Statistica Sinica Preprint No: SS- 2021-0221	
Title	Detecting Abrupt Changes in High-Dimensional Self-Exciting Poisson Processes
Manuscript ID	SS- 2021-0221
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0221
Complete List of Authors	Daren Wang, Yi Yu and Rebecca Willett
Corresponding Authors	Yi Yu
E-mails	yi.yu.2@warwick.ac.uk
Notice: Accepted version subject to English editing.	

Statistica Sinica

DETECTING ABRUPT CHANGES IN HIGH-DIMENSIONAL SELF-EXCITING POISSON PROCESSES

Daren Wang[‡], Yi Yu[†] and Rebecca Willett[*]

Department of ACMS, University of Notre Dame[‡]

Department of Statistics, University of Warwick[†]

Department of Statistics, University of Chicago[*]

Abstract: High-dimensional self-exciting point processes have been widely used in many application areas to model discrete event data in which past and current events affect the likelihood of future events. In this paper, we are concerned with detecting abrupt changes of the coefficient matrices in discrete-time high-dimensional self-exciting Poisson processes, which have yet to be studied in the existing literature due to both theoretical and computational challenges rooted in the non-stationary and high-dimensional nature of the underlying process. We propose a penalized dynamic programming approach which is supported by a theoretical rate analysis and numerical evidence.

Key words and phrases: Self-exciting Poisson process; High-dimensional statistics; Piecewise stationarity; Penalized dynamic programming.

1. Introduction

Self-exciting point processes (SEPPs) are useful in modelling many types of discrete event data in which past and current event help determine the likelihood of future events. Such data are common in spike trains recorded from biological networks [e.g. Brown et al., 2004, Pillow et al., 2008], interactions within a social network [e.g. Zhou et al., 2013, Hall and Willett, 2016], pricing changes within financial networks [e.g. Chavez-Demoulin and McGill, 2012, Aït-Sahalia et al., 2015], power failures in networked electrical systems [e.g. Ertekin et al., 2015], crime and military engagements [e.g. Stomakhin et al., 2011, Blundell et al., 2012] and a variety of other settings.

SEPPs were, arguably, first rigorously studied in a mathematical framework by Hawkes [1971], where the eponymous Hawkes process was proposed. Since the debut of the Hawkes process, there have been tremendous efforts poured into different aspects of understanding and utilizing the univariate Hawkes process; see Laub et al. [2015] and Reinhart [2019] for comprehensive and contemporary reviews. More recently, due to the availability of richer datasets and computational resources, attention has shifted to multivariate and even high-dimensional SEPPs, where different coordinates might correspond to different geographic locations, different neurons in a biological neural network, people in a social network, etc. See, for instance, Hall et al. [2016], Mark et al. [2018], Chavez-Demoulin and McGill [2012] and Ertekin et al. [2015].

In these high-dimensional settings, understanding how events in one coordinate influence the likelihood of events in another coordinate provides valuable insight into the underlying process. We call the collection of these influences between pairs of coordinates a “network”, and this paper describes novel methods for detecting abrupt changes in this network with theoretical performance bounds that characterize the accuracy of the change point estimation and how strong the signals must be to ensure reliable estimation.

While change point detection has a long and rich history, we are unaware of any preexisting change point methodology that can be used to detect changes in SEPPs in high dimensions. Some recent high-dimensional change point detection work is briefly discussed as follows. Wang et al. [2018] and Padilla et al. [2019] studied the change point detection in Bernoulli networks and dynamic random dot product graphs, respectively. Cho and Fryzlewicz [2015], Cho [2016], Matteson and James [2014], Wang and Samworth [2018], Dette and Gösmann [2018] and others investigated high-dimensional mean change problems. Wang et al. [2017], Aue et al. [2009] and others were concerned with high/multi-dimensional covariance structure changes. Safikhani and Shojaie [2017] and Wang et al. [2019] exploited the high-dimensional vector autoregressive models and provided change point detection results thereof. Li et al. [2017] focused on a low-dimensional Hawkes process setting in which the processes may be characterized by a small number of parameters.

Given the abundant existing literature, we see a vacuum in the research on high-dimensional integer valued time series change point detection, which on its own has already been of high demand in application areas. For example, in a biological neural network, the recorded data are spike trains recorded on neurons and are in the form of integers. It is of increasing interest to detect and understand the underlying changes in such a network. In a communication network, the data can be the number of emails sent by individuals from a large firm and are again in the form of integers. Estimating underlying changes in the communication network has been used for legal investigation among many other uses.

This paper describes a computationally- and statistically-efficient methodology for detecting changes in the network underlying SEPPs. At the heart of our method lies a penalized dynamic programming algorithm that estimates the times at which each change occurs when the underlying network is sparse, i.e. when the number of network edges is small relative to the number of pairs of network nodes. In this paper, we also apply our method to neuron spike train data sets to help pinpoint the times at which the functional networks might change due to the changes of the state of consciousness.

We would like to point out that in this paper, we address challenging theoretical issues that go well beyond simply combining the existing results, including two closely relevant papers Mark et al. [2018] and Wang et al. [2019].

- Mark et al. [2018] addressed penalized regression for SEPPs, a framework that does appear in this paper. However, there is *no* change in the underlying distribution in Mark et al. [2018]. In contrast, in this paper, it is essential that we characterize what happens when we perform penalized regression over an interval that does contain a change point, i.e. when there is more than one distribution governing the data generation. Such analysis goes beyond the scope of Mark et al. [2018] and is a major technical contribution of our submission. More generally, it has been known for decades in the change point detection community that consistency in regression settings does not generally translate to the consistency in change point detection.

- Wang et al. [2019] does consider sparse regression over a time series that might contain a change point, but it only considers linear models. In contrast, the SEPP model necessitates considering nonlinear models. There is a wealth of literature on GLMs showing that fitting linear models to GLM data without accounting for nonlinear link functions is highly problematic both theoretically and empirically. More specifically, a key technical task is to characterize the population quantity corresponding to fitting a nonlinear model to a time series containing a change point in high dimensions. This is a challenging task that has not been studied in any past paper of which we are aware and which requires non-trivial arguments that go well beyond combining or simply extending known results. In Section 3, we have preprocessed the data and applied the methods developed in Wang et al. [2019]. We have shown the limitations of applying methods designed for linear data to nonlinear data.

More comparisons with the aforementioned two papers can be found in Section 2.2.

1.1 Problem formulation

The detailed model considered in this paper is introduced as follows.

