Computational Statistics and Data Analysis 175 (2022) 107545

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

A unified framework on defining depth for point process using function smoothing

Zishen Xu, Chenran Wang, Wei Wu [*]

Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, United States of America

ARTICLE INFO

Article history:
Received 22 July 2021
Received in revised form 17 May 2022
Accepted 9 June 2022
Available online 15 June 2022

Keywords:
Statistical depth
Point process
Function smoothing
Proper metric
Spike trains

ABSTRACT

The notion of statistical depth has been extensively studied in multivariate and functional data over the past few decades. In contrast, the depth on temporal point process is still under-explored. The problem is challenging because a point process has two types of randomness: 1) the number of events in a process, and 2) the distribution of these events. Recent studies proposed depths in a weighted product of two terms, describing the above two types of randomness, respectively. Under a new framework through a smoothing procedure, these two randomnesses can be unified. Basically, the point process observations are transformed into functions using conventional kernel smoothing methods, and then the well-known functional *h*-depth and its modified, center-based version are adopted to describe the center-outward rank in the original data. To do so, a proper metric is defined on the point processes with smoothed functions. Then an efficient algorithm is provided to estimate the defined "center". The mathematical properties of the newly defined depths are explored and the asymptotic theories are studied. Simulation results show that the proposed depths can properly rank point process observations. Finally, the new methods are demonstrated in a classification task using a real neuronal spike train dataset.

1. Introduction

The statistical depth is a method to indicate the centrality of a data point with respect to a data cloud. It can provide a center-outward structure that can be used to understand the empirical distribution, similar to the empirical quantiles in univariate data. The concept of depth was firstly proposed by (Tukey, 1975) to handle multivariate data. Since then the notion of depth has been widely studied by mathematicians and statisticians. Different forms of depth were proposed and analyzed, based on different usages and criteria. The types of data that those depths were mainly applied to were multivariate data and functional data. Some well-known multivariate depths include the half-space depth (Tukey, 1975), the convex hull peeling depth (Barnett, 1976), the Oja depth (Oja, 1983), the simplicial depth (Liu et al., 1990), and the Mahalanobis depth (Liu and Singh, 1993). A study by Zuo and Serfling (2000) discussed desirable properties of multivariate depth, which include affine invariance, maximality at the center, monotonicity relative to the deepest point, and vanishing at infinity. As a generalization of multivariate data from finite dimension to infinite dimension, functional data received a lot of attentions recently and functional depths were also well studied. To name a few, Cuevas et al. (2007) proposed the *h*-depth, which applied the Gaussian kernel and $\mathbb{L}^2$ norm to form the depth. Cuesta-Albertos and Nieto-Reyes (2008)

[*] Corresponding author.
E-mail addresses: xufrankbear@gmail.com (Z. Xu), wwu@fsu.edu (W. Wu).

https://doi.org/10.1016/j.csda.2022.107545
0167-9473/© 2022 Elsevier B.V. All rights reserved.

extended the idea of half-space depth and defined the random Tukey depth. In 2009, the band depth and modified band depth were introduced by López-Pintado and Romo (2009), which were very commonly used in functional data problems. Similar to the work by Zuo and Serfling (2000), Nieto-Reyes et al. (2016); Gijbels and Nagy (2017) examined the desirable properties of functional depth, including distance invariance, maximality at center, strictly decreasing with respect to the deepest point, upper semi-continuity, and receptivity to convex hull width across the domain and continuity.

As a special type of data, observation from an orderly temporal point process is made up by an ascending sequence of event times. Such observation contains two types of randomness: the number of events and the time locations of these events. If we treat each observation as a vector of the event times, then the dimension of this vector will be a random variable. Given this dimension, this vector will also be random with ascending entries. We point out that the study on the notion of depth of point process is relatively new. So far the only existing work in this area was done in Liu et al. (2017), where the depth structure was built in two steps: For an observation s, 1) estimating the probability of getting the number of events $P(|s|)$; 2) given the number of events, compute the conditional depth $D(s \mid |s|)$. The estimation of probability was done through a normalized Poisson mass function and the conditional depth adopted the Mahalanobis depth for multivariate data. The final depth was the multiplication of these two with a weight power r: $D(s) = P(|s|)^r D(s \mid |s|)$. This depth structure involves both types of randomness and satisfies good mathematical properties such as invariant to time-shift and linear transformation, monotone on rays, and upper semi-continuous. However, the combination of the randomnesses on the number of events and the event time distribution is not natural and the impact of the number of events on the final depth is adjusted by the hyper-parameter r. The selection of the hyper-parameter could be tricky because an inappropriate value will make one of the two randomnesses dominates the depth value. In addition, the two-step procedure deals with the number of events and the event time independently. A more desirable method should be able to combine both steps in one framework, where the two types of randomness can be measured at the same time.

To deal with these issues, we propose another approach to process the point process observation: a “transformation” through smoothing method (Wand and Jones, 1994). The idea is to smooth the observed point process sequence with an appropriate kernel function. Through this operation, the event time vector will be replaced by a function curve. In Section 2.1.2, we will see that the point process observation and the function curve are one-to-one matched if the kernel function used for smoothing satisfies certain mild conditions. According to this bijective relation, we are able to apply the methods on functional data to the point process observations, such as metrics on functions and functional depths. In this way, both types of randomness are taken into account under one framework. For individual event time, it determines the location of the kernel function curve. Because we take the sum of the kernel functions in the smoothing procedure, a larger number of events will enhance the smoothed curve vertically. In addition, irrespective of the number of entries in the vector, its smoothed curve will be just one function. That is, vectors with different dimensions can be naturally compared. Therefore, this approach will not suffer the same issues in the framework proposed by Liu et al. (2017) and it has the advantage to utilize existing depth methods for functional data.

