Statistica Sinica Preprint No: SS-2024-0320	
Title	Model-free Multivariate Change Point Detection and Localization with Statistical Guarantee
Manuscript ID	SS-2024-0320
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0320
Complete List of Authors	Xin Xing, Zuofeng Shang, Hongyu Miao and Pang Du
Corresponding Authors	Pang Du
E-mails	pangdu@vt.edu
Notice: Accepted author version.	

Statistica Sinica

Model-free Multivariate Change Point Detection and Localization with Statistical Guarantee

Xin Xing, Zuofeng Shang, Hongyu Miao, Pang Du

Virginia Tech

New Jersey Institute of Technology

Florida State University

Virginia Tech

Abstract: Change point analysis involves the identification of potential shifts in the underlying data distribution and the precise estimation of the location of such a change point. Recent attention has been directed towards the adoption of nonparametric testing methods to address the former objective. However, existing research on the asymptotic behavior of change point tests is somewhat limited due to their reliance on an infinite series of nonparametric statistics. To address this limitation, we develop a CUSUM likelihood ratio test statistic based on nonparametric density estimation in the framework of reproducing kernel Hilbert spaces. Furthermore, we present a comprehensive non-asymptotic theoretical framework for nonparametric density estimation. We achieve an asymptotic control over type-I errors and can pinpoint the change point at an optimal rate. Besides sim-

Corresponding author: Pang Du, pangdu@vt.edu

ulations, our numerical studies also include a real-world application on neonatal seizure detection with multiple electrical signal channels.

Key words: multivariate change point detection; CUMSUM; RKHS

1. Introduction

In the past decades, rapid expansions of technology and increasing requirements from areas like finance [Zeileis et al., 2005], genetics [Chen and Gupta, 2012], and neuroscience [Toutounji and Durstewitz, 2018] have driven advances in statistical theories and methods for change point analysis. In this era of big data, change point analysis in sequences of multivariate and possibly non-Gaussian observations has received heightened interest. For instance, electrical recordings in neuroscience often involve multiple signal channels that exhibit non-Gaussian distributions, characterized by heavy tails or multimodality [Chae et al., 2021]. Detecting abrupt change points in such multivariate non-Gaussian data has been a longstanding challenge in the realms of statistics and machine learning [Carlstein et al., 1994, Polunchenko and Tartakovsky, 2012, Chen and Zhang, 2015].

Consider a sequence of multivariate observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ on the domain $\mathcal{X}$, with a potential change point denoted by $\tau \in \{1, \ldots, n\}$. It is assumed that up to time τ, the samples $\boldsymbol{x}_i$, $i = 1, \ldots, \tau$, are independent and identically distributed (i.i.d.) with a probability distribution denoted by P – referred to as the *background distribution* [Carlstein et al., 1994].

After the change point τ, the samples $\boldsymbol{x}_i$, $i > \tau$, are also i.i.d. but follow a distinct distribution Q known as the *post-change distribution* [Carlstein et al., 1994]. The precise location of the change point holds significant importance, as it provides a reliable foundation for further exploration into the underlying causes behind the observed changes. An example presented in this paper is the detection of seizures in newborns using electroencephalogram (EEG) data. Our empirical investigations reveal that our proposed methodology, capable of accommodating multiple electrical channels characterized by complex distributions, exhibits a substantial enhancement in both detection sensitivity and localization precision when compared to conventional univariate or parametric approaches.

Example: *EEG Seizure Detection.* Epileptic seizures [Litt and Echauz, 2002], manifest as sudden and uncontrolled electrical disruptions among brain cells, leading to temporary irregularities in muscle activity, behavioral patterns, or sensory perceptions. EEG [Teplan et al., 2002] is a medical diagnostic procedure employed to measure and record the electrical activity within the brain by utilizing sensors attached to the skin. In our study, EEG is utilized to pinpoint the onset of seizures. Given the intricate nature of EEG recordings, an effective automated detection method must incorporate both energy and frequency features inherent in the signals. The change

point analysis method proposed in this paper achieves this objective by employing a nonparametric approach to model the joint distributions of multiple features with enhanced accuracy and efficiency in seizure detection.

Change point analysis methods typically fall into two primary categories: parametric and nonparametric. In the parametric category, it is assumed that the underlying data distributions are members of a parametric family. In contrast, nonparametric methods offer a more versatile application spectrum, as they are not confined to specific distributional assumptions. Most literature on the detection of multiple change points in multivariate data are in the parametric category. A significant portion of these investigations are based on the cumulative sum statistic (CUSUM) in the realm of Gaussian distributions. Cho and Fryzlewicz [2015] pioneered a truncated CUSUM test, synergistically coupled with binary segmentation, to mitigate the issue of sparsity in datasets with high dimensions. Wang and Samworth [2018] investigated a variant of the CUSUM procedure tailored for high-dimensional sparse data constructs. Enikeeva and Harchaoui [2019] delved into the l_2 aggregation of CUSUM, presenting a nuanced approach to the methodology. Diverging from the conventional CUSUM-centric methodologies, Lavielle and Teyssiere [2006] offered an innovative array of techniques grounded in penalized Gaussian log-likelihood,

specifically designed for discerning shifts in covariance structures.

In the nonparametric domain, Einmahl and McKeague [2003] pioneered the utilization of non-parametric techniques derived from the empirical cumulative distribution function (CDF) for the identification of a single change point for univariate data. This approach was further augmented by Zou et al. [2014] to tackle the detection of multiple change points. Extending CDF-based methodologies to multivariate data presents challenges due to their dependency on the unreliable empirical CDF. In the realm of multivariate data, the efforts in kernel-based change point detection, as outlined in Harchaoui and Cappé [2007], were advanced by Arlot et al. [2019]. They unveiled the framework of a penalized kernel least squares estimator for the analysis of multivariate time series data. Furthermore, they presented an oracle inequality to rigorously validate the effectiveness of their proposed method. Around the same time, an innovative graph-based approach was introduced for testing and estimation of change points respectively in sequences of multivariate Chen and Zhang [2015] or non-Euclidean data Chu and Chen [2019]. Recently, Padilla et al. [2021b] and Padilla et al. [2023] considered nonparametric change point localization and inference for multivariate data with piecewise constant distributions, extending the univariate work in Padilla et al. [2021a].

Upon the establishment of the presence of a change point, the subsequent objective is the precise location of the change point. Recent endeavors have focused on addressing this in a multivariate setting. Garreau and Arlot [2018] presented an analytical exposition of an upper bound on the attainable localization rate of their proposed method. This theoretical threshold was further optimized from a computational perspective in subsequent studies Garreau and Arlot [2018], Celisse et al. [2018]. In general, the existing nonparametric detection methods predominantly revolves around the estimation consistency and rate of localization.

From the perspective of statistical inference, the change point analysis literature lacks rigorous quantitative assessments concerning the type-I error and localization accuracy in the realm of multivariate non-Gaussian data. This scarcity is particularly evident when pairing nonparametric inference with the complexities inherent in change point hypothesis tests. Notably, such testing often necessitates the introduction of an infinite series in nonparametric statistics, with CUSUM being a well-known example. The conventional CUSUM statistic is a cumulative aggregation of score statistics, which are typically derived from explicit parametric models, predominantly Gaussian in nature. However, these traditional methodologies might not adequately encapsulate the multifaceted nature inherent in con-

temporary datasets [Romano et al., 2021, Chen and Tian, 2010]. In this context, the penalized likelihood ratio (PLR) statistic we propose distinguishes itself by assuming merely a smooth underlying density function, thereby imbuing it with the versatility to accommodate a broader spectrum of complex distributions. It also fills in the longstanding gap on nonparametric CUSUM variants in the literature, a phenomenon attributable to the well-known challenges associated with nonparametric inference theories.

