and inhibitions. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1–31, 2021.

[16] Malem-Shinitski, N., Ojeda, C., and Opper, M. Variational bayesian inference for nonlinear hawkes process with gaussian process self-effects. *Entropy*, 24(3):356, 2022.

[17] Mei, H. and Eisner, J. The neural Hawkes process: A neurally self-modulating multivariate point process. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6754–6764, 2017.

[18] Neal, R. M. Probabilistic inference using markov chain monte carlo methods. 1993.

[19] Ogata, Y. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[20] Pinto, J. C. L., Chahed, T., and Altman, E. Trend detection in social networks using Hawkes processes. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 1441–1448. ACM, 2015.

[21] Rasmussen, J. G. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15:623–642, 2013.

[22] Sulem, D., Rivoirard, V., and Rousseau, J. Scalable variational bayes methods for hawkes processes. *arXiv preprint arXiv:2212.00293*, 2022.

[23] Wang, H., Xie, L., Xie, Y., Cuozzo, A., and Mak, S. Sequential change-point detection for mutually exciting point processes. *Technometrics*, 65(1):44–56, 2023.

[24] Zhang, R., Walder, C., and Rizoiu, M.-A. Variational inference for sparse Gaussian process modulated Hawkes process. *arXiv preprint arXiv:1905.10496v2*, 2019.

[25] Zhang, R., Walder, C. J., Rizoiu, M., and Xie, L. Efficient non-parametric Bayesian Hawkes processes. In *International Joint Conference on Artificial Intelligence*, pp. 4299–4305, 2019.

[26] Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. Efficient inference for non-parametric Hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21(241):1–31, 2020.

[27] Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. Fast multi-resolution segmentation for nonstationary Hawkes process using cumulants. *International Journal of Data Science and Analytics*, 10:321–330, 2020.

[28] Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. Efficient em-variational inference for nonparametric hawkes process. *Statistics and Computing*, 31(4):46, 2021.

[29] Zhou, F., Zhang, Y., and Zhu, J. Efficient inference of flexible interaction in spiking-neuron networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[30] Zhou, F., Kong, Q., Deng, Z., Kan, J., Zhang, Y., Feng, C., and Zhu, J. Efficient inference for dynamic flexible interactions of neural populations. *Journal of Machine Learning Research*, 23 (211):1–49, 2022.

A Derivation of CoBay-CPD

In this section we provide proof of data augmentation and Gibbs sampler, respectively.

A.1 Data Augmentation

Focus on the interval with change points $[t_{\tau_m}, t_m]$, the probability density (likelihood) of CoBay-CPD can be presented as:

$$p(\{t_i\}_{i=\tau_m}^m|\mathbf{w},\bar{\lambda}) = \prod_{i=\tau_m}^m \bar{\lambda}\sigma(h(t_i)) \exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right).$$

Substitute the augmentation of Pólya-Gamma Variables $\sigma(z) = \int_0^\infty e^{f(\omega,z)} p_{\text{PG}}(\omega|1,0)d\omega$ and the sigmoid symmetry property $\sigma(z) = 1 - \sigma(-z)$ to the above equation, we can obtain:

$$\exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right) = \exp\left(-\int_{t_{\tau_m}}^{t_m}\int_0^\infty (1 - e^{f(\omega,-h(t))})\bar{\lambda}p_{\text{PG}}(\omega|1,0)d\omega dt\right).$$

According to Campbell's theorem [9], the exponential integral term can be rewritten as

$$\exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right) = \mathbb{E}_{p_\lambda}\left[\prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}\right],$$

where $\Pi = (\omega_r, t_r)_{r=1}^R$ denotes a realization of a marked Poisson process on the interval $[t_{\tau_m}, t_m]$, with its probability measure denoted as p_λ and having an intensity $\lambda(\omega, t) = \bar{\lambda}p_{\text{PG}}(\omega|1,0)$.