Model 1. *Let $\{X(t)\}_{t=1}^{T} \subset \mathbb{Z}^M$ be a discrete-time Poisson process. For each $t \in \{1, \ldots, T\}$, let $\mathcal{X}(t) = (X(1), \ldots, X(t)) \in \mathbb{R}^{M \times t}$ consist of all the his-*

tory up to time t*. For each* $t \in \{1, \ldots, T-1\}$ *and* $m \in \{1, \ldots, M\}$*, suppose that given* $\mathcal{X}(t)$*, all coordinates* $\{X_m(t+1)\}$ *are conditionally independent and the conditional distribution of* $X_m(t+1)$ *is a Poisson distribution, i.e.*

$$X_m(t+1)|\mathcal{X}(t) \sim \textsc{Poisson}(\exp\{\lambda_m(t)\}), \tag{1.1}$$

where

$$\lambda_m(t) = v + A^*_m(t) g_t\{\mathcal{X}(t)\}, \tag{1.2}$$

the matrix $A^*(t) \in \mathbb{R}^{M \times M}$ *is the coefficient matrix at time point* t*,* $A^*_m(t)$ *is the* m*-th row of* $A^*(t)$ *and* $g_t(\cdot) : \mathbb{R}^{M \times t} \to \mathbb{R}^M$ *is an* M*-dimensional vector-valued function.*

Suppose that there exists an integer $K \geq 0$ *and time points* $\{\eta_k\}_{k=0}^{K+1}$*, called change points, satisfying* $1 = \eta_0 < \eta_1 < \ldots < \eta_K \leq T < \eta_{K+1} = T+1$ *and* $A^*(t) \neq A^*(t-1)$*, if and only if* $t \in \{\eta_k\}_{k=1}^{K}$*. Let the minimal spacing and the minimal jump size be defined as* $\Delta = \min_{k=1,\ldots,K+1}(\eta_k - \eta_{k-1})$ *and* $\kappa = \min_{k=1,\ldots,K} \|A^*(\eta_k) - A^*(\eta_k - 1)\|_{\mathrm{F}}$*, respectively, where* $\|\cdot\|_{\mathrm{F}}$ *denotes the Frobenius norm of a matrix.*

Compared to the abundance of the existing literature, we would like to highlight that Model 1 allows for change points in a high-dimensional integer-valued time series. In Model 1, we assume that up to time t, we observe a series of discrete events associated with M nodes. For each node

$m \in \{1, \ldots, M\}$, we model the marginal distribution of $X_m(t+1)$ using a point process with time-varying rate function $\exp(\lambda_m(t))$ that reflects how many events at time point $t+1$, node m is expected to participate. In order to incorporate the temporal dependence of the time series, we further assume that $\lambda_m(t)$ is a linear function of $\mathcal{X}(t) = [X_1(t), \ldots, X_m(t)]$. We remark that Model 1 resembles the high-dimensional vector autoregressive (AR) model, with the main difference being that all of our observations $\mathcal{X}(t)$ are vectors of integers. So we use generalized linear regression instead of linear regression to establish the temporal dependence between $\mathcal{X}(t)$ and $\mathcal{X}(t+1)$.

Remark 1 (The intercept)**.** *In Model 1, we assume that the intercept v stays constant across coordinates and the time course. We remark that in many applications, the intercept plays the role of background noise and it is common practice in the existing literature to treat it as a constant. On the other hand, allowing the intercept to vary across coordinates and/or time indeed increases the flexibility of the model. With additional assumptions imposed on the intercepts, the varying intercept case can be seen as a special case of our results through a simple change of variable argument. We will consider allowing for varying intercept in the future work.*

It is worth mentioning that $\{X(t)\}_{t=1}^{T}$ defined in Model 1 is an SEPP, where each $X_m(t)$ is conditionally distributed as a Poisson random variable. We therefore refer to (1.1) as a self-exciting Poisson process. When there is no ambiguity, we will also refer to self-exciting Poisson processes as SEPPs.

In fact, Model 1 is a generalization of a stationary SEPP process, which assumes that the coefficient matrices $A^*(t) = A^*(1)$, $t \in \{1, \ldots, T\}$. Stationary SEPP models have been well-studied in the existing literature, including Hall et al. [2018] and Mark et al. [2018], where it has been shown that the coefficient matrix of the point process can be estimated by an ℓ_1-penalized likelihood estimator.

Given $\{X(t)\}_{t=1}^{T}$ satisfying Model 1, our main task is to estimate $\{\eta_k\}_{k=1}^{K}$ accurately. To be specific, we seek estimators $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ such that as the sample size T diverges, with probability tending to 1, it holds that

$$\widehat{K} = K \quad \text{and} \quad \epsilon/\Delta = \Delta^{-1} \max_{k=1,\ldots,K} |\widehat{\eta}_k - \eta_k| \to 0. \tag{1.3}$$

For the change point estimators satisfying (1.3), we call them *consistent* change point estimators. We will also call ϵ the *localization error*.

To the best of our knowledge, we are the first to study the high-dimensional SEPPs with change points. In addition to the mathematical introduction of the model, we investigate the consistency of the abrupt change point location estimators, under minimal conditions. The proposed penalized dynamic programming approach in Section 2 is computationally efficient and tailored for this novel setting.

Notation. For any integer pair $(t_1, t_2) \in \mathbb{Z}^2$, let $[t_1, t_2]$ denote the integer interval $[t_1, t_2] \cap \mathbb{Z}$. Same notation applies to open intervals. For any matrix $A \in \mathbb{R}^{M \times M}$, let A_m denote the mth row of A and $A_{m,m'}$ denote the (m, m')th entry of A. With some abuse of notation, for any vector v and any matrix M, let $\|v\|_2$, $\|v\|_1$, $\|M\|_{\mathrm{F}}$ and $\|M\|_1$ be the ℓ_2- and ℓ_1-norms of v, the Frobenius norm of M and the ℓ_1-norm of $\mathrm{vec}(M)$, respectively, where $\mathrm{vec}(M)$ is the vectorized version of M by stacking all the columns of M. For any $v(t) : [1, T] \to \mathbb{R}^m$, let $\|Dv\|_0 = \sum_{t=2}^{T} I\{v(t-1) \neq v(t)\}$, where $I\{\cdot\} \in \{0, 1\}$ is the indicator function. For any set $S \subset \{(m, m') : m, m' = 1, \ldots, M\}$, let $A_S \in \mathbb{R}^{M \times M}$ satisfy $(A_S)_{m,m'} = A_{m,m'}$, if $(m, m') \in S$, and $(A_S)_{m,m'} = 0$, otherwise. Given any $A(t) : [1, T] \to \mathbb{R}^{M \times M}$ and any $I \subset [1, T]$, if $A(\cdot)$ is unchanged in I, then we denote $A(I) = A(t)$, $t \in I$.

2. The Penalized Dynamic Programming Algorithm

To detect the change points in Model 1, we propose the penalized dynamic programming (PDP) algorithm, stated in (2.7) with necessary notation in (2.4), (2.5) and (2.6). The PDP consists of two layers: estimation of the coefficient matrices $A^*(t)$, $t \in [1, T]$, and estimation of the change points.