Here we introduce a cumulative sum (CUSUM) statistic by employing a series of penalized likelihood ratios, which are constructed through density estimation in the framework of a reproducing kernel Hilbert space (RKHS). To ensure an accurate asymptotic control over the Type-I error, we establish a novel RKHS framework for the non-asymptotic theory of nonparametric density estimation. Under this framework, we calculate the localization rate and establish a minimax lower bound for the detection of change points. Additionally, we propose a method for localizing change points based on the largest likelihood ratio. Due to its involvement of an infinite sum of terms with complex dependence structures, traditional central limit theorems for independent data or data with simple dependence structures may not apply here. Therefore, this calls for the treatment of a non-asymptotic theory, which not only avoids this technical difficulty

but also provides more accurate guarantees and more precise error characterizations in finite sample settings rather than only asymptotically. By leveraging the non-asymptotic bound derived in our study, we determine the convergence rate of our proposed change point estimator.

The remaining sections of this paper are structured as follows. In Section 2, we introduce our hypothesis testing method for change point detection and derive the null asymptotic distribution for ensuring a valid control over type-I error rates. Section 3 presents our model-free change point localization estimate and provides the derivation of its convergence rate. Section 4 is dedicated to the presentation of simulation studies. In Section 5, we explore the application of seizure detection using real-time EEG data. Finally, Section 6 concludes with our final remarks. Proofs and additional numerical results can be found in the Supplementary Materials.

2. Change Point Detection

2.1 Model Setup

The observations are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ for $i = 1, \ldots, n$, where d is the dimension of the observations. There exists a potential change point $\tau \in \{1, \ldots, n\}$, such that the observations up to time τ are independent and identically distributed (i.i.d.) with the distribution P, while the

observations after time τ are also i.i.d. but with a distinct distribution Q.

Let η_p and η_q represent the logit-transformations of the probability density functions for P and Q respectively, that is, the densities of P and Q are $p(\boldsymbol{x}) = e^{\eta_p(\boldsymbol{x})}/\int_{\mathcal{X}} e^{\eta_p(\boldsymbol{x})}d\boldsymbol{x}$ and $q(\boldsymbol{x}) = e^{\eta_q(\boldsymbol{x})}/\int_{\mathcal{X}} e^{\eta_q(\boldsymbol{x})}d\boldsymbol{x}$ respectively. Note that η_p and η_q are identifiable up to a constant, since $\eta + c$ works equivalently for any constant c. Evidently, these transformations allow η_p and η_q to evade the positive and unity constraints. We do not restrict η_p and η_q to a specific parametric distribution family, as distributions in real applications often exhibit complexities, such as a heavy tail or multi-modality. Instead, we impose certain smoothness constraints on η_p and η_q, which are more realistic in many applications. Without loss of generality, we consider the m-th order Sobolev space on $[0,1]^d$, $\mathcal{H} = \{f \in L^2([0,1]^d) \mid f^{(\alpha)} \in L^2([0,1]^d), \quad \forall\, |\alpha| \leq m\}$, where $|\alpha| = \sum_{l=1}^{d} \alpha_l$. When $d = 1$, the associated kernel function is defined as $\mathcal{K}(X_i, X_j) = 1 + (-1)^{m-1} k_{2m}(X_i - X_j)$, where $k_{2m}(x)$ is the $2m$-th order scaled Bernoulli polynomial [Abramowitz and Stegun, 1948]. For $m = 2$, $k_4(x) = \frac{1}{24}((x-0.5)^4 - 0.5(x-0.5)^2 + \frac{7}{240})$ and the corresponding $\mathcal{K}$ is known as the homogeneous cubic spline kernel. When $d > 2$, Novak et al. [2018] showed that the kernel is $\mathcal{K}(X_i, X_j) = \int_{\mathbb{R}^d} [\prod_{l=1}^{d} \cos(2\pi(X_{il} - X_{jl})x_l)]/[1 + \sum_{0<|\alpha|\leq m} \prod_{l=1}^{d} (2\pi x_l)^{2\alpha_l}] d\boldsymbol{x}$.

Our first goal is to validate the presence or absence of a change point

through statistical hypothesis testing. In particular, our hypotheses are $H_0 : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim P$ versus

$$H_1 : \exists \tau \in (n_0, n - n_0) \text{ s.t. } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\tau-1} \sim P \text{ and } \boldsymbol{x}_\tau, \ldots, \boldsymbol{x}_n \sim Q,$$

where $n_0 = \alpha n$ for $\forall \alpha \in (0, 1)$. The introduction of n_0, a common practice in change point analysis, is to avoid a change point close to the boundary of the domain which would result in insufficient statistical power to differentiate between the distributions before and after the change. We will define a test statistic based on a series of penalized likelihood ratio test statistics derived for each possible candidate of the change point τ.

2.2 CUSUM Penalized Likelihood Ratio (CPLR) Test

Penalized likelihood density estimation methods have several significant advantages. Without any strong assumptions about the form of the distribution, they can handle a broad range of data distributions, making them more versatile than traditional parametric methods. The penalized likelihood approach automatically adjusts for the complexity of the model though a tuning parameter in the penalty term that determines the level of smoothness of the estimated density. Penalized likelihood estimators also have desirable asymptotic properties, such as consistency and asymptotic normality, under certain conditions.

Let the penalized likelihoods before and after τ be respectively

$$\ell_p^\tau(\eta) = -\sum_{i=1}^{\tau}\{\eta(\boldsymbol{x}_i) + \int_{\mathcal{X}} e^{\eta(\boldsymbol{x})}d\boldsymbol{x}\} + \tau\frac{\lambda}{2}J(\eta), \tag{2.1}$$

$$\ell_q^{n-\tau}(\eta) = -\sum_{i=\tau+1}^{n}\{\eta(\boldsymbol{x}_i) + \int_{\mathcal{X}} e^{\eta(\boldsymbol{x})}d\boldsymbol{x}\} + (n-\tau)\frac{\lambda}{2}J(\eta) \tag{2.2}$$

with a smoothing parameter $\lambda > 0$ and a penalty function $J(\cdot)$. Penalized likelihood method utilizes the penalty term to balance between bias and variance. In the absence of this penalty, models with great flexibility may succumb to overfitting the data, resulting in increased variance. By imposing a penalty term, overfitting is discouraged through the introduction of an acceptable degree of bias. Let $\widehat{\eta}_p^\tau$ and $\widehat{\eta}_q^{n-\tau}$ be the estimators of η_p and η_q defined as $\widehat{\eta}_p^\tau = \text{argmin}_{\ \eta\in\mathcal{H}}\ell_p^\tau(\eta)$ and $\widehat{\eta}_q^{n-\tau} = \text{argmin}_{\ \eta\in\mathcal{H}}\ell_q^{n-\tau}(\eta)$.

Under H_0, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. from the same distribution, i.e. $\eta_p = \eta_q$. The penalized likelihood becomes

$$\ell_0^n(\eta) = -\sum_{i=1}^{n}\{\eta(\boldsymbol{x}_i) + \int_{\mathcal{X}} e^{\eta(\boldsymbol{x})}d\boldsymbol{x}\} + n\frac{\lambda}{2}J(\eta). \tag{2.3}$$

Similarly, we define the minimizer $\widehat{\eta}_0$ of (2.3) as $\widehat{\eta}_0^n = \text{argmin}_{\ \eta\in\mathcal{H}}\ell_0^n(\eta)$.