Therefore, the likelihood of CoBay-CPD can be rewritten as :

$$\begin{aligned}
p(\{t_i\}_{i=\tau_m}^m|\mathbf{w},\bar{\lambda}) &= \prod_{i=\tau_m}^m \bar{\lambda}\sigma(h(t_i)) \exp\left(-\int_{t_{\tau_m}}^{t_m} \bar{\lambda}\sigma(h(t))dt\right) \\
&= \prod_{i=\tau_m}^m \left(\int_0^\infty \bar{\lambda}e^{f(\omega_i,h(t_i)}p_{\text{PG}}(\omega_i|1,0)d\omega_i\right)\mathbb{E}_{p_\lambda}\left[\prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}\right] \\
&= \int\int \prod_{i=\tau_m}^m [\lambda(t_i,\omega_i)e^{f(\omega_i,h(t_i))}]p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))} d\boldsymbol{\omega} d\Pi,
\end{aligned}$$

where $\boldsymbol{\omega}$ is the vector of ω_i. It is straightforward to see the integrand is the augmented likelihood:

$$p(\{t_i\}_{i=\tau_m}^m, \boldsymbol{\omega}, \Pi|\mathbf{w},\bar{\lambda}) = \prod_{i=\tau_m}^m [\lambda(t_i,\omega_i)e^{f(\omega_i,h(t_i))}]p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}.$$

A.2 Gibbs Sampler

Based on the augmented joint distribution

$$p(\{t_i\}_{i=\tau_m}^m, \boldsymbol{\omega}, \Pi, \mathbf{w}, \bar{\lambda}|\tau_m) = p(\{t_i\}_{i=\tau_m}^m, \boldsymbol{\omega}, \Pi|\mathbf{w}, \bar{\lambda}, \tau_m)p(\mathbf{w})p(\bar{\lambda}),$$

we can derive the conditional densities of all variables in closed form. By sampling from these conditional densities iteratively, we construct an analytical Gibbs sampler.

A.2.1 Derivation for ω

$$p(\boldsymbol{\omega}|\{t_i\}_{i=\tau_m}^m, \mathbf{w}) \propto \prod_{i=\tau_m}^m \left[\lambda(t_i,\omega_i)e^{f(\omega_i,h(t_i))}\right] \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))},$$

where $\prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}$ is constant as Π is given. In addition, combined $f(\omega,z) = z/2 - z^2\omega/2 - \log 2$ and $p_{\text{PG}}(\omega|b,0)\cdot e^{-\frac{c^2\omega}{2}} \propto p_{\text{PG}}(\omega|b,c)$ with the above derivation, we can deduce that,

$$p(\boldsymbol{\omega}|\{t_i\}_{i=\tau_m}^m, \mathbf{w}) \propto \prod_{i=\tau_m}^m \left[\bar{\lambda} p_{\text{PG}}(\omega_i|1,0) e^{h(t_i)/2 - h(t_i)^2\omega/2 - \log 2}\right] \propto \prod_{i=\tau_m}^m p_{\text{PG}}(\omega_i|1,h(t_i)).$$

A.2.2 Derivation for Π

The posterior of Π is dependent on $\{t_i\}_{i=\tau_m}^m$, $\mathbf{w}$ and $\bar{\lambda}$,

$$p(\Pi|\{t_i\}_{i=\tau_m}^m, \mathbf{w}, \bar{\lambda}) = \frac{p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}}{\int p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))} d\Pi},$$

where Campbell's theorem can be applied to convert the denominator, the equation above can be transformed as

$$\begin{aligned} p(\Pi|\{t_i\}_{i=\tau_m}^m, \mathbf{w}, \bar{\lambda}) &= \frac{p_\lambda(\Pi|\bar{\lambda}) \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))}}{\exp\left(-\int_{t_{\tau_m}}^{t_m}\int_0^\infty (1 - e^{f(\omega,-h(t))})\bar{\lambda} p_{\text{PG}}(\omega|1,0) d\omega dt)\right)} \\ &= \prod_{(\omega,t)\in\Pi} \left(e^{f(\omega,-h(t))})\bar{\lambda} p_{\text{PG}}(\omega|1,0)\right) \exp\left(-\int_{t_{\tau_m}}^{t_m}\int_0^\infty e^{f(\omega,-h(t))}\bar{\lambda} p_{\text{PG}}(\omega|1,0) d\omega dt)\right). \end{aligned}$$

The above posterior is in the likelihood form of a marked Poisson process with intensity

$$\Lambda(t,\omega|\{t_i\}_{i=\tau_m}^m, \mathbf{w}, \bar{\lambda}) = e^{f(\omega,-h(t))})\bar{\lambda} p_{\text{PG}}(\omega|1,0) = \bar{\lambda}\sigma(-h(t)) p_{\text{PG}}(\omega|1,h(t)).$$