For the coefficient matrix estimation, we let $\widehat{A}(I)$ be the penalized log-likelihood estimator of the coefficient matrix over an integer interval $I \subset [1,T]$, i.e.

$$\widehat{A}(I) = \operatorname*{argmin}_{A \in \mathcal{C}} H(A, I), \tag{2.4}$$

where $H(A, I)$ and $\mathcal{C}$ are the penalized log-likelihood function and the constrained domain of the coefficient matrices, respectively. To be specific, with a pre-specified tuning parameter $\lambda > 0$ and $I = [s, e]$, let

$$\begin{aligned} H(A, I) = \sum_{t=s}^{e-1} \sum_{m=1}^{M} \big(& \exp\left[v + A_m g_t\{\mathcal{X}(t)\}\right] \\ & - X_m(t+1)[v + A_m g_t\{\mathcal{X}(t)\}]\big) + \lambda |I|^{1/2} \|A\|_1 \end{aligned} \tag{2.5}$$

and

$$\mathcal{C} = \left\{ A \in \mathbb{R}^{M \times M} : \max_{m=1,\ldots,M} \|A_m\|_1 \leq 1 \right\}. \tag{2.6}$$

The loss function $H(\cdot, \cdot)$ is a penalized negative logarithmic conditional likelihood function, recalling that $X_m(t+1)$ given $\mathcal{X}(t)$ follows a Poisson distribution with intensity $\exp\left[v + A_m g_t\{\mathcal{X}(t)\}\right]$. The penalty term $\lambda |I|^{1/2}$ in (2.5) is introduced in a way such that the tuning parameter λ is independent of the interval length. The term $|I|^{1/2}$ reflects the order of the standard error of the sum of $|I|$ marginal log-likelihood functions. We elaborate on this scaling factor and its derivation in Lemma S8 and its proof.

The constraint on $\mathcal{C}$ is to ensure that the SEPP process as vector-valued time series is stable [see e.g. Lütkepohl, 2005]. As for stationary SEPP estimation, Mark et al. [2018] proposed a constraint similar to (2.6).

Given the above framework, we can now consider estimating change points by setting

$$\widehat{\mathcal{P}} = \underset{\mathcal{P}}{\operatorname{argmin}} \left\{ \sum_{I \in \mathcal{P}} H(\widehat{A}(I), I) + \gamma |\mathcal{P}| \right\}, \tag{2.7}$$

where $\gamma > 0$ is a tuning parameter, the minimization is over all possible interval partitions of $[1, T]$ and $\mathcal{P}$ denotes one such partition. To be specific, an interval partition has the form $\mathcal{P} = \{I_k,\, k = 1, \ldots, K_{\mathcal{P}}\}$ and satisfies $I_{k'} \cap I_k = \emptyset$ and $\bigcup_{k=1}^{K_{\mathcal{P}}} I_k = [1, T]$. Once $\widehat{\mathcal{P}}$ is at hand, we let $\widehat{K} = |\widehat{\mathcal{P}}| - 1 \geq 0$, $\eta_{\widehat{K}+1} = T + 1$ and

$$\widehat{\mathcal{P}} = \left\{ \{1, \ldots, \widehat{\eta}_1 - 1\}, \ldots, \{\widehat{\eta}_k, \ldots, \widehat{\eta}_{k+1} - 1\}_{k=1}^{\widehat{K}} \right\}.$$

We call $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ the change point estimators induced by $\widehat{\mathcal{P}}$.

The optimization problem in (2.7) is known as the minimal partitioning problem on a linear chain graph and can be solved using dynamic programming [e.g. Friedrich et al., 2008] with the worst case computational cost of order $O\{T^2 \mathrm{Cost}(T)\}$, where $\mathrm{Cost}(T)$ denotes, in our case, the computational cost of computing $\widehat{A}(I)$ in the interval I with $|I| = T$. Using coordinate decent, one can achieve $\mathrm{Cost}(T) = O(TM^2)$. We remark that there has been a line of attack on the computational aspect of optimizing the minimal partition problem, including Killick et al. [2012] and Maidstone et al. [2017], among others. Some variants [e.g. the PELT algorithm proposed in Killick et al., 2012] of the minimal partition problem can have a linear computational cost, under stronger model assumptions. We remark that in practice, one may use these variants to solve (2.7), but the theoretical results in this paper only hold when the minimal partition algorithm is executed.

For completeness, we summarize the PDP procedure in Algorithm 1 below. The quantities and functions involved are defined in (2.4), (2.5) and (2.6).

2.1 Localization rate of the PDP estimators

In order to establish the consistency of the change point estimators resulting from the PDP procedure detailed in Algorithm 1, we first impose Assumption 1.

Assumption 1. *Let $\{X(t)\}_{t=1}^{T} \subset \mathbb{Z}^M$ be a discrete-time* SEPP *generated according to Model 1 and satisfying the following.*

A1. *There exists a subset $S \subset \{(m, m') : m, m' = 1, \ldots, M\}$ such that for all $t \in [1, T]$, $A^*_{m,m'}(t) = 0$, if $(m, m') \notin S$. Let $d = |S|$.*

Algorithm 1 Penalized Dynamic Programming. PDP($\{X(t)\}_{t=1}^{n}, \lambda, \gamma$)

```
INPUT: Data $\{X(t)\}_{t=1}^{T}$, tuning parameters $\lambda, \gamma > 0$.
  Set $\mathcal{B} = \emptyset$, $\mathfrak{p} = \underbrace{(0, \ldots, 0)}_{T}$, $B = \underbrace{(\infty, \ldots, \infty)}_{T}$ and $B_0 = -\gamma$. Denote $B_i$ to
  be the $i$-$th$ entry of $B$.
  for $r$ in $\{1, \ldots, T]\}$ do
      for $l$ in $\{1, \ldots, r]\}$ do
          $b \leftarrow B_{l-1} + \gamma + H(\widehat{A}(I), I))$, where $I = [l, \ldots, r]$;
          if $b < B_r$ then
              $B_r \leftarrow b$;
              $\mathfrak{p}_r \leftarrow l - 1$.
          end if
      end for
  end for
  To compute the change point estimates from $\mathfrak{p} \in \mathbb{N}^T$, $k \leftarrow T$.
  while $k > 0$ do
      $h \leftarrow \mathfrak{p}_k$;
      $\mathcal{B} = \mathcal{B} \cup h$;
      $k \leftarrow h$.
  end while
OUTPUT: The estimated change points $\mathcal{B}$.
```

A2. *It holds that*

$$\max_{t=1,\ldots,T} \max_{m=1,\ldots,M} \|A_m^*(t)\|_1 \leq 1.$$

A3. *For any $\xi > 0$, there exist absolute constants $C_{\Delta,1}, C_{\Delta,2} > 0$ such that*

$$\Delta \geq C_{\Delta,1} T \quad \textit{and} \quad \Delta \geq C_{\Delta,2} \log^{2+\xi}(TM) d^2 \max\{\kappa^{-2}, \kappa^{-4}\}.$$

A4. *There exist absolute constants $p \in \mathbb{Z}^+$ and $\omega > 0$ such that for any t, the matrix*

$$\mathbb{E}[g_t\{\mathcal{X}(t)\} g_t\{\mathcal{X}(t)\}^\top | \mathcal{X}(t-p)] - \omega I_M$$

is positive definite, where $I_M \in \mathbb{R}^{M \times M}$ is an identity matrix. In addition, v and $\|g_t(\cdot)\|_\infty$, for all t, are uniformly upper bounded by an absolute constant $C_g > 0$.