The rest of this manuscript is organized as follows. In section 2, we will introduce the details of transforming a vector to a smoothed curve. We will also propose a proper metric for the space of point process observations, followed by a discussion of the properties for this metric. Based on this metric, we will define new depth methods for the point process observations. In section 3, we will examine the asymptotic theory of the proposed depths. In section 4, we will apply our methods to simulations and experimental datasets to validate the effectiveness of the new framework. Finally in section 5, a summary of this paper will be given, followed by the future work. All mathematical proofs and algorithmic details are given in appendices in the supplementary material.

2. Methods

2.1. *Equivalent representation via function smoothing*

Our goal is to rank point process observations on a finite interval $[0, T]$ via a kernel smoothing method. We will at first introduce the type of smoothing kernels for this purpose.

2.1.1. *Kernel functions*

The basic idea of kernel smoothing on an orderly temporal point process is to assign a kernel to each observed point and then sum over all the assigned kernels to get a smooth function. Often the kernel is a probability density function of a given distribution such as a Gaussian kernel. The kernel function will depend on the time interval $[0, T]$ for the point process and we denote it as $K(\cdot; T)$. In general, we propose to use any kernel function which satisfies the following four basic requirements. These requirements are needed in order to achieve good mathematical properties for the new depth function:

1. Continuous and non-negative: $K(\cdot; T)$: $(-\infty, \infty) \to [0, \infty)$ is continuous;
2. Positive at zero: $K(0; T) > 0$;
3. Linear independence with shifting: for any n ($n \in \mathbb{N}$) different values $t_1 < t_2 < \cdots < t_n$, we have: $\sum_{i=1}^{n} \alpha_i K(u - t_i; T) \doteq 0$ for any $u \in [0, T]$ $\iff$ $\alpha_1 = \cdots = \alpha_n = 0$;
4. Scale invariance: for any $x \in [0, T]$ and constant $\alpha > 0$, $K(\alpha u; \alpha T) \doteq K(u; T)$;

We define a kernel function to be "proper" if it satisfies the above four conditions. To have an example of a proper function, we can consider the following Gaussian kernel function:

$$K_G(u;T) = c_1 e^{-\frac{c_2}{T^2}u^2}, \tag{1}$$

where c_1, c_2 are two positive hyper-parameters. One can control c_1 to scale the kernel function value to a desired range, and control c_2 to change the width of the kernel function, thus adjust the smoothness of the kernel smoothing output. The following lemma shows that this kernel satisfies all the 4 conditions, where the proof is given in Appendix A.

Lemma 1. *The Gaussian kernel in Eqn. (1) is a proper kernel.*

2.1.2. Proper metric

Before moving on to the depth function, we look for a proper metric to measure the difference between two observed point processes. Since the two point processes, in general, may have different numbers of events, multivariate measurements cannot be applied directly due to inconsistency in dimensions. As smoothing functions are included in this depth, here we will use the distance on smoothed processes as the distance on point processes. Let the time interval of the point process be $[0, T]$. We define the set of observed point process with given dimension $l > 0$ as: $\Omega_l = \{x = (x_1, x_2, \cdots, x_l) \in \mathbb{R}^l \mid 0 \le x_1 \le x_2 \le \cdots \le x_l \le T\}$. For the case of $l = 0$, there is no observed event in $[0, T]$, so Ω_0 is the set of 0-length vector, i.e. $\Omega_0 = \{\phi_0\}$ where ϕ_0 is the event time vector of no event. Then $\Omega = \cup_{l=0}^{\infty}\Omega_l$ is the space of all point processes. Each process $x = (x_1, x_2, \cdots, x_l) \in \Omega_l$ with $l > 0$ can be represented using a Dirac delta function in the form: $x(\cdot) = \sum_{i=1}^{l} \delta(\cdot - x_i)$. Let $K(\cdot; T)$ be a proper smoothing kernel. Then the smoothed process is a function on $[0, T]$ in the form:

$$f_x(u) = \sum_{i=1}^{l} K(u - x_i; T). \tag{2}$$

In general, the space of smoothed processes for l events with $l > 0$ is $\mathbb{F}_l = \{f_x : [0, T] \to \mathbb{R} \mid f_x(u) = \sum_{i=1}^{l} K(u - x_i; T)$ where $x = (x_1, x_2, \cdots, x_l) \in \Omega_l\}$. For $l = 0$, the smoothed process is $f_{\phi_0}(u) \doteq 0$ and $\mathbb{F}_0 = \{f_{\phi_0}\}$. The space of all smoothing processes is then $\mathbb{F} = \cup_{l=0}^{\infty}\mathbb{F}_l$. We point out that the correspondence between a point process and its smoothed version is one-to-one. This is given in the following lemma (the proof is in Appendix B).

Lemma 2. *For any proper smoothing kernel $K(\cdot; T)$, the smoothing procedure given in Eqn. (2) is a bijective mapping from Ω to $\mathbb{F}$.*

Now we are ready to define a metric on the point process space. This definition is based on the classical $\mathbb{L}^p$ norm on the smoothed processes. This is formally given as follows:

Definition 1. For any two point processes $s, t \in \Omega$ on $[0, T]$ and the correspondent $f_s, f_t \in \mathbb{F}$ given by Eqn. (2), we define a distance function $d_{K,p}$ between s and t as:

$$d_{K,p}(s,t) = \|f_s - f_t\|_p = \left(\int_0^T |f_s(u) - f_t(u)|^p du\right)^{1/p}, \tag{3}$$

where $\|.\|_p$ $(p \ge 1)$ indicates the classical $\mathbb{L}^p$ norm on $[0, T]$.