A.2.3 Derivation for $\bar{\lambda}$

$$\begin{aligned} p(\bar{\lambda}|\{t_i\}_{i=\tau_m}^m, \Pi) &\propto \prod_{i=\tau_m}^m [\lambda(t_i,\omega_i) e^{f(\omega_i,h(t_i))})] p_\lambda(\Pi|\bar{\lambda})\cdot 1/\bar{\lambda} \\ &\propto \prod_{i=\tau_m}^m [\bar{\lambda} p_{\text{PG}}(\omega|1,h(t_i))] p_\lambda(\Pi|\bar{\lambda})\cdot 1/\bar{\lambda}, \end{aligned}$$

where $p_{\text{PG}}(\omega|1,h(t)) = \prod_{(\omega,t)\in\Pi} \bar{\lambda} p_{\text{PG}}(\omega|1,0) e^{-\int_{t_{\tau_m}}^{t_m}\int_0^\infty \bar{\lambda} p_{\text{PG}}(\omega|1,0) d\omega dt}$. So the above equation are transformed as

$$p(\bar{\lambda}|\{t_i\}_{i=\tau_m}^m, \Pi) \propto \bar{\lambda}^{(m-\tau_m+1+|\Pi|-1)} e^{-\int_{t_{\tau_m}}^{t_m} dt\cdot\bar{\lambda}}.$$

Let $N_m = m - \tau_m + 1$, $R = |\Pi|$ is the number of points on the marked Poisson process, and $T_m = t_m - t_{\tau_m}$, then we can get

$$p(\bar{\lambda}|\{t_i\}_{i=\tau_m}^m, \Pi) = p_{\text{Ga}}(\bar{\lambda}|N_m + R, T_m).$$

A.2.4 Derivation for $\mathbf{w}$

For $\mathbf{w}$, we utilize the Gaussian prior $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0},\mathbf{K})$ where $\mathbf{K}$ is the prior covariance matrix. Then the derivation for $\mathbf{w}$ is as follows

$$p(\mathbf{w}|\{t_i\}_{i=\tau_m}^m, \boldsymbol{\omega}, \Pi) \propto \prod_{b=1}^{B+1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w_b-\mu)^2}{2\sigma^2}} \prod_{i=\tau_m}^m e^{f(\omega_i,h(t_i))} \prod_{(\omega,t)\in\Pi} e^{f(\omega,-h(t))},$$

where $h(t) = \mu + \sum_{b=1}^B w_b \sum_{t_i<t} \tilde{\phi}_b(t-t_i)$ and $f(\omega,z) = z/2 - z^2\omega/2 - \log 2$ and $p_{\text{PG}}(\omega|b,0)\cdot e^{-\frac{c^2\omega}{2}} \propto p_{\text{PG}}(\omega|b,c)$. Therefore, we obtain

$$p(\mathbf{w}|\{t_i\}_{i=\tau_m}^m, \boldsymbol{\omega}, \Pi) = N(\mathbf{w}|\mathbf{m},\boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = [\boldsymbol{\Phi}\mathbf{D}\boldsymbol{\Phi}^\top + \mathbf{K}^{-1}]^{-1}$, where $\mathbf{D}$ is a diagonal matrix with $\{\omega_i\}_{i=\tau_m}^m$ in the first $m-\tau_m+1$ entries and $\{\omega_r\}_{r=1}^R$ in the following R entries, and $\boldsymbol{\Phi} = [\{\boldsymbol{\Phi}(t_i)\}_{i=\tau_m}^m, \{\boldsymbol{\Phi}(t_r)\}_{r=1}^R]$; $\mathbf{m} = \boldsymbol{\Sigma}\boldsymbol{\Phi}\mathbf{v}$, where the first $m-\tau_m+1$ entries of $\mathbf{v}$ are $1/2$, and the following R entries of $\mathbf{v}$ are $-1/2$.