Model 1 and Assumption 1 completely characterize the problem with model parameters M (the dimensionality of the time series), d (the sparsity parameter indicating an upper bound of the number of nonzero entries in all the coefficient matrices), Δ (the minimal spacing between change points), and κ (the minimal jump size), along with the sample size T. The consistency we are to establish is based on allowing M and d to diverge and κ to vanish as the sample size T diverges unbounded.

The number of parameters at each time point is of order M^2, which is allowed to well exceed the sample size. A sparsity constraint therefore comes into force in Assumption **A1**, which is a standard assumption in the high-dimensional statistics literature. Note that the set S is the union of all (m, m') pairs with a nonzero entry in any coefficient matrix. Assumption **A2** echoes the imposition of the constraint domain $\mathcal{C}$ (2.6) in the optimization (2.4), to ensure the stationarity of the SEPP. In fact, the constant one in the upper bound can be relaxed to any absolute constant and is set to be one in this paper for identification issue. To be specific, what goes into the model is the product of $A_m(t)$ and $g_t\{\mathcal{X}(t)\}$, and the latter is assumed to be upper bounded in sup-norm in Assumption **A4**.

Assumption **A3** can be regarded as a signal-to-noise assumption. It is required that the minimal spacing Δ is at least of a constant fraction of the total sample size, which implies the number of change points is of order $O(1)$. This might appear to be strong compared to other change point detection literature, however, the problem we are facing here is challenging due to the nonlinearity of the SEPP model. In order to estimate the change points accurately, one needs to estimate the underlying distribution. In the analysis, one needs to deal with intervals, say I, containing more than one underlying distributions, and control the estimation error $\|\widehat{A}_I - A_I^*\|_{\mathrm{F}}$, where $\widehat{A}_I$ is the penalized estimator and A_I^* is the population coefficient matrix for the whole interval I. With nonlinear models, such as the SEPP model considered here, it is hard to characterize A_I^*. As a consequence, we resort to the current minimal spacing condition, which is still the sharpest in the existing literature. The number of change points can grow with n if we assume knowledge of the minimal spacing between change points, Δ. In this case, we can repeat our proposed PDP method in every segment of length $C\Delta$, where $C > 1$ is an absolute constant. Thus we focus in the below on the setting where Δ is unknown.

In fact, Assumption **A3** is a mild condition and covers some challenging scenarios. For instance, Assumption **A3** holds if $M \asymp \exp(T^{1/2})$, $d \asymp T^{1/4}$ and $\kappa \asymp \log(T)$. The quantity ξ can be set arbitrarily small and it ensures the consistency of the estimator which will be explained after Theorem 1.

Assumption **A4** can be interpreted as the restricted eigenvalue condition for SEPP processes. We refer readers to Section 4 of Mark et al. [2018] for a number of common self-excited point process models satisfying Assumption **A4**.

In what follows, we show the consistency of PDP in Theorem 1.

Theorem 1. *Let $\{X(t)\}_{t=1}^T \subset \mathbb{Z}^M$ be an SEPP generated from Model 1 and satisfying Assumption 1. Let $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ be the change point estimators from the PDP algorithm detailed in Algorithm 1 with tuning parameters*

$$\lambda = C_\lambda \log(TM) \quad \text{and} \quad \gamma = C_\gamma \log^2(TM) d \left(1 + d\kappa^{-2}\right), \tag{2.8}$$

where $C_\lambda, C_\gamma > 0$ are absolute constants, depending only on p, ω, $C_{\Delta,1}$, $C_{\Delta,2}$ and C_g. We have that

$$\mathbb{P}\left\{\widehat{K} = K \quad \textit{and} \quad \max_{k=1,\ldots,K} |\widehat{\eta}_k - \eta_k| \leq C_\epsilon d^2 \log^2(TM) \max\{\kappa^{-2}, \kappa^{-4}\}\right\}$$
$$\geq 1 - 2(TM)^{-1},$$

where $C_\epsilon > 0$ is an absolute constant only depending on p, ω, $C_{\Delta,1}$, $C_{\Delta,2}$ and C_g.

The proof of Theorem 1 is deferred to Section S2, where it can be seen that the order of the estimation error is of the form

$$\frac{\lambda^2 d}{\kappa^2} + \frac{\lambda^2 d^2}{\kappa^4} + \frac{\gamma}{\kappa^2}.$$

Due to the signal-to-noise ratio condition in Assumption **A3**, we have that

$$\frac{\max_{k=1,\ldots,K} |\widehat{\eta}_k - \eta_k|}{\Delta} \lesssim \frac{d^2 \log^2(TM) \max\{\kappa^{-2}, \kappa^{-4}\}}{\Delta}$$
$$\lesssim \frac{d^2 \log^2(TM) \max\{\kappa^{-2}, \kappa^{-4}\}}{d^2 \log^{2+\xi}(TM) \max\{\kappa^{-2}, \kappa^{-4}\}} \to 0,$$

as $T \to \infty$. This explains the role of the quantity ξ in Assumption **A3** and shows the consistency of the PDP algorithm. In fact, if we let $d = 1$ and assume $\kappa > 1$, then the localization error we derived here coincides with the optimal localization error in the univariate mean change point detection problem [e.g. Wang et al., 2020].