We have three penalized likelihood estimators $\widehat{\eta}_p^\tau$, $\widehat{\eta}_q^{n-\tau}$, and $\widehat{\eta}_0^n$ obtained from maximizing penalized likelihoods. Let $\widehat{\eta}_a^b$ be a generic estimator with $(a,b) = (p,\tau), (q,n-\tau)$, and $(0,n)$ corresponding respectively to observations of indices $S_p = \{1,\ldots,\tau\}$, $S_q = \{\tau+1, n\}$, and $S_0 = \{1,\ldots,n\}$. We assume that the domain $\mathcal{X}$ is endowed with a probability distribution $\mathbb{P}$

and $\mathcal{H}$ is a reproducible kernel Hilbert space (RKHS). Denote the positive semi-definite (PSD) kernel function associated with $\mathcal{H}$ by $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Due to the integration in the penalized likelihoods, the Representer Theorem [Wahba, 1990] does not apply here and the exact solution is not computable [Gu, 2013]. We consider an efficient approximation in Gu [2013] by calculating the minimizer of the penalized likelihood functional in the data-adaptive finite-dimensional subspace $\text{span}\{\mathcal{K}(\boldsymbol{x}_i, \cdot)\}_{i \in S_a}$ of $\mathcal{H}$. The function $\eta_a^b \in \text{span}\{\mathcal{K}(\boldsymbol{x}_i, \cdot)\}_{i \in S_a}$ has the form $\eta_a^b(\cdot) = \sum_{i \in S_a} \mathcal{K}(\boldsymbol{x}_i, \cdot) c_{a,i}^b = (\zeta_a^b)^T \boldsymbol{c}_a^b$, where $\zeta_a^b = [\mathcal{K}(\boldsymbol{x}_i, \cdot)]_{i \in S_a}^T$ is the vector of functions obtained from kernel $\mathcal{K}$ with its first argument fixed at $\boldsymbol{x}_i$, and $\boldsymbol{c}_a^b \in \mathbb{R}^{S_a}$ is the coefficient vector. Under the RKHS framework, $J(\eta) = \langle \eta, \eta \rangle_{\mathcal{H}}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. Then we have $J(\eta_a^b) = (\boldsymbol{c}_a^b)^T Q_a^b \boldsymbol{c}_a^b$ where $Q_a^b \in \mathbb{R}^{|S_a| \times |S_a|}$ is the empirical kernel matrix with its (i, j)-th entry of Q_a^b as $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This representation converts the infinite-dimensional minimization of the penalized likelihood with respect to η to the finite-dimensional optimization problem with respect to the coefficient vector $\boldsymbol{c}_a$ by solving

$$\widehat{\boldsymbol{c}}_a^b = \underset{\boldsymbol{c}_a^b}{\text{argmin}} \left\{ -\frac{1}{|S_a|} \mathbf{1}_a^T Q_a^b \boldsymbol{c}_a^b + \int_{\mathcal{X}} \exp\{(\zeta_a^b)^T \boldsymbol{c}_a^b\} d\mathbf{x} + \frac{\lambda}{2} (\boldsymbol{c}_a^b)^T Q_a^b \boldsymbol{c}_a^b \right\} \quad (2.4)$$

where $\mathbf{1}_a$ is an $|S_a| \times 1$ vector of ones. The objective function (2.4) is strictly convex. Its optimization with respect to $\boldsymbol{c}_a^b$ can be performed via a standard convex optimization procedure such as the Newton-Raphson

algorithm. The integrals in (2.4) can be calculated by numerical integration (see Section 7.4.2 in Gu [2013]). When $|S_a|$ is large, the representation (2.4) involves a large number of coefficients, which may lead to numerical instability. To tackle this, one may consider only a subsample of $\{\boldsymbol{x}_i : i \in S_a\}$ as the knots [Kim and Gu, 2004]. As shown in Kim and Gu [2004], the subsample knots can maintain the minimax optimality through properly selected subsample size. The guideline in Gu [2013] is to select a subsample sample size proportional to a theoretically specified fractional power of the full sample size. In general, we have $\widehat{\eta}_a^b = (\zeta_a^b)^T \widehat{\boldsymbol{c}}_a^b$ and recall that $(a, b) = (p, \tau), (q, n-\tau)$, and $(0, n)$ as the penalized likelihood estimate with respect to (2.1), (2.2), and (2.3) correspondingly.

For a given τ, we define the *penalized likelihood ratio* (PLR) as, $PLR^\tau = \ell_p^\tau(\widehat{\eta}_p^\tau) + \ell_q^{n-\tau}(\widehat{\eta}_q^{n-\tau}) - \ell_0(\widehat{\eta}_0)$. We calculate the PLR^τ for each $\tau = n_0, \ldots, n - n_0$ and construct our test statistic as $CPLR = \sum_{\tau=n_0}^{n-n_0} PLR^\tau$. It represents a novel extension of the traditional CUSUM methods to a nonparametric setting under the RKHS framework. Each individual PLR term in the statistic essentially performs kernel density comparisons between two distributions Gretton et al. [2012]. But the cumulative nature of the CPLR statistic poses significant theoretical and computational challenges. It is an infinite series of PLR statistics whose asymptotic distribution cannot

be adequately characterized by traditional theories. So it is necessary to adopt the non-asymptotic theory which provides finite-sample bounds that remain valid across the entire sequence of PLR statistics. This allows a rigorous control of Type-I error rates, uniform convergence guarantees, and precise characterization of localization accuracy.

As demonstrated in later numerical studies, CPLR is particularly effective in detecting small and persistent shifts that other methods might overlook. This makes CPLR ideal for quality control and health monitoring applications where accurate detection of changes is crucial. We next derive the asymptotic null distribution of the CPLR statistic which gives the type-I error control of the detection limits. Also, the CPLR is robust when the exact distribution family of data is not normal or not even known.

2.3 Control the Type-I error

We first introduce some notation related to the true densities. Recall that η_p and η_q are the true logit-transformed densities before and after τ_0 respectively. Under the null hypothesis H_0, η_p equals η_q, denoted collectively by η_0. Now we establish an infinite series of non-asymptotic bounds for PLR statistics. Non-asymptotic convergence rate for nonparametric regression was first proposed in Yang et al. [2017]. However, such a non-asymptotic re-

sult is still lacking for density estimation, which is filled in by our work here. Our work differs significantly from Yang et al. [2017] in the more complex objective, additional normalization, more subtle regularization properties and more complicated information geometry. Then, we derive the asymptotic distribution of our proposed test statistics under the null hypothesis.

For any $\eta, \widetilde{\eta} \in \mathcal{H}$, define

$$\langle \eta, \widetilde{\eta} \rangle = V(\eta, \widetilde{\eta}) + \lambda J(\eta, \widetilde{\eta}), \tag{2.5}$$

where $V(\eta, \widetilde{\eta}) = \mathbb{E}\{\eta(\boldsymbol{X})\widetilde{\eta}(\boldsymbol{X})\}$ with the expectation taken over $\boldsymbol{X}$, and J is a bilinear form corresponding to the roughness penalty. Note that $\mathcal{H}$, endowed with the inner product (2.5), is an RKHS. So $\mathcal{H}$ has a reproducing kernel $K(\cdot, \cdot)$ with $\langle K_{\boldsymbol{x}}, \eta \rangle = \eta(\boldsymbol{y})$ for any $\eta \in \mathcal{H}$, where $K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot)$. Next, we construct the eigensystem under the inner product (2.5).

Assumption 1 *There exist sequences of functions* $\{\xi_p\}_{p=1}^{\infty} \subset \mathcal{H}$ *and non-negative eigenvalues* $\{\rho_p\}_{p=1}^{\infty}$ *such that* $V(\xi_p, \xi_{p'}) = \delta_{p,p'}$, $J(\xi_p, \xi_{p'}) = \rho_p \delta_{p,p'}$, *for all* $p, p' \geq 1$, *and that any* $\eta \in \mathcal{H}$ *can be written as* $\eta = \sum_{p=1}^{\infty} V(\eta, \xi_p)\xi_p$.