Note that the smoothing kernel $K(\cdot; T)$ and $\mathbb{L}^p$ norm can influence the distance value. However, we can prove that the distance $d_{K,p}$ in Eqn. (3) is a proper metric on Ω for any K and p. This is given in the following theorem (the proof is in Appendix C).

Theorem 1. *If the smoothing kernel K satisfies the four conditions to be proper, then the function $d_{K,p}$ in Definition 1 is a proper metric in the point process space Ω. That is, it satisfies non-negativity, identity of indiscernible, symmetry, and triangle inequality.*

Based on Lemma 2 and Theorem 1, we know that the point process space Ω and smoothed process space $\mathbb{F}$ are one-to-one, and an $\mathbb{L}^p$ norm on $\mathbb{F}$ can be used to define a proper metric on Ω. This metric helps solve the problem in Ω where a conventional vector metric cannot be directly used as different processes may have different cardinalities. Based on this result, we need to answer two fundamental questions:

1. If two point processes have the same cardinality and close corresponding event times, will the distance $d_{K,p}$ between them be close?

2. Conversely, if the distance $d_{K,p}$ between two point processes is close, will they have the same cardinality and close corresponding events times?

Question 1 examines if the $d_{K,p}$ distance is continuous with respect to the event times. We claim that this is true and the conclusion is stated in Proposition 1 as follows. The detailed proof is given in Appendix D.

Proposition 1. *For any $k \in \mathbf{N}$, suppose $y = (y_1, y_2, \cdots, y_k)$ is an observed point process in Ω_k. Let $x^{(n)} = \left(x_1^{(n)}, x_2^{(n)}, \cdots, x_k^{(n)}\right)$, $n = 1, 2, \cdots$ be a sequence of processes in Ω_k. If $\lim_{n\to\infty} x^{(n)} = y$, or equivalently, $\lim_{n\to\infty} x_i^{(n)} = y_i$, $i = 1, 2, \cdots, k$, then for any observed point process $z = (z_1, z_2, \cdots, z_l) \in \Omega$,*

$$\lim_{n\to\infty} d_{K,p}\left(x^{(n)}, z\right) = d_{K,p}(y, z), \quad \textit{and} \quad \lim_{n\to\infty} d_{K,p}\left(x^{(n)}, y\right) = 0.$$

Question 2 examines the inverse continuity. We claim that this inverse continuity is also true and the result is stated in Proposition 2 (see Appendix E for a detailed proof).

Proposition 2. *For any $k \in \mathbf{N}$, suppose $y = (y_1, y_2, \cdots, y_k)$ is an observed point process in Ω_k. Let $x^{(n)} = \left(x_1^{(n)}, x_2^{(n)}, \cdots, x_{k_n}^{(n)}\right)$ be a sequence of processes in Ω_{k_n}, $n = 1, 2, \cdots$. If $\lim_{n\to\infty} d_{K,p}\left(x^{(n)}, y\right) = 0$, then $\lim_{n\to\infty} x^{(n)} = y$. That is,*

$$1)\ k_n = k \text{ for } n \text{ sufficiently large}, \quad \textit{and} \quad 2) \lim_{n\to\infty} x_i^{(n)} = y_i, i = 1, 2, \cdots, k.$$

2.2. Desirable properties for depth on point process

In general, a depth function is defined to provide a measure of centrality of a given data point within a data cloud or a probability distribution. To examine if such goal is achieved, one often examines desirable mathematical properties corresponding to the centrality measurement. For instance, Zuo and Serfling (2000) proposed the desirable properties for multivariate depths. Later, Nieto-Reyes et al. (2016) provided the desirable properties for functional depths. The commonly studied properties are (1) linear invariance, (2) vanishing at infinity, (3) maximality at the center, and (4) monotonicity.

Motivated by previous studies on desirable properties for multivariate and functional depths, we propose the following five desirable properties for the depth on point process: Suppose $D(s; P_S)$ is the depth function for observed event time vector s with respect to the probability space $(\Omega, \mathcal{F}, P_S)$ for a random point process S on interval $[0, T]$, we expect that D satisfies

P1 Continuity: $D(s; P_S)$ is continuous with respect to s in Ω.

P2 Linear invariance: For any event time vector $s = (s_1, s_2, \cdots, s_k)$ in interval $[0, T]$, we transform it to $\tilde{s} = as + b = (as_1 + b, as_2 + b, \cdots, as_k + b)$ in interval $[b, aT + b]$ where $a \in \mathbb{R}^+, b \in \mathbb{R}$ are two constants. Then $D(\tilde{s}; P_{\tilde{S}}) = D(s; P_S)$, where $\tilde{S} = aS + b$ denotes a linear transformation on the point process S.

P3 Vanishing at infinity: When the number of events goes to infinity, the depth should go to 0: $D(s; P_S) \to 0$ as $|s| \to \infty$.

P4 Unique maximum at the center: There exists $s_c \in \Omega$, such that $D(s_c; P_S) = \max_{s\in\Omega} D(s, P_S)$. Also, for any $x(\neq s_c) \in \Omega$, $D(x; P_S) < D(s_c; P_S)$. We refer to s_c as the "center".

P5 Monotonic decreasing from the center: For any $s_1.s_2 \in \Omega$, suppose the center is s_c. If $d_{K,p}(s_1, s_c) < d_{K,p}(s_2, s_c)$, then we have $D(s_1, P_S) > D(s_2, P_S)$.