B Algorithm

B.1 Gibbs Sampler

Algorithm 1 Gibbs Sampler

Input: $t_{\tau_m:m}$, basis functions $\{\widetilde{\phi}_b(\cdot)\}_{b=1}^B$, covariance $\mathbf{K}$;
Output: Parameter posterior samples $\{\mathbf{w}^{(k)}, \bar{\lambda}^{(k)}\}_{k=1}^K$;

1: **for** iteration **do**
2: Sample $\boldsymbol{\omega}$ by Eq. (5a);
3: Sample Π through thinning by Eq. (5b);
4: Sample $\bar{\lambda}$ by Eq. (5c);
5: Sample $\mathbf{w}$ by Eq. (5d).
6: **end for**

B.2 CoBay-CPD

Algorithm 2 $(m+1)$-th round of CoBay-CPD

Input: $t_{\tau_m:m+1}$, basis functions $\{\widetilde{\phi}_b(\cdot)\}_{b=1}^B$, covariance $\mathbf{K}$;
Output: Change point at the current position or not;

1: Sample $\{\mathbf{w}^{(k)}, \bar{\lambda}^{(k)}\}_{k=1}^K$ by Appendix B.1;
2: Sample $\{t_{m+1}^{(k)} \sim p(t_{m+1}|t_{\tau_m:m}, \mathbf{w}^{(k)}, \bar{\lambda}^{(k)})\}_{k=1}^K$ and create a confidence interval of $\{t_{m+1}^{(k)}\}$, e.g., the 5% and 95% quantiles denoted as t_{m+1}^l and t_{m+1}^r. If the actual t_{m+1} falls within the interval $[t_{m+1}^l, t_{m+1}^r]$, classify it as not a change point; otherwise, a change point.

C Analysis of Complexity

Assuming the length of the entire sequence is N, the average length of $t_{\tau_m:m}$ is M, the average length of the latent marked Poisson process is R, the average number of points within the interval of T_ϕ is N_ϕ, and the number of Gibbs iterations is L, the computation complexity of CoBay-CPD is $\mathcal{O}(N(MN_\phi B + LRN_\phi B + LC_{\text{TH}} + L(M+R)(B+1)^2 + L(B+1)^3))$, where C_{TH} represents the complexity of the thinning algorithm in marked poisson process, not in the prediction step. This is because thinning algorithm in prediction step only samples one point at a time, which is very fast, so its computational complexity can be ignored. We ignore the complexities of other sampling operations since they are fast. The first term corresponds to the precomputation of $\mathbf{\Phi}(t)$ on $t_{\tau_m:m}$, the second term to the computation of $\mathbf{\Phi}(t)$ on Π, the third term to the sampling of the marked Poisson process, the fourth and fifth terms to the computation of mean and covariance. By limiting the maximum length of $t_{\tau_m:m}$ and T_ϕ, we can reduce the values of M, N_ϕ and R, thereby accelerating the computation of $\mathbf{\Phi}(t)$. Moreover, as the length of $t_{\tau_m:m}$ decreases, C_{TH} will decrease as well.