Two tuning parameters are involved, where λ is used in the optimization (2.5) to recover the sparsity in estimating high-dimensional coefficient matrices, and γ is involved in optimizing (2.7) to penalize the over-partitioning. The order of λ required in (2.8) is a logarithmic quantity in T and M, which is resulted from a union bound argument applied to a sub-exponential concentration bound. The requirement on γ is essentially that $\gamma \asymp \lambda^2 \left(d + d^2\kappa^{-2}\right)$, which can be intuitively explained as an upper bound on the difference between $H(\widehat{A}(I_1), I_1) + H(\widehat{A}(I_2), I_2)$ and $H(\widehat{A}(I_1 \cup I_2), I_1 \cup I_2)$, where I_1 and I_2 are two relatively long, non-overlapping and adjacent intervals, and there is no true change point near the shared endpoint of I_1 and I_2. In this case, one would not wish to partition $I_1 \cup I_2$ into I_1 and I_2. If we only focus on the log-likelihood functions, over-estimating will result in that

$$H(\widehat{A}(I_1), I_1) + H(\widehat{A}(I_2), I_2) < H(\widehat{A}(I_1 \cup I_2), I_1 \cup I_2).$$

The penalty we impose through γ will therefore avoid this over-partitioning.

2.2 Comparisons with related work

In a broad sense, there have been numerous existing papers on different aspects of SEPPs. Another related area is the analysis of piecewise-stationary time series models, where we also see a vast volume of existing papers. The two most related papers are Mark et al. [2018], which is concerned with a stationary, high-dimensional SEPP, and Wang et al. [2019], which studies a piecewise-stationary high-dimensional linear process.

Mark et al. [2018] studied a stationary version of Model 1 with $K = 0$. The penalized estimator of the coefficient matrix developed there is almost identical to the ones summoned in our problem in (2.4). There are a few fundamental differences between this paper and Mark et al. [2018]. (1) Due to the piecewise-stationarity assumed in Model 1, when estimating the coefficient matrices in (2.4) and (2.5), it is possible that there exists a true change point in the interval of interest and the estimator we seek is an estimator of a mixture of different true coefficient matrices. (2) We provide a more refined analysis as an improved version of Mark et al. [2018], for instance, the optimization constrain domain $\mathcal{C}$ defined in (2.6) is a cleaner version of its counterpart in Mark et al. [2018]; a subspace compatibility condition is required in Mark et al. [2018] to control the ratio of different norms of the coefficient matrix, and this assumption is shown to be redundant in our new analysis.

The other closest-related work is Wang et al. [2019], where the change point localizing problem in the piecewise-stationary vector autoregressive models is investigated and a penalized dynamic programming approach was deployed there. The main differences between this paper and Wang et al. [2019] come from the underlying model. The vector autoregressive model is a linear model in the sense that given $\mathcal{X}(t)$ the history data up till time point t, the conditional expectation of $X(t+1)$ is a linear combination of the columns of $\mathcal{X}(t)$, which is not the case here. The self-exciting point process is a nonlinear model, and as we have mentioned, the logarithm of the conditional intensity is a linear function of the history. Another key difference is that Wang et al. [2019] are concerned with sub-Gaussian innovation sequences, while the counting processes we study here determine the heavy-tail properties of the data.

3. Numerical Experiments

In this section, we further examine the performances of the PDP algorithm by numerical experiments, with simulated data analyzed in Section 3.1 and a real data set in Section 3.2.

3.1 Simulated data analysis

We generate data according to Model 1 and Assumption 1. In particular, we adopt the setting in Mark et al. [2018] and assume that the design function

$g_t(\cdot)$ is defined to be

$$g_t\{\mathcal{X}(t)\} = \left(\min\{\mathcal{X}_1(t), C_g\}, \ldots, \min\{\mathcal{X}_M(t), C_g\}\right)^\top \in \mathbb{R}^M, \qquad (3.9)$$

where $C_g > 0$ is a constant, $\mathcal{X}(t)$ is an $M \times t$ matrix and $\mathcal{X}_m(t)$ denotes the mth row of $\mathcal{X}(t)$, $m \in \{1, \ldots, M\}$. For the two tuning parameters λ and γ defined in (2.5) and (2.7), respectively, with the theoretical guidance in Theorem 1, we fix $\lambda = 90 \log(TM)$ and $\gamma = \log^2(M)/2$ in all experiments in this section.

Remark 2 (The robustness of γ)**.** *Note that the tuning parameter γ is crucial in terms of determining the number of estimated change points. In our analysis, we in fact conducted identical analysis to a range of γ in all simulation settings. To be specific, we let $\gamma \in \{0.2, 0.5, 1, 1.3, 2\} \times \log^2(M)$ and they returned identical numerical results. We therefore omit presenting them separately but we remark the robustness of the choice of γ in our algorithms.*

Since the piecewise-stationary SEPP model is first introduced here, we do not have direct competitors. For illustration purpose, however, we compare our PDP algorithm with the SBS-MVTS algorithm [Cho and Fryzlewicz, 2015], E-Divisive procedure [Matteson and James, 2014] and VARDP algorithm [Wang et al., 2019], all of which are designed to detect abrupt change points in multivariate time series, but none is designed specifically for the scenarios we are studying here. Having said this, there are some reasons we choose these competitors. The SBS-MVTS algorithm can identify covariance changes in the high-dimensional autoregressive time series and the E-Divisive procedure can estimate of both the number and locations of change points under mild assumptions on the first or second moments of the underlying distributions. Since Poisson random variables have the same means and variances, these two competitors may be able to detect the changes in Poisson processes with piecewise-constant parameters. The VARDP adopts the same ℓ_0-penalization framework and can detect change points in the regression coefficients in the high-dimensional vector autoregressive models. In order to apply the VARDP, we add independent noise (Uniform$[0, 0.01]$) to every univariate data point $X_i(t)$, $i \in \{1, \ldots, M\}$ and $t \in \{1, \ldots, T\}$, then apply the logarithm transform on the resulting data. We remark that there is an optional local refinement (LR) second step to VARDP. It improves the results of VARDP, provided that VARDP produces consistent estimators.

In all the simulated experiments, the tuning parameters for SBS-MVTS algorithm and E-Divisive procedure are selected according to the information-type criteria and permutation tests in the R [R Core Team, 2017] packages wbs [Baranowski and Fryzlewicz, 2019] and ecp [Nicholas A. James and Matteson, 2019], respectively. The tuning parameters for VARDP are selected based on a cross-validation procedure (`https://github.com/darenwang/vectordp`).

Let $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ and $\{\eta_k\}_{k=1}^{K}$ be a collection of change point estimates and a collection of true change points, respectively. We evaluate the estimators' performances by the absolute error $|K - \widehat{K}|$ and their Hausdorff distance. The Hausdorff distance between two sets $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$\mathcal{D}(\mathcal{A}, \mathcal{B}) = \max\{d(\mathcal{A}|\mathcal{B}),\, d(\mathcal{B}|\mathcal{A})\}, \tag{3.10}$$

where

$$d(\mathcal{A}|\mathcal{B}) = \max_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} |a - b|.$$

In the sequel, we consider three settings. Recall that T is the total number of time points, M is the dimensionality of the time series and C_g is the threshold used in the design function $g_t(\cdot)$, which is specified in (3.9). Every setting is repeated 100 times. Additional setting details are listed below.