P4 and ***P5*** are critically important as they properly characterize the notion of "center-outward". We point out that the "center" in ***P4*** was originally defined as a point with symmetry in the data cloud. For example, when dealing with zero-mean multivariate normal samples, we will consider the origin as center, and the corresponding Mahalanobis depth is uniquely maximized at this point. However, for point process data in a finite domain $[0, T]$, it is difficult to define a geometrically symmetric central point in Ω. Notwithstanding, we expect the proposed depth has a unique maximum point, which results in our notion of "center".

Here we examine if the above five properties are satisfied by the generalized Mahalanobis depth on point process (Liu et al., 2017). Note that the generalized Mahalanobis depth a point process s is defined as a weighted product

$$D(s) := \left(\frac{\Psi^{|s|}/|s|!}{\Psi^{\lfloor\Psi\rfloor}/\lfloor\Psi\rfloor!}\right)^r \left(\frac{1}{1 + (s - \mu_{|s|})^T \Sigma_{|s|}^{-1} (s - \mu_{|s|})}\right),$$

where $r > 0$ is the weight power, $\Psi > 0$ is the total intensity and $\lfloor\cdot\rfloor$ denotes the floor function, and $\mu_{|s|} \in \mathbb{R}^{|s|}$ and $\Sigma_{|s|} \in \mathbb{R}^{|s|\times|s|}$ are conditional (on cardinality) mean and covariance, respectively. We can see that 1) If two point processes are close to each other, they should have the same number of events and close event times in the time order. As the classical Mahalanobis depth is continuous with respect to the input vector, the generalized Mahalanobis depth on point process satisfies ***P1***; 2) Linear transformation on the time will not change the number of events, so $P(|s|)^r$ will stay the same. Because the classical Mahalanobis distance is linear invariant, ***P2*** is also satisfied; 3) The Poisson mass function will go to 0 as the input goes to infinity, so $P(|s|) \to 0$ as $|s| \to \infty$. Thus, ***P3*** holds, too; 4) $D(s \mid |s|)$ has maximum value to be 1. However, this maximum in general can be achieved at multiple processes. For example, if the total intensity Ψ is an integer, both the population mean with dimension Ψ and the population mean with dimension $\Psi - 1$ will achieve the maximum depth. Thus, ***P4*** does not hold as the solution to the maximum is not unique; 5) As a result, ***P5*** does not hold, either. In summary, the generalized Mahalanobis depth on point process only satisfies ***P1*** to ***P3***.

Similar to the discussion about the Mahalanobis depth, we will refer to ***P1*** - ***P5*** to explore these properties for other depth functions in the following sections of this paper.

2.3. h-Depth on point process

We have defined a proper metric in the point process space Ω. This metric is based on a smoothing procedure where a point process is equivalently represented by a function in $\mathbb{F}$. In this section, we will exploit a commonly used functional depth, called h-depth, to define an "h-depth" for point process in Ω.

At first, we review the notion of h-depth (Cuevas et al., 2007) for functional data. Assume X is a functional random variable in the probability space $(\Lambda, \mathcal{F}_\Lambda, P_\Lambda)$, where Λ is a subset of $\mathbb{L}^2([0,T])$. Let $\|\cdot\|$ denote the conventional $\mathbb{L}^2$ norm. For any function $\eta \in \Lambda$, its h-depth is defined as:

$$HD(\eta; P_\Lambda) = \mathbb{E}\big[G_h(\|\eta - X\|)\big], \tag{4}$$

where $G_h(u) = \exp\big(-\frac{u^2}{2h}\big)$ is a modified Gaussian kernel with positive hyper-parameter h. To simplify notation, we use $HD(\eta)$ to represent $HD(\eta; P_\Lambda)$ when the measure P_Λ is implicitly known. Given a set of i.i.d. random functions $X_1, X_2, \cdots, X_n \in \Lambda$, the sample version of the h-depth of η is given as

$$HD_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} G_h(\|\eta - X_i\|) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\eta - X_i\|^2}{2h}\right).$$

One important property of the h-depth in Eqn. (4) is its continuity. This is given in the following lemma, where the proof can be found in Appendix F.

Lemma 3. *If the functional random variable X satisfies $\mathbb{E}\|X\| < \infty$, then $HD(\cdot)$ is a continuous function on Λ.*

Based on the notion of h-depth on functional data (in Eqn. (4)) and the proper metric on point process (in Eqn. (3)), we can formally introduce the h-depth to point process observations as follows:

Definition 2. Let S be a random point process on $[0,T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$. The h-depth of any $s \in \Omega$ is defined as:

$$D(s; P_S) = \mathbb{E}\big[G_h(\|f_s - f_S\|)\big], \tag{5}$$

where f_s and f_S are smoothed curves for s and S (by Eqn. (3)) and $G_h(u) = \exp\big(-\frac{u^2}{2h}\big)$.

Definition 2 provides the population version of the depth. In practice, we should follow the sample version of h-depth on the given observations, which is given below.

Definition 3. Let $\{S_i\}_{i=1}^N$ be a sample of event time vectors from a point process on $[0,T]$. The empirical h-depth of any $s \in \Omega$ is defined as:

$$\hat{D}\big(s; \{S_i\}_{i=1}^N\big) = \frac{1}{N}\sum_{i=1}^{N} G_h(\|f_s - f_{S_i}\|), \tag{6}$$

where f_s and f_{S_i} are smoothed curves for s and S_i, $i = 1, 2, \cdots, N$, and $G_h(u) = \exp\big(-\frac{u^2}{2h}\big)$.

In Definitions 2 and 3, h is positive and serves as a variance hyper-parameter. Changing the value of h will influence the depth values and possibly the ranking orders. With the definitions for both the population and empirical versions of the h-depth on point process, we now examine the mathematical properties based on the discussion in section 2.2. The basic result can be summarized in the following proposition.