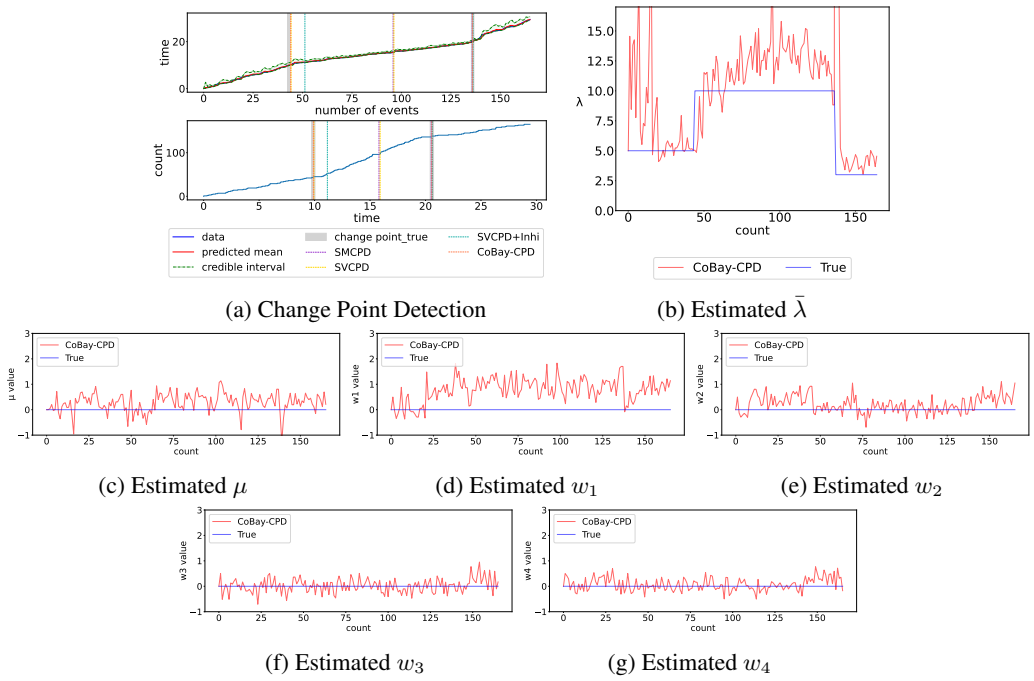
D Synthetic Data Experiment

D.1 Data Processing

We generate a synthetic data concatenated by three segments of Hawkes process data. In these three segments of Hawkes process data, we assume 4 scaled beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 50, \tilde{\beta} = 50, \text{scale} = 6, \text{shift} = \{-2, -1, 0, 1\})$ as the basis functions with support $[0, T_\phi = 6]$ and $\mu = 0$ as the baseline activation. However, they have different intensity upperbounds, namely $\bar{\lambda}_1 = 5$, $\bar{\lambda}_2 = 10$ and $\bar{\lambda}_3 = 3$. We use the thinning algorithm to generate a sequence according to the intensity function specified above.

D.2 Result Presentation

In experiment, we assume the basis functions are same as the ground truth, and set the other hyperparameters as 90% confidence interval and $\mathbf{K} = 0.5\mathbf{I}$. The estimation of $\bar{\lambda}$ and the estimated $\mathbf{w} = [\mu, w_1, \ldots, w_4]^\top$ from CoBay-CPD are shown in Fig. 1. We can see that, the estimated parameter $\bar{\lambda}$ and $\mathbf{w} = [\mu, w_1, \ldots, w_4]^\top$ of synthetic data from CoBay-CPD closely oscillates around the true value, indicating the accuracy of parameter estimation by CoBay-CPD.



(a) Change Point Detection (b) Estimated $\bar{\lambda}$

(c) Estimated μ (d) Estimated w_1 (e) Estimated w_2

(f) Estimated w_3 (g) Estimated w_4

Figure 1: Synthetic data: (a) the change point detection results of CoBay-CPD and alternatives, illustrating the change point detection performance; (b) the estimated $\bar{\lambda}$ from CoBay-CPD, indicating the accuracy of parameter estimation of CoBay-CPD;(c)-(g) the estimated parameter (a) μ, (b) w_1, (c) w_2, (d) w_3 and (e) w_4 of synthetic data from CoBay-CPD.

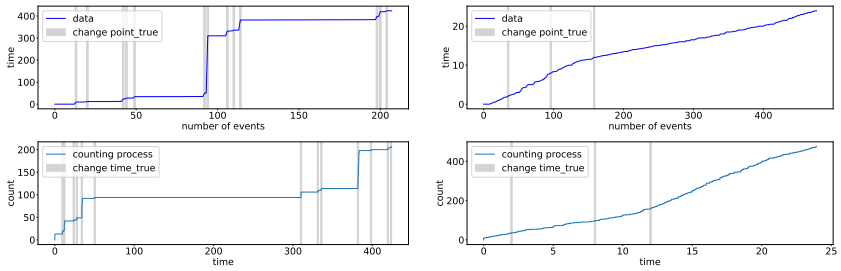
E Real-world Data Experiment

E.1 Data Processing

The preprocessing details of two real-world datasets are shown below.