(a) One change point and varying jump size. Fix $T = 450$, $M = 30$,

$C_g = 6$ and the intercept $v = 1/2$, which is defined in (1.2). Let

$$A^*(t) = \begin{cases} (\rho v_1, \rho v_2, 0_{M\times(M-2)}) \in \mathbb{R}^{M\times M}, & t \in [1, 150], \\ (\rho v_2, \rho v_1, 0_{M\times(M-2)}) \in \mathbb{R}^{M\times M}, & t \in [151, 450], \end{cases}$$

where $v_1 \in \mathbb{R}^M$ with odd coordinates being 1 and even coordinates being -1, $v_2 = -v_1$, $0_{M\times(M-2)} \in \mathbb{R}^{M\times(M-2)}$ is an all zero matrix and $\rho \in \{0.15, 0.20, 0.25, 0.30, 0.35\}$.

(b) Two change points and varying minimal spacing. Let

$$T \in \{180, 240, 300, 360, 420\},$$

$M = 40$, $C_g = 8$ and the intercept $v = 1/4$. Let the coefficient matrices satisfy $(A^*(t))_{ij} = 0$, $|i-j| > 1$, $t \in [1, T]$,

$$(A^*(t))_{ij} = \begin{cases} \begin{cases} 0.15 & t \in [1, T/3] \cup (2T/3, T], \\ -0.15 & t \in (T/3, 2T/3], \end{cases} & i = j, \\ \begin{cases} -0.15 & t \in [1, T/3], \\ 0.15 & t \in (T/3, T], \end{cases} & i - j = -1, \\ \begin{cases} 0.15 & t \in [1, 2T/3], \\ -0.15 & t \in (2T/3, T], \end{cases} & i - j = 1. \end{cases}$$

(c) Two change points and varying dimension. Let $T = 450$, $C_g = 4$, $v = 1/5$ and $M \in \{15, 20, 25, 30, 35\}$. Let

$$A(t) = \begin{cases} (v_1, v_2, v_3, 0_{M\times(M-3)}), & t \in [1, 150], \\ (v_2, v_3, v_3, 0_{M\times(M-3)}), & t \in [151, 300], \\ (v_3, v_2, v_1, 0_{M\times(M-3)}), & t \in [301, 450], \end{cases}$$

where $v_1, v_2, v_3 \in \mathbb{R}^M$ are

$$v_1 = (-0.075, 0.15, 0.3, -0.3, 0, \ldots, 0)^\top,$$

$$v_2 = (\underbrace{0, \ldots, 0}_{4}, 0.375, -0.225, -0.075, 1.5, 0.225, 0, \ldots, 0)^\top,$$

$$v_3 = (\underbrace{0, \ldots, 0}_{8}, -0.15, -0.075, 0.45, -0.225, 0, \ldots, 0)^\top.$$

We collect the simulation results in Tables 1, 2 and 3, for Settings (a), (b) and (c), respectively. Each cell contains the mean and standard errors of 100 repetitions. The Hausdorff distances are visualized in Figure 1 to improve readability. These three settings have ranged over various situations. It is clearly that PDP outperforms both competitors in all settings on both metrics. We notice that, PDP outperforming VARDP demonstrates that it is crucial to develop nonlinear-data-specific methods. Merely pre-processing data and applying linear-model-specific methods are not reliable.

Table 1: Simulation results of Setting (a). Each cell is in the form of mean(standard error). For the metrics, $\mathcal{D}$ denotes the Hausdorff distance defined in (3.10) and $|\widehat{K} - K|$ denotes the absolute errors in estimating the numbers of the change points. PDP uniformly outperforms the other methods across a range of ρ values, reflecting the jump size.

	Metric	$\rho = 0.15$	$\rho = 0.20$	$\rho = 0.25$	$\rho = 0.30$	$\rho = 0.35$
PDP	$\mathcal{D}$	3.1(9.8)	1.1(1.0)	0.7(0.5)	0.6(0.5)	0.6(0.5)
SBS		282.6(69.1)	226.5(119.9)	114.7(130.8)	47.3(52.9)	9.3(21.3)
ECP		151.0(0.0)	151.0(0.0)	151.0(0.0)	151.0(0.0)	151.0(0.0)
VAR		131.16(6.33)	44.84(7.61)	100.12(6.16)	67.68(6.86)	60.28(8.01)
VAR(LR)		123.76(8.00)	30.24(7.52)	104.36(6.03)	72.68(7.19)	52.68(8.21)
PDP	$\|\widehat{K} - K\|$	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)
SBS		0.9(0.2)	0.7(0.4)	0.4(0.5)	0.5(0.5)	0.1(0.3)
ECP		300.0(0.0)	300.0(0.1)	300.0(0.5)	296.4(16.2)	287.2(31.4)
VAR		0.82(0.05)	0.22(0.06)	4.52(0.32)	1.20(0.14)	0.78(0.13)
VAR(LR)		0.82(0.05)	0.22(0.06)	4.52(0.32)	1.20(0.14)	0.78(0.13)

Table 2: Simulation results of Setting (b). Each cell is in the form of mean(standard error). For the metrics, $\mathcal{D}$ denotes the Hausdorff distance defined in (3.10) and $|\widehat{K} - K|$ denotes the absolute errors in estimating the numbers of the change points. PDP uniformly outperforms the other methods across a range of T values, reflecting the minimal spacing.

	Metric	$T = 180$	$T = 240$	$T = 300$	$T = 360$	$T = 420$
PDP	$\mathcal{D}$	11.5(6.2)	3.7(4.6)	2.5(4.6)	2.8(4.3)	1.2(3.6)
SBS		177.0(21.1)	233.3(38.1)	270.1(85.5)	243.8 (156.1)	263.5(185.2)
ECP		61.0(0.0)	81.0(0.0)	101.0(0.0)	121.0(0.0)	141.0(0.0)
VAR		58.08(1.16)	76.20(1.72)	87.92(4.08)	106.52(4.75)	126.04(4.84)
VAR(LR)		56.96(1.35)	76.00(1.66)	87.36(3.74)	104.36(5.02)	122.12(5.12)
PDP	$\lvert\widehat{K} - K\rvert$	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)
SBS		2.0(0.2)	1.9(0.3)	1.9(0.4)	1.6(0.7)	1.6(0.6)
ECP		178.9(0.3)	238.9(0.3)	298.9(0.3)	358.8(0.4)	418.8(0.4)
VAR		1.92(0.06)	2.04(0.08)	1.94(0.15)	1.96(0.18)	2.30(0.28)
VAR(LR)		1.92(0.06)	2.04(0.08)	1.94(0.15)	1.96(0.18)	2.30(0.28)

Table 3: Simulation results of Setting (c). Each cell is in the form of mean(standard error). For the metrics, $\mathcal{D}$ denotes the Hausdorff distance defined in (3.10) and $|\widehat{K} - K|$ denotes the absolute errors in estimating the numbers of the change points. PDP uniformly outperforms the other methods across a range of M values, the dimension of the time series.