Proposition 3. *Let $D(\cdot)$ denote the h-depth on point process in Definition 2. Then it satisfies the following properties:*

- ***P1***: *If $\mathbb{E}(|S|) < \infty$, then the depth $D(s; P_S)$ is continuous with respect to s.*
- ***P2***: *If the hyper-parameter h is proportional to the interval length T, i.e. $h = CT$ for some constant $C > 0$, then the depth $D(s; P_S)$ is invariant with respect to a linear transformation on the time interval.*
- ***P3***: $D(s; P_S) \to 0$ *when* $|s| \to \infty$.

The detailed proof is given in Appendix G. Note that ***P4*** and ***P5*** are not satisfied because h-depth in general may not have a unique maximum point. Therefore, a center-outward decreasing depth structure will not be guaranteed.

2.4. Modified h-depth on point process

The center has been a critical notion in statistical depths. However, as we have pointed out in Sec. 2.3, the h-depth for a point process may not have a unique center. In this section, we propose to modify the h-depth by including a “center”-based process in the definition.

2.4.1. Definition and properties

Many commonly used depth functions have the “center” as the point with maximum depth. Based on this idea, we propose a center-based new depth function on point process using the smoothing method. The formal definition is given as follows:

Definition 4. Let s_c be a given “center” point process in Ω on $[0, T]$. For any $s \in \Omega$, its center-based h-depth is defined to be:

$$D(s; s_c) = \exp\left(-\frac{\|f_s - f_{s_c}\|^2}{2h}\right), \tag{7}$$

where $h > 0$ is a hyper-parameter. f_s and f_{s_c} are smoothed processes of s and s_c, respectively.

Remark 1: We point out that Definition 4 is a modified version of Definition 2. In fact, the classical h-depth of a process s takes the form $\mathbb{E}\big[\exp(-\frac{\|f_s - f_S\|^2}{2h})\big]$. If the order of the exponential function and the expectation is switched, then we have:

$$\begin{aligned}
&\exp\left[\mathbb{E}(-\frac{\|f_s - f_S\|^2}{2h})\right] \\
&= \exp\left(-\frac{\mathbb{E}\|f_s - \mathbb{E}f_S\|^2 + \mathbb{E}\|\mathbb{E}f_S - f_S\|^2 + 2\mathbb{E} < f_s - \mathbb{E}f_S, \mathbb{E}f_S - f_S >}{2h}\right) \\
&= \exp\left(-\frac{\|f_s - \mathbb{E}f_S\|^2 + \mathbb{E}\|\mathbb{E}f_S - f_S\|^2}{2h}\right) \\
&\propto \exp\left(-\frac{\|f_s - \mathbb{E}f_S\|^2}{2h}\right).
\end{aligned}$$

By approximating the center f_{s_c} by using the expectation $\mathbb{E}f_S$, we obtain the center-based h-depth $D(s; s_c)$.

Remark 2: The $\mathbb{L}^2$ norm in Definition 4 can be generalized to $\mathbb{L}^p$, with $1 \le p < \infty$.

Different from the h-depth in section 2.3, h in Definition 4 serves as a scaling hyper-parameter to make the resulting depth values to a desired range. Because every $\mathbb{L}^2$ distance is divided by the same h, the h constant will have impact on the output depth values, but will not influence the ranking orders.

With a center process given in the definition, this modified h-depth for point process is expected to satisfy more desirable mathematical properties than the classical one. Indeed, all the desirable properties in section 2.2 are satisfied and the details are given in the following proposition.

Proposition 4. *Let $D(\cdot; s_c)$ denote the center-based h-depth on point process in Definition 4 and the metric in Ω is the $\mathbb{L}^2$ norm $d_{K,2}(s, t) = \|f_s - f_t\|$. Then $D(\cdot; s_c)$ satisfies the following desirable mathematical properties in section 2.2:*

- ***P1***: *If $|s| < \infty$, then depth $D(s; s_c)$ is continuous with respect to s.*
- ***P2***: *If the hyper-parameter h is proportional to the interval length T, i.e. $h = CT$ for some constant $C > 0$, then the depth $D(s; s_c)$ is invariant with respect to a linear transformation on the time interval.*
- ***P3***: $D(s; s_c) \to 0$ *when* $|s| \to \infty$.
- ***P4***: $D(s_c; s_c) = \max_{s \in \Omega} D(s; s_c)$ *and* $\forall t (\neq s_c) \in \Omega$, $D(t; s_c) < D(s_c; s_c)$.
- ***P5***: *For any $s_1, s_2 \in \Omega$, if $d_{K,2}(s_1, s_c) < d_{K,2}(s_2, s_c)$, then $D(s_1; s_c) > D(s_2; s_c)$.*

The detailed proof is provided in Appendix H. Note that this is a clear advantage over the h-depth method in section 2.3, where only properties ***P1***-***P3*** are satisfied.

2.4.2. Estimation of the center

In practice, there is no prior knowledge of the center of point process, so a proper estimation of the center will be needed. As Ω and $\mathbb{F}$ are not conventional vector spaces, we adopt the common notion of the Karcher mean in a metric space (Grove and Karcher, 1973). Since we have a proper metric between any two point processes in Definition 1, the Karcher mean can be defined as follows:

Definition 5. Let S be a random point process on $[0, T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$. The Karcher mean of S is defined to be

$$\mu_K = \underset{t\in\Omega}{\arg\min}\, \mathbb{E}\big[d^2_{K,2}(t, S)\big], \tag{8}$$

where $d_{K,2}(\cdot,\cdot)$ is the metric with kernel function $K(\cdot;T)$ and $\mathbb{L}^2$ norm in Definition 1.