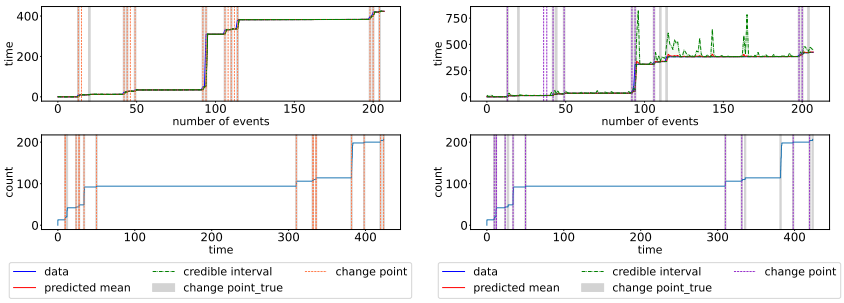
WannaCry Cyber Attack In May 2017, the WannaCry virus infected more than 200,000 computers worldwide, causing at least hundreds of millions of dollars in damage, and received much attention. The WannaCry Cyber Attack data contains 208 traffic logs information observations. Each observation contains the relevant timestamp. In this paper, the points where timestamps surge are taken as the ground truth change points, shown in Fig. 2a.

NYC Vehicle Collisions The New York City vehicle collision dataset comprises approximately 1.05 million vehicle collision records, each containing information about the time and location of the collision. For our experiments, we select the records from October 14th, 2017, which contains 477 vehicle collision records. We utilize the change points detected in [27] as the ground truth, shown in Fig. 2b.



(a) WannaCry Data (b) NYC Vehicle Collisions Data

Figure 2: (a) The WannaCry data with ground-truth change points (grey lines). (b) The NYC Vehicle Collisions data with ground-truth change points (grey lines). The upper plot illustrates the increasing of timestamps as events accumulate. The lower plot reverses the axes, representing a counting process.



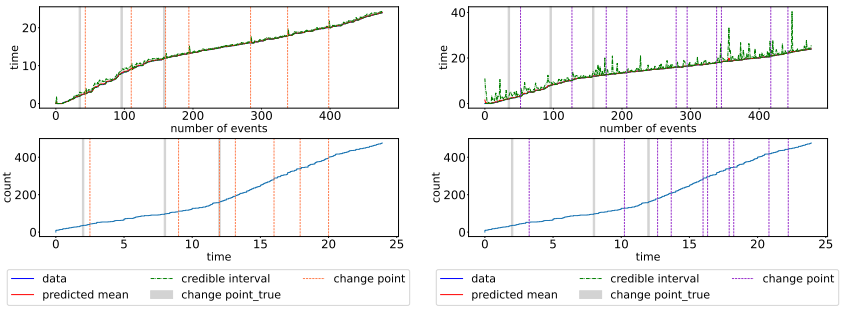
(a) CoBay-CPD (b) SMCPD

(c) SVCPD (d) SVCPD+Inhibition

Figure 3: The WannaCry data. The upper plot illustrates the increasing of timestamps as events accumulate. The lower plot reverses the axes, representing a counting process. The change point detection result of (a) CoBay-CPD, (b) SMCPD, (c) SVCPD, (d) SVCPD+Inhibiton.

E.2 Results Presentation

For WannaCry, we adopt a prior distribution $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \mathbf{K})$ for CoBay-CPD, where $\mathbf{K}$ is a diagonal matrix with diagonal entries of 0.5. Moreover, we choose 90% confidence interval and 4 scaled shifted beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 50, \tilde{\beta} = 50, \text{scale} = 6, \text{shift} = \{-2, -1, 0, 1\})$ as basis functions. The complete graph of experimental results of four methods on WannaCry Cyber Attack Dataset is shown in Fig. 3. Figures 3a to 3d display the change point detection outcomes of different methods applied to WannaCry data. The blue line is the real data, the red solid line is the mean of the predicted points, the green dotted line is the confidence interval, and the orange, purple, yellow and turquoise line are the detected change point location.



(a) CoBay-CPD

(b) SMCPD

(c) SVCPD

(d) SVCPD+Inhibiton

Figure 4: The NYC Vehicle Collisions data. The upper plot illustrates the increasing of timestamps as events accumulate. The lower plot reverses the axes, representing a counting process. The change point detection result of (a) CoBay-CPD, (b) SMCPD, (c) SVCPD, (d) SVCPD+Inhibiton.