Assumption 1 assumes an eigensystem that simultaneously diagonalizes the bilinear forms V and J. Such eigenvalue/eigenfunction assumptions are typical in the smoothing spline literature [Shang and Cheng, 2013,

Xing et al., 2023]. Define $h = \{\sum_{i=1}^{\infty} 1/(1+\lambda\rho_i)^2\}^{-1}$. Note that $h^{-1} = \mathcal{O}(\lambda^{-1/2m}(\log(1/\lambda))^{1/d})$ as λ goes to zero.

Let D, D^2 be the first- and second-order Frechét derivatives of the likelihood functional $\ell_a^b(\eta) = -\sum_{i\in S_a}\{\eta(\boldsymbol{x}_i) + \int_{\mathcal{X}} e^{\eta(\boldsymbol{x})}d\boldsymbol{x}\} + b\frac{\lambda}{2}J(\eta)$. We also denote its minimizer by $\widehat{\eta}_a^b$. Notice that the only difference between the likelihood functionals in (2.1), (2.2) and (2.3) is the number of observations involved. Thus the Fréchet derivatives of (2.1), (2.2) and (2.3) can be derived similarly through replacing t by the corresponding number of observations. For any $\eta, \Delta\eta_1, \Delta\eta_2, \Delta\eta_3 \in \mathcal{H}$,

$$\begin{aligned} D\ell_a^b(\eta)\Delta\eta_1 &= \sum_{i\in S_a}\Delta\eta_1(\boldsymbol{x}_i) + \int_{\mathcal{X}}\Delta\eta_1(\boldsymbol{x})e^{\eta(\boldsymbol{x})}d\boldsymbol{x} + t\lambda J(\eta,\Delta\eta_1) \\ &= \left\langle \sum_{i=1}^{t} K_{\boldsymbol{x}_i} + \mathbb{E}_\eta K_X + tW_\lambda\eta, \Delta\eta_1 \right\rangle, \end{aligned} \tag{2.6}$$

$D^2 l^t(\eta)\Delta\eta_1\Delta\eta_2 = \int_{\mathcal{X}}\Delta\eta_1(\boldsymbol{x})\Delta\eta_2(\boldsymbol{x})e^{\eta(\boldsymbol{x})}d\boldsymbol{x} + t\lambda J(\Delta\eta_1,\Delta\eta_2)$, where the second equality of (2.6) is due to the RK property and that $\int_{\mathcal{X}}\Delta\eta(\boldsymbol{x})e^{\eta(\boldsymbol{x})}d\boldsymbol{x} = \boldsymbol{E}_\eta\Delta\eta_1(X) = \boldsymbol{E}_\eta\langle K_{\boldsymbol{X}}, \Delta\eta_1\rangle = \langle \boldsymbol{E}_\eta K_{\boldsymbol{X}}, \Delta\eta_1\rangle$.

For $\epsilon > 0$ and constant $c_0 > 0$, we define the quantity

$$\begin{aligned} A(h,\epsilon) = &\frac{32\sqrt{6}}{\sqrt{\log(1.5)}}c_K^{-1}c_0^m h^{-(2m-1)/2}\Psi\Big(\frac{1}{2}c_K c_0^{-m}h^{(2m-1)/2}\epsilon\Big) \\ &+ \frac{32\sqrt{6}}{\sqrt{\log(1.5)}}\sqrt{\log(1+\exp(2c_0(c_K h^{(2m-1)/2}\epsilon)^{-1/m})}. \end{aligned}$$

Here $c_K = \sup_{\mathbf{x}\in\mathcal{X}}\sqrt{hK(\boldsymbol{x},\boldsymbol{x})}$, the function $\Psi(r) = \int_0^r \sqrt{\log(1+\exp(x^{-1/m}))}dx$

is from Dudley's entropy integral controlling the upper bound of a concentration inequality, and c_0 satisfies $\log N(\epsilon, \mathcal{G}, \|\cdot\|_{\sup}) \leq c_0 h^{-\frac{2m-1}{2m}} \epsilon^{-1/m}$, where $N(\epsilon, \mathcal{G}, \|\cdot\|)$ is the ϵ-packing number. The existence of such c_0 follows from Shang and Cheng [2013]. For simplicity of notation, we also denote $A(h) = A(h, 2)$ which plays an important role in deriving the non-asymptotic bound for our penalized maximum likelihood estimator $\widehat{\eta}_a$.

Lemma 1 *Under H_0 with Assumption 1, for any positive M, r_n, and h satisfying $c_K^2 \sqrt{M} r h^{-1/2} A(h) \leq 1/2$, we have $\sup_{\eta_0 \in \mathcal{H}} P\big(\|\widehat{\eta}_0 - \eta_0\| \geq \delta_t(M, r)\big) \leq 2\exp(-Mthr^2)$, where $\delta_t(M, r) = 2h^m + c_K(\sqrt{2M} r + (th)^{-1/2})$.*

The proof of Lemma 1 is based on the large deviation bounds between $\widehat{\eta}_0$ and η_0. As shown in our proof, this bound holds uniformly over a "unit ball" in $\mathcal{H}$. It also provides a rigorous framework to quantify the error of non-parametric density estimation in finite samples, a general interest in the density estimation field. Next, we use the score functional defined in (2.6) to approximate the penalized likelihood estimator.

Lemma 2 *Under H_0, for any positive h, r, M satisfying $c_K^2 \sqrt{M} r h^{-1/2} A(h) \leq 1/2$, it holds that $\sup_{\eta \in \mathcal{H}} P_f(\|\widehat{\eta}_0 - \eta_0 - (\frac{1}{t}\sum_i K_{\boldsymbol{X}_i} - \mathbb{E}[K_{\boldsymbol{X}_i}] - W_\lambda \eta_0)\| \geq \gamma_n(M, r)) \leq 2\exp(-Mthr^2)$. where $\gamma_t(M, r) = c_K^2 \sqrt{M} r h^{-1/2} A(h) \sigma_t(M, r)$, with $\sigma_t(M, r)$ defined in Lemma 1.*

Lemma 2 allows us to approximate our CPLR statistic by a series of score functions. We then use Lemmas 1 and 2, with t replaced by $n, n-\tau, \tau$, to derive the asymptotic distribution of CPLR in the following Theorem. The proofs for the lemmas are in the supplementary materials.

Theorem 1 *Let $m \geq 1$ and Assumption 1 be satisfied. Then under H_0,*

$$\frac{2CPLR - \operatorname{tr}(\Delta)h^{-1}}{\sqrt{2\operatorname{tr}(\Delta)h^{-1}}} \xrightarrow{d} N(0,1),\ n \to \infty,$$

where $\Delta = \sum_{\tau=n_0}^{n-n_0} \boldsymbol{e}^{\tau}(\boldsymbol{e}^{\tau})^T$, $\boldsymbol{e}_i^{\tau} = \sqrt{(n-\tau)/\tau}$ for $1 \leq i \leq \tau$, $\boldsymbol{e}_i^{\tau} = -\sqrt{\tau/(n-\tau)}$ for $\tau+1 \leq i \leq n$.