In general the solution of the minimization, if existing, may not be unique, so μ_K is a set of point processes. Based on the Karcher mean in Definition 5, its empirical version is given as follows:

Definition 6. Let $\{S_i\}_{i=1}^N$ be a collection of N observed point processes on $[0, T]$. The empirical Karcher mean of this process, based on $\{S_i\}_{i=1}^N$, is defined to be

$$\bar{S}_K^{(N)} = \underset{t\in\Omega}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N} d^2_{K,2}(t, S_i). \tag{9}$$

The empirical Karcher mean also may not be unique, and therefore $\bar{S}_K^{(N)}$ is a set of solution points, too. For simplicity, we define the following notation:

$$SSD(t; S_1, S_2, \cdots, S_N) = \sum_{i=1}^{N} \|f_t - f_{S_i}\|_2^2,$$

where SSD stands for "Sum of Squared Distance" and $t \in \Omega$ is the event time vector. With this definition, the empirical Karcher mean is just $\bar{S}_K^{(N)} = \arg\min_{t\in\Omega} SSD(t; S_1, S_2, \cdots, S_N)$. For any element in $\bar{S}_K^{(N)}$, its dimension has a finite upper bound. The result is formally given as follows, where the proof is given in Appendix I.

Theorem 2. *Any solution in the empirical Karcher mean of a point process has an upper bound in dimension, i.e.* $\exists D_K^{(N)} \in \mathbb{N}^+$ *such that* $\dim(\bar{S}) \le D_K^{(N)}, \forall \bar{S} \in \bar{S}_K^{(N)}$.

We will introduce a theoretically-proven convergent algorithm to estimate $\bar{S}_K^{(N)}$ where this upper bound $D_K^{(N)}$ provides an effective searching range. Such algorithm is a combination of simulation annealing and line search. We will at first briefly review the simulation annealing and line search, and then introduce an approach to combine these two methods.

- RJMCMC annealing

  Simulated annealing (Geman and Geman, 1984; Van Laarhoven and Aarts, 1987) is a method that makes optimization feasible if we are able to generate samples. To find the global minimum of function $f(u)$, we can build a density function $\pi_i(u) \propto \exp\big[-\frac{1}{T_i} f(u)\big]$ where T_i is a pre-determined decreasing sequence of positive numbers called temperature and $\lim_{i\to\infty} T_i = 0$ (e.g. $T_i = C/\log(1+i)$ for constant $C > 0$). Under weak regularity assumptions on f, the density π_i will be concentrated on the global minimum points of f. After a large number of iterations of computing π_i and sampling from it, we will have samples passing the global minimum point.

  Simulation annealing provides a method to find the minimum of a function, but to apply it on the optimization of SSD, a tool that is able to generate vectors with different dimensions will be needed. In this paper, we adopt the reversible-jump Markov Chain Monte Carlo (RJMCMC) method, which was first introduced by Green (1995). Its main idea is the Metropolis Hasting method (Metropolis et al., 1953), but generalizes to allow the candidate to have different dimensions. In this way, it can be used to get samples from densities where the dimension of the random vector is not fixed. One condition to use the RJMCMC method is that the number of possible dimensions should be finite. According to Theorem 2, the range of dimension is $\big\{0, 1, \cdots, D_K^{(N)}\big\}$, so RJMCMC is feasible for the center estimation.

  Combining the two methods, we are able to find the Karcher mean through sampling from $\pi_i(t) \propto \exp\big[-\frac{1}{T_i} SSD(t; S_1, S_2, \cdots, S_N)\big]$. Because of RJMCMC, the annealing process can search over different dimensions automatically. An example RJMCMC annealing algorithm is provided in Appendix J, where the input to the algorithm can be adjusted before application.

Algorithm 1 Combined method to find empirical Karcher mean.

Input: the observed event time vectors $S_1, S_2, \cdots, S_N$.
For RJMCMC annealing: The maximum number of iterations n_{max}; Initial value x_0; Initial dimension k_0 to be the dimension of x_0; the upper bound of dimension $D_K^{(N)}$.
For line search: The batch size B; the learning rate r; the maximum number of epochs ep_{max}; convergence indicator ϵ.
For combination: The number of dimensions to be kept in RJMCMC annealing: $d_r \in \mathbb{N}$.
(1) Pre-train:
Use RJMCMC annealing (such as Algorithm 2 in Appendix J) with input $S_1, S_2, \cdots, S_N$, n_{max}, x_0, k_0 and $D_K^{(N)}$ to find the top d_r number of events $\{k_{0,i}\}_{i=1}^{d_r}$ and the corresponding event time vectors $\{x_{0,i}\}_{i=1}^{d_r}$ that produce the smallest SSD values.
(2) Optimization:
Use line search (such as Algorithm 3 in Appendix L) with input $S_1, S_2, \cdots, S_N$, B, r, ep_{max}, ϵ as parameters, and the output from RJMCMC annealing: $\{k_{0,i}\}_{i=1}^{d_r}$ and $\{x_{0,i}\}_{i=1}^{d_r}$ as searching dimensions and initial values, to find the minimum solution of the SSD: $\bar{S}_K^{(N)}$.
Output: $\bar{S}_K^{(N)}$ is the empirical Karcher mean.

- Line search
According to Theorem 2, the dimensions to search are in a finite range $\left\{0,\ 1,\ \cdots,\ D_K^{(N)}\right\}$. Therefore, we can alternatively do optimization in each dimension and then compare the outputs across dimensions. In this case, the optimization can be done using an efficient "line search", where the gradient of the SSD can be computed explicitly in a given dimension, as shown in the following proposition.