For NYC Vehicle Collisions, we choose 4 scaled shifted beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 10, \tilde{\beta} = 30, \text{scale} = 6, \text{shift} = \{-2, -1, 0, 1\})$ as basis functions, 90% confidence interval and $\mathbf{K} = 0.5\mathbf{I}$, which is same as that in WannaCry. The change points in the NYC Vehicle Collisions are not as obvious as those in the WannaCry, making change point detection a more challenging task for this dataset. The complete graph of experimental results of four methods on NYC Vehicle Collisions Dataset is shown in Fig. 4. The blue line is the real data, the red solid line is the mean of the predicted points, the green dotted line is the confidence interval, and the orange, purple, yellow and turquoise line are the detected change point location. Figures 4a to 4d show the change point detection outcomes of four methods for the NYC data. Notably, SVCPD detects fewer change points, while SMCPD identify an excessive number.

F Stress Tests

F.1 Test 1: Number of Change Points

We conduct a stress test with the number of change points: $1, 2, 3$. We generate three sets of synthetic data concatenated by some segments of Hawkes process data. In all these segments of Hawkes process data, we assume 4 scaled beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 50, \tilde{\beta} = 50, \text{scale} = 6, \text{shift} = \{-2, -1, 0, 1\})$ as the basis functions with support $[0, T_\phi = 6]$ and $\mu = 0$ as the baseline activation. However, they have different intensity upperbounds. For $\#$ of change points $= 1$, we let $\bar{\lambda}_{11} = 5$, $\bar{\lambda}_{12} = 10$; For $\#$ of change points $= 2$, we let $\bar{\lambda}_{21} = 5$, $\bar{\lambda}_{22} = 10$, and $\bar{\lambda}_{23} = 3$; For $\#$ of change points $= 3$, we let $\bar{\lambda}_{31} = 5$, $\bar{\lambda}_{32} = 10$, $\bar{\lambda}_{33} = 3$, and $\bar{\lambda}_{34} = 8$. We use the thinning algorithm to generate these sequences according to the intensity specified above.

F.2 Test 2: Difference between Adjacent Parameters

We conduct a stress test with $\Delta\bar{\lambda} = 0.1, 1, 5$. We generate three sets of synthetic data by concatenating two segments of Hawkes process data. Within each segment, we assume four scaled beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 50, \tilde{\beta} = 50, \text{scale} = 6, \text{shift} = -2, -1, 0, 1)$ as the basis functions with support $[0, T_{\phi} = 6]$, and $\mu = 0$ as the baseline activation. However, they possess different intensity upper bounds. Specifically, for $\Delta\bar{\lambda} = 0.1$, we set $\bar{\lambda}_{11} = 10$ and $\bar{\lambda}_{12} = 10.1$; for $\Delta\bar{\lambda} = 1$, we set $\bar{\lambda}_{21} = 10$ and $\bar{\lambda}_{22} = 9$; for $\Delta\bar{\lambda} = 5$, we set $\bar{\lambda}_{31} = 10$ and $\bar{\lambda}_{32} = 5$. We use the thinning algorithm to generate these sequences according to the intensity specified above.

F.3 Test 3: Closeness between Adjacent Change Points

We conduct a stress test with $\Delta_t = 5, 10, 15$. Three sets of synthetic data are generated by concatenating three segments of Hawkes process data. Within each segment, we assume four scaled beta densities: $\tilde{\phi}_{1,2,3,4} = \text{Beta}(\tilde{\alpha} = 50, \tilde{\beta} = 50, \text{scale} = 6, \text{shift} = -2, -1, 0, 1)$ as the basis functions with support $[0, T_{\phi} = 6]$, and $\mu = 0$ as the baseline activation. They have different intensity upper-bounds $\bar{\lambda}_1 = 10$, $\bar{\lambda}_2 = 5$, and $\bar{\lambda}_3 = 15$. We use the thinning algorithm to generate these sequences according to the intensity specified above. We adjust the data length of the second segment from 5 to 10 to 15, thereby controlling the interval Δ_t between two adjacent change points from 5 to 10 to 15.

F.4 Stree Tests Supplements

Due to the page limit, we only presented the stress test results for our own method in the main paper. However, based on a suggestion from an anonymous reviewer to include the stress test results for the baselines, we have provided them in Tables 5 to 7.

Table 5: The FNR, FPR and MSE of CoBay-CPD and other baselines on synthetic dataset with different number of change points.