		Setting (c)				
	Metric	$M = 15$	$M = 20$	$M = 25$	$M = 30$	$M = 35$
PDP	$\mathcal{D}$	3.3(5.0)	3.6(5.5)	3.2(4.5)	5.0(12.4)	6.1(13.2)
SBS		401.4(112.8)	378.2(129.9)	411.3(101.7)	377.7(134.1)	375.4(134.5)
ECP		151.0(0.0)	151.0(0.0)	151.0(0.0)	151.0(0.0)	151.0(0.0)
VAR		134.24(4.19)	146.84(2.17)	148.40(0.95)	149.08(0.84)	146.78(2.79)
VAR(LR)		131.44(6.42)	146.64(2.87)	149.72(0.10)	149.88(0.07)	149.56(0.34)
PDP	$\lvert\widehat{K} - K\rvert$	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)
SBS		1.8(0.4)	1.7(0.5)	1.9(0.3)	1.8(0.4)	1.8(0.4)
ECP		448.6(0.5)	449.0(0.3)	449.0(0.1)	449.0(0.0)	449.0(0.0)
VAR		1.48(0.20)	1.74(0.07)	1.86(0.05)	1.94(0.03)	1.89(0.76)
VAR(LR)		1.48(0.20)	1.74(0.07)	1.86(0.05)	1.94(0.03)	1.89(0.76)

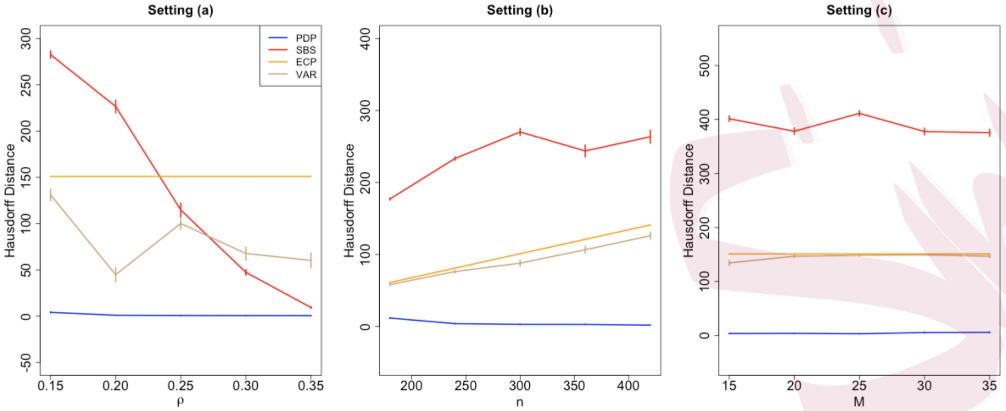

Figure 1: A visualization of the mean Hausdorff distance metric $\mathcal{D}$ (3.10) in Tables 1, 2 and 3. The methods concerned are: Algorithm 1, PDP; SBS, SBS-MVTS; ECP, E-Divisive; VAR, VARDP. In each panel, the y-axis represents the mean Hausdorff distance across 100 repetitions and the x-axis represents the varying parameter in each setting. PDP uniformly outperforms the other two methods across a range of parameter values, including ρ (reflecting the jump size), T (the number of samples), and M (the dimension of the time series).

3.2 Real data example

We consider the neuron spike train data set previously analyzed in Watson et al. [2016a]. The three chosen data sets are from Watson et al. [2016b] and each consists of wake-sleep episodes of multi-neuron spike train recording sessions of one laboratory animal. Each wake-sleep episode includes at least 7 minutes of wake time followed by at least 20 minutes of sleep time. Note that the wake and sleep periods were recorded so the true change point in each dataset is the end of the wake period. For each data set, we first compute the Firing Rate (FR) of each neuron using a 5-second discretization time window and then apply Algorithm 1 with $\lambda = 800$ and $\gamma = \log^2(M)/2$, the same as in Section 3.1. For comparison, we also apply the SBS-MVTS algorithm [Cho and Fryzlewicz, 2015], E-Divisive procedure [Matteson and James, 2014] and the VARDP algorithm [Wang et al., 2019].

The subjects concerned are 20140528_565um, BWRat17_121912 and BWRat19_032413. The numbers of neurons, i.e. the dimensions of the time series M, are 24, 33 and 41, respectively. The total numbers of 5-second time intervals, i.e. the total number of time points T considered in Model 1, are 3750, 2995 and 3920, respectively. The true change points are at point 788, 1184 and 2001, respectively.

The results are summarized in Table 4 and are depicted in Figure 2.

Since the results of VARDP and its local refinement are very close, in Figure 2, we only depict those of VARDP. Since E-Divisive procedure outputs too many estimators, we omit them from Figure 2. As we can see from the table and the figure, our PDP algorithm consistently outperforms the other algorithms in these real data examples.

Table 4: The results of three algorithms on multi-neuron spike train data sets. For the metrics, $\mathcal{D}$ denotes the Hausdorff distance defined in (3.10) and $|\widehat{K} - K|$ denotes the absolute errors in estimating the numbers of the change points. PDP uniformly outperforms the other methods.

Subject	Metric	PDP	SBS	ECP	VAR	VAR(LR)
20140528_565um	$\mathcal{D}$	38	382	2966	708	646
	$\|\widehat{K} - K\|$	0	0	740	1	1
BWRat17_121912	$\mathcal{D}$	84	140	1816	1162	1138
	$\|\widehat{K} - K\|$	0	0	595	8	8
BWRat19_032413	$\mathcal{D}$	1	99	1996	1889	1989
	$\|\widehat{K} - K\|$	0	0	773	12	12