Proposition 5. *If the dimension of the input event time vector t is given and the distance function is $d_{K,2}$, then the gradient of SSD with respect to t is:*

$$\frac{\partial SSD}{\partial t}(t; S_1, S_2, \cdots, S_N) = -2\int_0^T K'(u-t)\Big[N\vec{1}_{|t|}^T K(u-t;T) - \sum_{i=1}^N \vec{1}_{|S_i|}^T K(u-S_i;T)\Big]du,$$

where t and S_i are event time vectors, and $|t|$ is the dimension of vector t. All operations are done element-wise.
If the modified Gaussian kernel $K_G(u;T) = c_1 e^{-\frac{c_2}{T^2}u^2}$ is used for smoothing, then the gradient can be simplified as:

$$\frac{\partial SSD}{\partial t}(t; S_1, S_2, \cdots, S_N) = \frac{4c_1^2 c_2}{T^2}\left[Ng(t\vec{1}_{|t|}^T, \vec{1}_{|t|}t^T)\vec{1}_{|t|} - \sum_{i=1}^N g(t\vec{1}_{|t|}^T, \vec{1}_{|S_i|}S_i^T)\vec{1}_{|S_i|}\right],$$

where the function $g(\cdot,\cdot)$ has the following form:

$$g(u,v) = e^{-\frac{c_2}{2T^2}(u-v)^2}\left\{\frac{T^2}{4c_2}\Big[e^{-\frac{2c_2}{T^2}(\frac{u+v}{2})^2} - e^{-\frac{2c_2}{T^2}(T-\frac{u+v}{2})^2}\Big] - \sqrt{\frac{\pi}{8c_2}}T(u-v)\Big[\Phi\big(\frac{2\sqrt{c_2}}{T}(T-\frac{u+v}{2})\big) - \Phi\big(-\frac{\sqrt{c_2}(u+v)}{T}\big)\Big]\right\}.$$

Here, Φ is the cumulative distribution function for standard Normal distribution $N(0,1)$; $\vec{1}_d = [1, 1, \cdots, 1]^T \in \mathbb{R}^d$. All operations are done element-wise.

The computational detail in Proposition 5 is provided in Appendix K. Because of this result, gradient-based methods such as stochastic gradient descent can be applied to conduct the optimization in each given dimension. By comparing optimization result across all dimensions, we will get a solution to the empirical Karcher mean. An example line search algorithm using the gradient in Proposition 5 is provided in Appendix L.

- Combined method
Both RJMCMC annealing and line search can be used to estimate the empirical Karcher mean, but either has apparent disadvantages. Although the RJMCMC annealing can search over dimensions automatically, it often converges extremely slowly to the optimal solution. Similarly, the line search can converge fast to the solution in a given dimension, but it needs to search every dimension in a large range. As a result, the line search is also an inefficient method to obtain empirical Karcher mean.

To improve the efficiency, we propose to combine these two methods in a sequential way: do RJMCMC annealing in a given number of iterations first as a "pre-train" process to locate a range of optimal dimensions, and then use line search to do optimization within the narrowed dimension range. Moreover, the output event time vectors from the RJMCMC annealing can be used as the initial values for the line search. In this way, the number of iterations in the line search will be reduced, which could also help avoid local optimal points. This combined algorithm is shown in Algorithm 1. The application of this algorithm on simulations and real data will be given in section 4.

3. Asymptotic theory

In this section, we will show that the empirical depth converges to the population depth as the sample size goes to infinity for h-depth and modified h-depth, defined in section 2. Firstly, for the h-depth, we have the following result:

Theorem 3. *Based on Definitions 2 and 3, let $D(\cdot; P_S)$ represent the h-depth for a point process on $[0, T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$ and $\hat{D}\big(\cdot; \{S_i\}_{i=1}^N\big)$ is the corresponding independent sample version with sample size N. Then, for any input $s \in \Omega$,*

$$\hat{D}\big(s; \{S_i\}_{i=1}^N\big) \to D(s; P_S) \text{ a.s.} \tag{10}$$

This theorem implies the convergence of the sample h-depth to the population h-depth, and can be easily proven using the Strong Law of Large Numbers.

For the modified h-depth, the convergence of the depth value will be based on the convergence of the estimated center. As defined in section 2.4.2, the Karcher mean can be treated as the solution in Ω that minimizes the average of squared distances, where the distance is $d_{K,2}$ with $\mathbb{L}^2$ norm. In fact, we can generalize the Karcher mean definition by using any $d_{K,p}$ distance for $p \geq 1$, given in the following form:

Definition 7. Let S be a random point process on $[0, T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$. The generalized Karcher mean of S is defined to be

$$\mu_K = \underset{t \in \Omega}{\arg\min}\, \mathbb{E}\big[d_{K,p}^2(t, S)\big], \tag{11}$$

where $d_{K,p}(\cdot, \cdot)$ is the metric with kernel function $K(\cdot; T)$ and $\mathbb{L}^p$ norm with $p \geq 1$.