The theorem is built on the functional expansion of the likelihood functional. Its complete proof is in the supplementary materials. The behavior of the CPLR in its asymptotic form is determined by a series of eigenvalues, denoted by $\{\rho_\nu\}_{\nu=1}^{\infty}$. It also possesses an excellent property called the Wilks phenomenon, that is, the asymptotic distribution is not influenced by nuisance parameters that are not of direct interest but can impact the estimation of the parameters of interest. Such independence from nuisance parameters facilitates more accurate statistical inference and widens its applicability across diverse statistical modeling contexts. In practice, the constant h can be approximated using the empirical eigenvalues of the kernel matrix. Therefore, our decision rule for identifying a change point is

$$\Phi(\alpha) = \mathbb{1}(|2n \cdot PLR_{n,\lambda} - \theta_\lambda| \geq z_{1-\alpha/2}\sqrt{2}\sigma_\lambda) \tag{2.7}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

3. Change point localization

In change point analysis, upon the rejection of the null hypothesis, the next task is to determine the exact location of the change point. We adopt the penalized likelihood estimation approach to this task. The essence of this method is captured in the comprehensive likelihood function:

$$\ell^\tau(\widehat{\eta}_p^\tau, \widehat{\eta}_q^{n-\tau}) = \ell_p^\tau(\widehat{\eta}_p^\tau) + \ell_q^\tau(\widehat{\eta}_q^{n-\tau}) \tag{3.8}$$

where $\widehat{\eta}_p^\tau$ and $\widehat{\eta}_q^{n-\tau}$ represent the maximum likelihood (or minimum negative log-likelihood) estimators, with their values contingent upon τ. We define our maximum likelihood estimator for τ as:

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmin}}\, \ell^\tau(\widehat{\eta}_p^\tau, \widehat{\eta}_q^{n-\tau}). \tag{3.9}$$

3.1 Covergence rate of the estimation of τ

In the context of the change point localization problem, it becomes essential to quantify the precision of the estimation. The primary challenge in

localization arises from the fact that it involves the selection of the maximum value from a sequence of penalized likelihood ratio (PLR) estimates. To address this inherent challenge, we have identified a relatively straightforward and intuitive approach for establishing a connection between the complexity of the density estimation problem and the task of achieving consistent change point localization. This relationship provides valuable insights into the problem. In particular, under $H_1 : P \neq Q$, the true underlying densities are represented by their respective logit-transformations η_p^τ and $\eta_q^{n-\tau}$ and the logit-transformation of the underlying true density for the whole data by $\eta_0 = \log(\frac{\tau_0}{n}e^{\eta_p} + \frac{n-\tau_0}{n}e^{\eta_q})$. Here, the function η_p^τ equals η_p, for $\tau \in (0, \tau_0)$, and $\log(\frac{\tau}{\tau_0}e^{\eta_p} + \frac{n-\tau_0}{\tau_0}e^{\eta_q})$, for $\tau \in (\tau_0, n)$. The function $\eta_q^{n-\tau}$ equals $\log(\frac{\tau_0-\tau}{n-\tau}e^{\eta_p} + \frac{n-\tau_0}{n-\tau}e^{\eta_q})$ for $\tau \in (0, \tau_0)$, and η_q for $\tau \in (\tau_0, n)$.

The following theorem, based on the non-asymptotic results in Lemmas S.1 and S.2, characterizes the asymptotic accuracy of $\widehat{\tau}$.

Theorem 2 *With Assumption 1 and* $h^{m-1/2} < \min\{1/(6c_K), 1/(4c_K e)\}$, $h^{m+1/2} \leq 8c_K/5$, *we have* $|\widehat{\tau} - \tau_0|/n = O((nh)^{-1} + \lambda)$.

The detailed proof for Theorem 2 is in the supplementary materials. This theorem establishes the convergence rate of $|\widehat{\tau} - \tau_0|/n$ to zero, and this rate is no slower than $(nh)^{-1} + \lambda$. For the m-th order Sobolev space on $[0, 1]^d$, the optimal orders of λ and h are respectively $n^{-\frac{2m}{2m+d}}$

and $O(n^{-1/(2m+d)})$, balancing the bias and variance to achieve the minimax rate $n^{-2m/(2m+d)}$ for density estimation and localization error. Remarkably, we observe that this rate aligns with the minimax distinguishability rate for testing the comparison of two smooth density functions [Xing et al., 2023]. Additionally, it is noteworthy that this localization rate surpasses the optimal sample complexity required for estimating all the underlying densities.

4. Simulation Study

We now illustrate the simulated outcomes of the suggested approach utilizing a variety of univariate and multivariate distributions. We first appraise the proposed hypothesis test through the empirical size and power derived from 200 replicates. Following that, we bring forth the outcomes of numerical simulations exploring the convergence of the change point estimator across diverse sequence lengths. The mean estimation error is reported with the standard deviations of our proposed estimate duly presented.

We incorporate methodologies such as the pruned exact linear time (PELT) method [Killick et al., 2012], the Bayesian analysis of change point problems (BCP) [Barry and Hartigan, 1993], the nonparametric multiple change point analysis of multivariate data (ECP) [Matteson and James,

2014], the graph-based change-point detection [Chen and Zhang, 2015], KolmogorovSmirnov detector (KSD) [Padilla et al., 2021a,b]. However, we should note that implementing a univariate change point process to each margin, as executed by the changepoint, cpm, and bcp packages, will not detect the changes. We mainly focus on multivariate settings (Settings 1-4). We put the univaraite settings (Setting 3-7) as a reference in our Supplementary material.

For multivariate signals, we take into account changes in the distribution of covariance structures. Similar to the single-variable scenarios, we let the sequence length, denoted by n, range from 500 to 2000, and we position the change point at $n\theta$, where θ is randomly sampled from a uniform distribution within the interval (0.2, 0.8). We perform 200 replications to compute the empirical power and size. In this context, we generate covariates $X \in \mathbb{R}^d$ from two distinct multivariate distributions, each with two different settings for the dimension d: specifically, $d = 2$ and $d = 10$.

Setting 1 (Change in Covariance): X_i, $i = 1, \ldots, n$ are i.i.d. generated as $X_i \sim N_d(0, I_d)$ if $i \leq n\theta$ and $N_d(0, \Sigma_\rho)$ if $i > n\theta$, where I_d is the identity matrix and Σ_ρ has diagonal entries one and off-diagonal entries ρ set as 0.2.

Setting 2 (Change in Skewness): $X_i \in \mathbb{R}^d, i = 1, \ldots, n$ have different skewnesses before and after change point. In particular, $X_{ij} \sim N(0, 1)$ if

$i \leq n\theta$ and $= Z_{ij}^2 - 1$ if $i > n\theta$ for $j = 1, \ldots, d$, where $Z_{ij} \sim N_d(0, I_d)$.