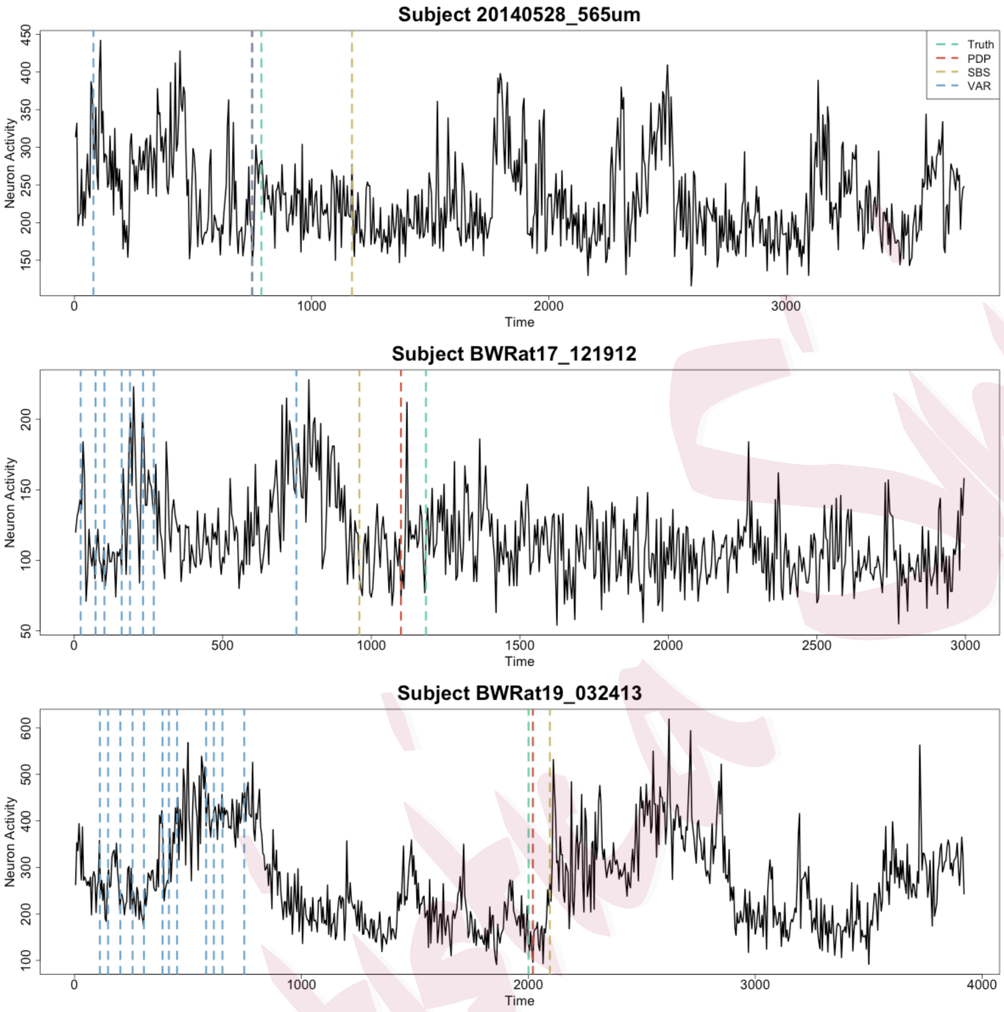

Figure 2: The true change points and the estimators provided by PDP and SBS-MVTS in the multi-neuron spike train data sets. Each panel corresponds to a subject. The y-axis represents the sum of the FRs across all neurons and the x-axis represents the ordered time intervals. The estimators of the E-Divisive procedure are not included because the corresponding $\widehat{K}$'s are too large. PDP uniformly outperforms the other methods. See Table 4 for detailed information.

4. Discussions

In this paper, we studied piecewise-stationary discrete-time high-dimensional self-exciting Poisson processes, which, or at least the theoretical properties of which were not studied in the literature. The number of stationary segments in the whole time series is assumed to be an unknown constant. All the other model parameters are allowed to be functions of the sample size T. We proposed a computationally-efficient and theoretically-guaranteed algorithm.

In the numerical experiments, we fix tuning parameters. One future research direction is to investigate data-driven methods for tuning parameter selection. Possible methods include variants of stationary bootstrap [Politis and Romano, 1994] or variants of information criteria [e.g. Chen and Chen, 2012].

Another future research direction is to extend the techniques we derived in this paper to other popular time series models. For instance, one key feature of the SEPPs we are concerned in this paper is the varying variance structure and heavy tail behaviours. These share similarities with the GARCH models, which are widely used in finance.

Supplementary Materials

All the proofs are in the Supplementary Materials.

References

Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger JA Laeven. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606, 2015.

Alexander Aue, Siegfried Hörmann, Lajos Horváth, and Matthew Reimherr. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087, 2009.

Rafal Baranowski and Piotr Fryzlewicz. *wbs: Wild Binary Segmentation for Multiple Change-Point Detection*, 2019. URL `https://cran.r-project.org/web/packages/wbs/index.html`. R package version 1.4.

Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. *Advances in neural information processing systems*, 25, 2012.

Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, 2004.

Valérie Chavez-Demoulin and JA McGill. High-frequency financial data modeling using hawkes processes. *Journal of Banking & Finance*, 36(12): 3415–3426, 2012.

Jiahua Chen and Zehua Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574, 2012.

Haeran Cho. Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, 10(2):2000–2038, 2016.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2): 475–507, 2015.

Holger Dette and Josua Gösmann. Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, 12(2):2578–2636, 2018.

Şeyda Ertekin, Cynthia Rudin, and Tyler H McCormick. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.

Felix Friedrich, Angela Kempe, Volkmar Liebscher, and Gerhard Winkler. Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008.

Eric C Hall and Rebecca M Willett. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346, 2016.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Eric C Hall, Garvesh Raskutti, and Rebecca M Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4):2401–2422, 2018.

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

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

Shuang Li, Yao Xie, Mehrdad Farajtabar, Apurv Verma, and Le Song. Detecting changes in dynamic events over networks. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2):346–359, 2017.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Network estimation from point process data. *IEEE Transactions on Information Theory*, 65(5):2953–2975, 2018.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

Wenyu Zhang Nicholas A. James and David S. Matteson. *ecp: Non-Parametric Multiple Change-Point Analysis of Multivariate Data*, 2019.

URL `https://cran.r-project.org/web/packages/ecp/index.html`. R package version 3.1.2.

Oscar Hernan Madrid Padilla, Yi Yu, and Carey E Priebe. Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv preprint arXiv:1911.07494*, 2019.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Alex Reinhart. Self-exciting point processes, 2019. URL `https://www.refsmmat.com/notebooks/self-exciting-point-processes.html`.

Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional non-stationary var models. *arXiv preprint arXiv:1711.07357*, 2017.

Alexey Stomakhin, Martin B Short, and Andrea L Bertozzi. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27(11):115013, 2011.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*, 2018.

Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*, 2019.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961, 2020.

Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83, 2018.

Brendon O Watson, Daniel Levenstein, J Palmer Greene, Jennifer N Gelinas, and György Buzsáki. Network homeostasis and state dynamics of neocortical sleep. *Neuron*, 90(4):839–852, 2016a.

Brendon O Watson, Daniel Levenstein, J Palmer Greene, Jennifer N Gelinas, and György Buzsáki. Multi-unit spiking activity recorded from rat frontal cortex (brain regions mpfc, ofc, acc, and m2) during wake-sleep episode wherein at least 7 minutes of wake are followed by 20 minutes of sleep, 2016b. URL `http://dx.doi.org/10.6080/K02N506Q`.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013.