Model	1			2			3		
	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)
SMCPD	0.33 ± 0.47	0.63 ± 0.63	0.08 ± 0.03	0.38 ± 0.41	0.76 ± 0.26	0.07 ± 0.01	0.67 ± 0.19	1.23 ± 0.35	0.08 ± 0.01
SVCPD	0.67 ± 0.47	0.63 ± 0.95	0.06 ± 0.01	0.50 ± 0.35	0.76 ± 0.26	0.06 ± 0.00	0.50 ± 0.32	1.74 ± 0.65	0.15 ± 0.05
SVCPD+Inhi	0.33 ± 0.47	1.88 ± 0.63	0.08 ± 0.01	0.33 ± 0.24	0.60 ± 0.00	0.16 ± 0.01	0.28 ± 0.23	1.84 ± 0.50	0.09 ± 0.00
CoBay-CPD	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.43 \pm 0.60}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.13 \pm 0.22}$	$\mathbf{0.46 \pm 0.26}$	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.11 \pm 0.14}$	$\mathbf{0.31 \pm 0.50}$	$\mathbf{0.07 \pm 0.01}$

Table 6: The FNR, FPR and MSE of CoBay-CPD and other baselines on synthetic dataset with different difference between adjacent $\bar{\lambda}$'s ($\Delta\bar{\lambda}$).

Model	0.1			1			5		
	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)
SMCPD	1.00 ± 0.00	$\mathbf{1.20 \pm 0.00}$	0.06 ± 0.01	0.50 ± 0.50	0.70 ± 0.70	0.06 ± 0.02	0.33 ± 0.47	0.63 ± 0.63	0.08 ± 0.03
SVCPD	1.00 ± 0.00	2.41 ± 0.98	0.05 ± 0.01	0.83 ± 0.37	3.29 ± 1.05	0.06 ± 0.01	0.67 ± 0.47	0.63 ± 0.95	0.06 ± 0.01
SVCPD+Inhi	$\mathbf{0.67 \pm 0.47}$	1.41 ± 0.83	0.06 ± 0.00	0.33 ± 0.47	1.17 ± 0.52	0.06 ± 0.00	0.33 ± 0.47	1.88 ± 0.63	0.08 ± 0.01
CoBay-CPD	1.00 ± 0.00	1.61 ± 0.57	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.25 \pm 0.43}$	$\mathbf{0.35 \pm 0.60}$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.43 \pm 0.60}$	$\mathbf{0.04 \pm 0.00}$

Table 7: The FNR, FPR and MSE of CoBay-CPD and other baselines on synthetic dataset with different closeness between two change points (Δt).

Model	5			10			15		
	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)	FNR(↓)	FPR(% ↓)	MSE(↓)
SMCPD	0.42 ± 0.34	$\mathbf{0.75 \pm 0.75}$	0.03 ± 0.01	0.67 ± 0.24	$\mathbf{0.23 \pm 0.52}$	0.05 ± 0.01	0.17 ± 0.24	1.00 ± 0.83	0.07 ± 0.01
SVCPD	0.42 ± 0.19	1.24 ± 0.56	0.03 ± 0.01	0.75 ± 0.25	0.46 ± 0.65	0.05 ± 0.01	$\mathbf{0.08 \pm 0.19}$	3.01 ± 1.15	0.06 ± 0.01
SVCPD+Inhi	0.58 ± 0.19	1.24 ± 1.33	0.05 ± 0.01	0.25 ± 0.38	$\mathbf{0.23 \pm 0.52}$	0.05 ± 0.00	0.17 ± 0.24	2.01 ± 1.33	0.06 ± 0.00
CoBay-CPD	$\mathbf{0.33 \pm 0.24}$	1.00 ± 0.70	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	0.93 ± 0.65	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.08 \pm 0.19}$	$\mathbf{0.80 \pm 0.57}$	$\mathbf{0.03 \pm 0.00}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. See the Abstract and Introduction sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation of the work is discussed in the Limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For all theoretical results, the paper provides the corresponding proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results in the paper. See the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the data and code in the supplemental material to reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details necessary to understand the results. See the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the statistical significance of the experiments. In the tables presenting the experimental results, we provide the standard deviations across multiple runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted on a normal laptop and we provide the time of execution needed to reproduce the experiments. See the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See the Broader Impacts section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All code, models, and datasets mentioned in the text are appropriately cited with their original papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper, such as code, are well documented. The documentation is provided alongside the assets in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.