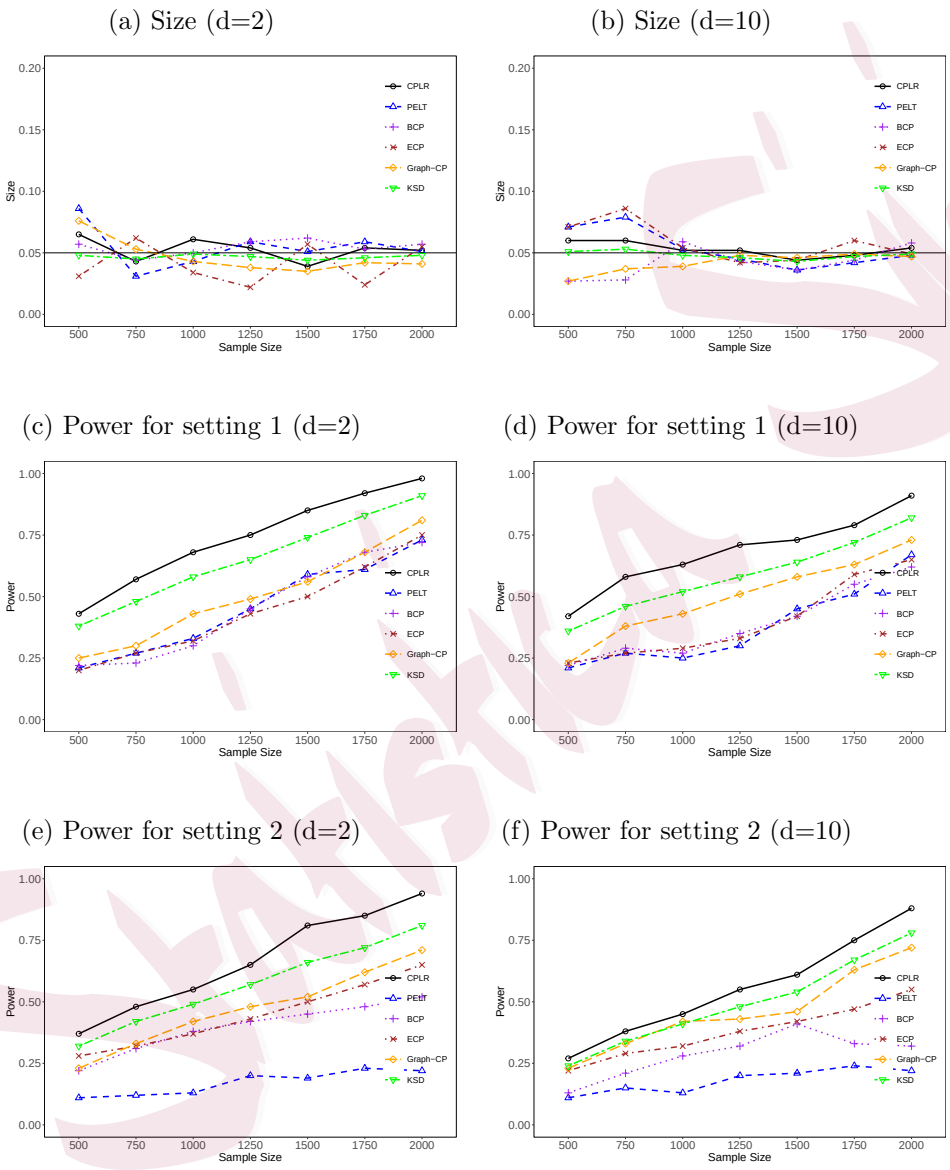

Figure 1: Empirical power and size for setting 1-2. Here the design matrix is generated from the multivariate normal distribution.

In the simulation settings explored, we establish the significance level at 0.05, focusing on this value as our threshold for statistical significance. The results are in Figure 1(a-b). We observe the empirical size of the proposed test under different dimensional, finding it to be approximately 0.05 across different sequence lengths. This consistency highlights the test's reliability in maintaining the desired significance level, regardless of the specific length of the sequence under examination.

We proceeded to assess the empirical power performance of our method under setting 1-2, employing different dimensions ($d = 2$ and $d = 10$) to provide a comprehensive evaluation. This investigation is depicted in Figure 1(c,d), where we observe that our proposed approach consistently outperforms our competitors in terms of power. This outcome vividly illustrates the test's remarkable sensitivity in detecting shifts in correlation among covariates within the dataset. Such sensitivity is crucial when it comes to identifying subtle alterations in data patterns, making our method exceptionally well-suited for scenarios where detecting nuanced changes is of paramount importance. Figure 1(e,f) further shows the versatility of our proposed test by showcasing its capacity to detect changes in skewness within the data. Moreover, it is worth noting that as the dimensionality of the data increases, we observe a corresponding increase in the difficulty

of detection. This observation highlights the real-world adaptability of our method across various conditions and shows its capability to maintain statistical rigor. In conclusion, these simulation results serve as compelling evidence of the robustness and practicality of our proposed test. They illustrate its adaptability to changing conditions and its unwavering commitment to statistical precision. The ability to fine-tune signal strength and observe corresponding variations in test power offers valuable insights into the test's behavior under a range of hypothetical scenarios. This, in turn, enhances its potential for application in diverse fields where precise change detection is a paramount requirement.

In Tables 1 and 2, we present a comprehensive analysis of the convergence behavior of our proposed change point estimate as it pertains to various dimensions with CPLR, and KSD. It is evident that as the sample size expands, CPLR has smaller estimation error and faster convergence rate comparing with KSD. This observation shows the efficacy of our approach, showcasing its ability to provide increasingly accurate change point estimates as more data becomes available. Furthermore, as the dimensionality of the dataset increases, we notice a gradual slowdown in the rate of convergence. The data presented in this table serves as robust empirical evidence, reinforcing the dependability of our change point estimate. It

	d=2		d=10	
n	CPLR	KSD	CPLR	KSD
500	$0.056_{(0.065)}$	$0.089_{(0.078)}$	$0.112_{(0.136)}$	$0.145_{(0.152)}$
750	$0.060_{(0.047)}$	$0.083_{(0.059)}$	$0.092_{(0.111)}$	$0.125_{(0.134)}$
1000	$0.042_{(0.043)}$	$0.067_{(0.055)}$	$0.083_{(0.094)}$	$0.116_{(0.118)}$
1250	$0.036_{(0.040)}$	$0.054_{(0.048)}$	$0.061_{(0.072)}$	$0.089_{(0.095)}$
1500	$0.029_{(0.028)}$	$0.047_{(0.035)}$	$0.056_{(0.060)}$	$0.078_{(0.082)}$
1750	$0.016_{(0.021)}$	$0.032_{(0.028)}$	$0.038_{(0.032)}$	$0.055_{(0.045)}$
2000	$0.006_{(0.005)}$	$0.018_{(0.015)}$	$0.016_{(0.014)}$	$0.028_{(0.023)}$

Table 1: Convergence Analysis for Multivariate Setting 1. Table lists the averaged relative distance between change point estimates and the underlying truth when n ranges from 500 to 2000. CPLR shows superior localization accuracy compared to KSD.

demonstrates that our method's accuracy experiences substantial improvement with larger sample sizes, thereby fortifying the reliability of our approach in the context of detecting changes in the underlying data structure. This insight is invaluable for practitioners and researchers alike, as it highlights the practical advantages of utilizing our methodology in situations where data quality and precision are paramount concerns.

	d=2		d=10	
n	CPLR	KSD	CPLR	KSD
500	$0.061_{(0.062)}$	$0.094_{(0.075)}$	$0.122_{(0.113)}$	$0.156_{(0.128)}$
750	$0.049_{(0.054)}$	$0.072_{(0.067)}$	$0.089_{(0.086)}$	$0.118_{(0.101)}$
1000	$0.043_{(0.045)}$	$0.065_{(0.058)}$	$0.067_{(0.072)}$	$0.095_{(0.089)}$
1250	$0.034_{(0.034)}$	$0.051_{(0.042)}$	$0.051_{(0.053)}$	$0.074_{(0.068)}$
1500	$0.024_{(0.032)}$	$0.038_{(0.041)}$	$0.031_{(0.030)}$	$0.049_{(0.043)}$
1750	$0.020_{(0.016)}$	$0.031_{(0.024)}$	$0.027_{(0.026)}$	$0.040_{(0.034)}$
2000	$0.006_{(0.008)}$	$0.015_{(0.012)}$	$0.013_{(0.012)}$	$0.025_{(0.019)}$

Table 2: Convergence Analysis for Multivariate Setting 2. Table lists the averaged relative distance between change point estimates and the underlying truth when the sequence length ranges from 500 to 2000. CPLR shows superior localization accuracy compared to KSD.

5. Real Data Analysis

Epileptic seizures, characterized by sudden, uncontrolled electrical disturbances in the brain, are the most prevalent neurological dysfunction in newborns, necessitating immediate medical intervention. Among full-term babies, the observed incidence of neonatal seizures is approximately 3 per 1000, and increases to 50 per 1000 in specific cases [van Rooij et al., 2013].

The clinical diagnosis of neonatal seizures can be highly unreliable. This is primarily due to the subtle nature of the clinical signs, which can be ex-

ceedingly difficult to detect. Traditional methods of detection often fail to identify these complex signs, leading to delays in medical intervention. Most current change point detection tools fall short in addressing these intricacies. They either make assumptions of a multivariate Gaussian distribution or are designed as nonparametric methods focusing on a single feature. Both of these approaches miss critical aspects of the problem's complexity. Our proposed method innovatively addresses these shortcomings by considering both non-Gaussian and multivariate features, filling the existing gap in seizure detection techniques. By doing so, we have achieved significant improvements in both the detection power and accuracy of change point estimation. These enhancements reflect a critical advancement in neonatal seizure detection, potentially leading to more timely and effective interventions. Our approach not only demonstrates a clear methodological innovation but also carries profound implications for the healthcare of newborns at risk of neurological dysfunction.

In this study, we selected two features that are shown to be most predictable for seizure as in Greene et al. [2008]. The first feature is Root mean squared EEG amplitude (RMS Amp) which is an estimate of the cerebral function monitor (CFM) output. CFM is perhaps the most widely used tool for the detection and diagnosis of seizures in the neonatal ICU. The second

feature is the line (curve) length proposed in Katz [1988] as a potential feature for epileptic seizure detection in adults. To validate the independence assumption critical to our statistical modeling framework, we conducted an empirical assessment using the Ljung-Box test [Ljung and Box, 1978], which detects autocorrelation in time series by examining whether any group of autocorrelations differs significantly from zero. As shown in Supplementary S.2, the empirical evidence from our Ljung-Box test analysis conclusively demonstrates that down-sampling from 2000Hz to 100Hz effectively eliminates temporal autocorrelation while maintaining clinically sufficient temporal resolution.

The data in this study were collected at a high sampling rate of 2000Hz, implying that there is a mere 1/2000-second time gap between consecutive sampling points. This dataset comprises a total of 137 marked seizure records, each with a seizure duration of approximately 5 seconds. In clinical practice, achieving detection accuracy within a 1-second timeframe is deemed sufficiently reliable for accurate clinical diagnosis. To optimize the efficiency of detection, we employed a down-sampling technique, reducing the data to a more manageable 100Hz (with a 1/100-second time gap between consecutive sampling points). For our approach, we employed a window size of 4 seconds and moved the window every 0.01 seconds to calculate

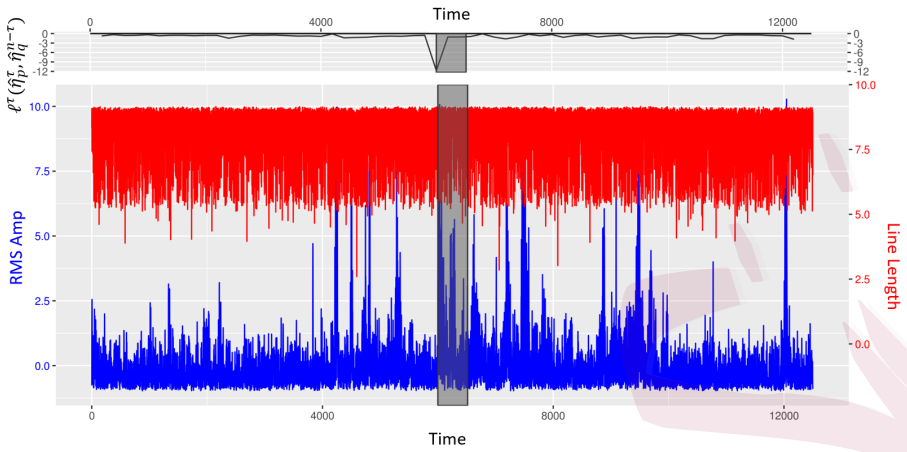

Figure 2: This plot shows a record with p-value = 0.0013 from our proposed CPLR test. The upper panel displays statistics defined in (3.8) for change point localization, which is based on two-dimensional covariates, namely RMS Amp and Line Length. In the lower panel, we provide a graphical representation of the signal characteristics, showcasing the RMS amplitude (in blue) and line length (in red). Additionally, we highlight the presence of seizures on-site by shading the corresponding regions in both panels.

our test statistics denoted by PLR^{τ}. Subsequently, the Cumulative Sum (CUSUM) statistics, referred to as CPLR, were computed for values of τ ranging from 0.02 seconds to the total length of each record denoted by T.

Figure 2 offers a visual representation of the testing and estimation results achieved using our proposed detection method on one of the records from the dataset. In the lower panel of Figure 2, the original features appear noisy. Our change point estimate is determined as the location corresponding to the smallest p-value, which is clearly depicted as a spike

in the upper panel of Figure 2. Remarkably, this estimate closely aligns with the commencement of the seizure, thus serving as a compelling validation of the accuracy and efficacy of our proposed estimation method.

Method	RMS Amp	Line Length	RMS Amp & Line Length
CPLR	**0.81**	**0.76**	**0.94**
PELT	0.52	0.34	0.61
BCP	0.71	0.51	0.82
ECP	0.62	0.61	0.75
Graph-CP	0.61	0.41	0.76

Table 3: Empirical powers evaluated for CPLR, PELT, BCP, ECP, and Graph-CP.

In our research, we rigorously applied our proposed method alongside several competitors, including PELT, BCP, ECP, and Graph-CP, to a dataset comprising 137 meticulously annotated seizure records. Our evaluation process was twofold: we initially assessed the performance of each covariate independently and subsequently examined their combined impact when considered jointly. The comprehensive results, as detailed in Table 3, unequivocally demonstrate the superiority of our proposed CPLR method in terms of power. Whether applied to covariates individually or in combination, our method consistently exhibited the highest power. Remarkably, when both covariates were considered simultaneously in a joint analysis, we

observed a substantial enhancement in power by an impressive margin of 16%. This significant power boost vividly shows the critical importance of employing multivariate change point detection algorithms, as it enables us to harness the rich interplay of multiple covariates to achieve more accurate and effective change point detection in complex datasets.

	Method	RMS Amp	Line Length	RMS Amp & Line Length
$\widehat{r}(s)$	CPLR	0.528 (0.128)	0.813 (0.521)	0.151 (0.192)
	KSD	0.982 (0.245)	1.256 (0.634)	0.889 (0.318)

Table 4: Average relative distances (std. dev.) between change point estimates using one or two features and the true seizure starting points for CPLR and KSD methods.

Furthermore, we conducted an in-depth analysis by comparing the relative distance $\widehat{r}(s)$, in the bivariate case with the univariate cases. When we exclusively utilized either RMS Amp or Line Length as covariates, the average value of $\widehat{r}$ across the 137 records exceeded the threshold of 0.05. However, when we incorporated both of these two features simultaneously, the detection error exhibited a remarkable reduction, as highlighted in Table 3. On average, our proposed estimation method achieved a proximity of approximately 0.15 seconds to the true onset time of the seizure, demonstrating its impressive accuracy and precision in pinpointing critical events within the data. When using both RMS amplitude and line length features

jointly, CPLR achieves an average relative distance of 0.151 seconds compared to KSDs 0.889 seconds, representing a substantial improvement in localization precision. This finding shows the substantial benefits of employing a multivariate approach that considers multiple covariates, yielding superior results compared to a univariate analysis.

6. Discussion

Our proposed method has contributed significantly by establishing a foundation for deriving the null asymptotic distribution within the nonparametric framework. Additionally, it has facilitated precise localization through the characterization of convergence rates. This feature is particularly valuable as it ensures accurate pinpointing of change points within the data. The capacity of nonparametric multivariate change point detection to tap into the richness of multivariate data, while simultaneously enhancing sensitivity, positions it as an indispensable tool for researchers, analysts, and practitioners alike. It empowers them to unearth concealed insights, identify anomalies, and make well-informed decisions in scenarios characterized by complex, interrelated datasets. Extending our RKHS-based framework to temporally dependent data represents a natural and valuable direction for future work, with a potential approach being the time-dependent densi-

ties introduced by Song et al. [2009]. In addition, we focus on single change point detection in this paper, the proposed CPLR methodology can be naturally extended to multiple change points through a binary segmentation framework [Fryzlewicz, 2014, Kovács et al., 2023].

Our localization result in Theorem 2 can be refined with the incorporation of a "jump size" defined as $\Delta^2 = \|\eta_p - \eta_q\|_H^2 = \int(\eta_p - \eta_q)^2 d\mu + \lambda J(\eta_p - \eta_q)$. Building on Theorem 2, the localization error rate can be refined as $|\widehat{\tau} - \tau_0|/n = O(\Delta^{-2} \cdot \max\{(nh_p)^{-1}, (nh_q)^{-1}\} + \lambda)$. This rate suggests that a larger distributional change can lead to a much more precise localization while small jumps can make localization more difficult. This explains why clear seizure events in our EEG application are localized within 0.15 seconds while subtle changes require longer observation windows.

As a nonparametric approach, our method focuses only on change points not close to the boundaries. This is a common limitation for nonparametric methods, since they generally require sufficient data on both sides of the change point to deliver accurate estimation of the distributions or distributional parameters on both sides. On the other hand, parametric methods can often bypass this requirement through parametric assumptions and thus be used to detect a change point near the boundaries.

Our theoretical results cannot be easily extended to high-dimensional

cases ($d > n$). In high dimensions, the kernel matrix conditioning deteriorates, density estimation rates become impractical at $O(n^{-2m/(4m+d)})$, and the non-asymptotic bounds in Lemma 1 require sample complexity that scales exponentially with the dimension. Following Raskutti et al. [2012] on sparse additive models, a potential extension would require imposing sparsity assumptions where only $s \ll d$ coordinates are relevant for change point detection. Through decomposing the d-dimensional RKHS into univariate components with tensor product kernels and applying ℓ_1 penalties to enforce sparse coefficient selection, a convergence rate of $|\widehat{\tau} - \tau_0|/n = O((ns\log d)^{-1} + \lambda\sqrt{s\log d/n})$ may be achievable under the condition $s\log d \ll n$. However, such extensions would face significant computational challenges requiring variable selection, sparse optimization solvers, and careful hyperparameter tuning. Numerically, degraded detection power and localization accuracy may occur in the high dimensional setting. Substantial methodological innovations incorporating structural assumptions about sparsity and additive decomposability would be necessary.

Supplementary Material

The online Supplementary Material contains the proofs of the main theorems as well as some auxiliary results.

Acknowledgements

The authors are grateful of the Assoicate Editor and two anonymous reviewers for their insightful comments that have significantly improved the paper. Xing's research was partially supported by U.S. NSF Grant DMS-2311215. Shang's research was supported by NSF DMS 1821157.

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162), 2019.

Daniel Barry and John A Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

Edward G Carlstein, David Siegmund, et al. Change-point problems. IMS, 1994.

Alain Celisse, Guillemette Marot, Morgane Pierre-Jean, and GJ Rigaill. New efficient algorithms for multiple change-point detection with reproducing kernels. *Computational Statistics & Data Analysis*, 128:200–220, 2018.

Uikyu Chae, Hyogeun Shin, Nakwon Choi, Mi-Jung Ji, Hyun-Mee Park, Soo Hyun Lee, Jiwan Woo, Yakdol Cho, Kanghwan Kim, Seulkee Yang, et al. Bimodal neural probe for highly

co-localized chemical and electrical monitoring of neural activities in vivo. *Biosensors and Bioelectronics*, 191:113473, 2021.

Hao Chen and Nancy Zhang. Graph-based change-point detection. *Annals of Statistics*, 43: 139–176, 2015.

Jie Chen and Arjun K Gupta. Parametric statistical change point analysis: with applications to genetics, medicine, and finance. 2012.

Zhanshou Chen and Zheng Tian. Modified procedures for change point monitoring in linear models. *Mathematics and computers in simulation*, 81(1):62–75, 2010.

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.

Lynna Chu and Hao Chen. Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Annals of Statistics*, 47:382–414, 2019.

John HJ Einmahl and Ian W McKeague. Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290, 2003.

Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12(2):4440–4486, 2018.

BR Greene, S Faul, WP Marnane, G Lightbody, I Korotchikova, and GB Boylan. A comparison of quantitative eeg features for neonatal seizure detection. *Clinical Neurophysiology*, 119 (6):1248–1261, 2008.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

Zaid Harchaoui and Olivier Cappé. Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE, 2007.

Michael J. Katz. Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, 18(3):145 – 156, 1988.

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.

Young-Ju Kim and Chong Gu. Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical*

Methodology), 66(2):337–356, 2004.

Solt Kovács, Peter Bühlmann, Housen Li, and Axel Munk. Seeded binary segmentation: a general methodology for fast and optimal changepoint detection. *Biometrika*, 110(1):249–256, 2023.

Marc Lavielle and Gilles Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

Brian Litt and Javier Echauz. Prediction of epileptic seizures. *The Lancet Neurology*, 1(1): 22–30, 2002.

Greta M Ljung and George EP Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

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.

Erich Novak, Mario Ullrich, Henryk Woźniakowski, and Shun Zhang. Reproducing kernels of sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Analysis and Applications*, 16(05):693–715, 2018.

Carlos Misael Madrid Padilla, Haotian Xu, Daren Wang, Oscar Hernan Madrid Padilla, and Yi Yu. Change point detection and inference in multivariable nonparametric models under mixing conditions. *NeurIPS 2023*, 2023.

Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric change point analysis. *Electronic Journal of Statistics*, 15(1), 2021a.

Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2021b.

Aleksey S Polunchenko and Alexander G Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and computing in applied probability*, 14:649–684, 2012.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(13):389–427, 2012. URL `http://jmlr.org/papers/v13/raskutti12a.html`.

Gaetano Romano, Idris Eckley, Paul Fearnhead, and Guillem Rigaill. Fast online changepoint detection via functional pruning cusum statistics. *arXiv preprint arXiv:2110.08205*, 2021.

Zuofeng Shang and Guang Cheng. Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638, 2013.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th annual international conference on machine learning*, pages 961–968, 2009.

Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2): 1–11, 2002.

Hazem Toutounji and Daniel Durstewitz. Detecting multiple change points using adaptive regression splines with application to neural recordings. *Frontiers in neuroinformatics*, page 67, 2018.

Linda G.M. van Rooij, Lena Hellstrm-Westas, and Linda S. de Vries. Treatment of neonatal seizures. *Seminars in Fetal and Neonatal Medicine*, 18(4):209 – 215, 2013.

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

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.

Xin Xing, Zuofeng Shang, Pang Du, Ping Ma, Wenxuan Zhong, and Jun S Liu. Minimax nonparametric multi-sample test under smoothing. *Statistica Sinica*, 2023.

Yun Yang, Zuofeng Shang, and Guang Cheng. Non-asymptotic theory for nonparametric testing. *arXiv preprint arXiv:1702.01330*, 2017.

A. Zeileis, F. Leisch, C. Kleiber, and K. Hornik. Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, 20(1):99–121, 2005.

Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. 2014.