Definition 8. Let $\{S_i\}_{i=1}^N$ be a collection of N independent point processes on $[0, T]$. The empirical generalized Karcher mean of this process, based on $\{S_i\}_{i=1}^N$, is defined as

$$\bar{S}_K^{(N)} = \underset{t \in \Omega}{\arg\min}\, \frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t, S_i). \tag{12}$$

It is apparent that Definitions 5 and 6 are special cases with $p = 2$ in the generalized definitions, respectively. Then the convergence theorem is given as follows:

Theorem 4. *S is a random point process on $[0, T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$, where the number of events $|S|$ has a constant upper bound $D > 0$. $\{S_i\}_{i=1}^N$ is the set of independent event time vectors from S. Then we have:*

1. *The minimum average of squared distance in $\{S_i\}_{i=1}^N$ will converge to the minimum expected squared distance in $(\Omega, \mathcal{F}, P_S)$ almost surely, in other words,*

$$\min_{t \in \Omega} \frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t, S_i) \to \min_{t \in \Omega} \mathbb{E}\big[d_{K,p}^2(t, S)\big] \quad \text{almost surely.} \tag{13}$$

2. *The empirical generalized Karcher mean converges to the generalized Karcher mean almost surely, in other words,*

$$\bar{S}_K^{(N)} \to \mu_K \quad \text{almost surely.} \tag{14}$$

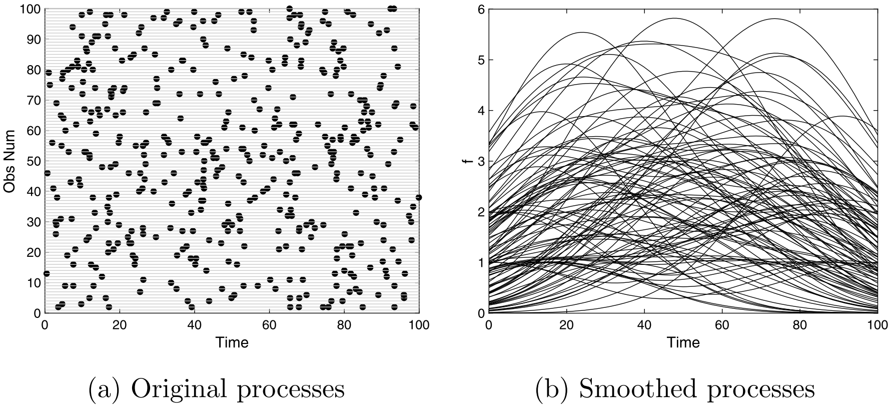
The almost-surely convergence from set $\bar{S}_K^{(N)}$ to set μ_K means $\operatorname{Lim\,sup}_{n\to\infty} \bar{S}_K^{(N)} \subset \mu_K$ a.s., where $\operatorname{Lim\,sup}_{n\to\infty} \bar{S}_K^{(N)}$ is the Kuratowski upper limit (Kuratowski, 2014):

$$\operatorname*{Lim\,sup}_{n\to\infty} \bar{S}_K^{(N)} = \Big\{x \in \Omega \quad \big| \quad \liminf_{N\to\infty} d_{K,p}\big(x, \bar{S}_K^{(N)}\big) = 0\Big\},$$

where $d_{K,p}\big(x, \bar{S}_K^{(N)}\big) = \inf\big\{d_{K,p}(x, \bar{S}) \quad | \quad \bar{S} \in \bar{S}_K^{(N)}\big\}$.

3. *Let $D(\cdot; s_c)$ be the modified h-depth in Definition 4 with center s_c. Suppose the Karcher mean contains a unique element: $\mu_K = \big\{s_c^{(p)}\big\}$. $\hat{s}_c^{(N)}$ is any element from the empirical Karcher mean $\bar{S}_K^{(N)}$ based on $\{S_i\}_{i=1}^N$. Then, for any $s \in \Omega$,*

$$D\big(s; \hat{s}_c^{(N)}\big) \to D\big(s; s_c^{(p)}\big) \text{ almost surely.} \tag{15}$$



(a) Original processes

(b) Smoothed processes

Fig. 1. Simulation on HPP(0.045). (a) 100 realizations from HPP(0.045) on interval [0, 100]. Each row represents one realization, where each dot is one event. (b) Smoothed processes of these 100 realizations using a modified Gaussian kernel.

Remark 1: In Eqns. (11), (12), *and* (13), min *is used instead of* inf *because the minimum value can be achieved in both the empirical version and the population version.*
Remark 2: The theorem will still hold when the square power of the distance in Eqn. (13) *and in the definitions of* $\bar{S}_K^{(N)}$ *and* μ_K *are generalized to any* $r \geq 1$, *i.e. replacing* $d^2_{K,p}$ *by* $d^r_{K,p}$ *with any* $r \in [1, \infty)$ *in the Karcher mean and empirical Karcher mean.*

The proof of Theorem 4, together with the two remarks, is shown in Appendix M. With the first two points of the theorem, we can confirm that the minimum of the SSD converges to the minimum of the population expected squared distance, irrespective of the optimal solution sets in sample and population. It also indicates the convergence of the estimated center set to its population version. The last point of the theorem suggests that the modified h-depth with sample Karcher mean as center converges to the one with population Karcher mean as the center.

4. Application results

In this section, we will apply the proposed methods, h-depth and modified h-depth, on point process observations in simulations and real experimental data.

4.1. Simulation studies

4.1.1. Homogeneous Poisson process (HPP)

We will at first illustrate the depth methods on observations from HPP(λ) in a finite time interval, where λ is a constant mean value. 100 independent realizations from HPP(0.045) on $[0, 100]$ are generated, where the raster plot of these processes is shown in Fig. 1(a). Basically, the number of events in each process follows a Poisson distribution with mean 4.5 (varying from 1 to 10 in these 100 realizations), and the event time in every observation is uniformly distributed on [0, 100]. We then smooth the realizations using a kernel in Eqn. (1) with $c_1 = 1, c_2 = 10$, so that the output smoothing curves have proper function values (within $[1, 100]$) and also proper smoothness (number of peaks less than 5). The smoothed processes are shown in Fig. 1(b).

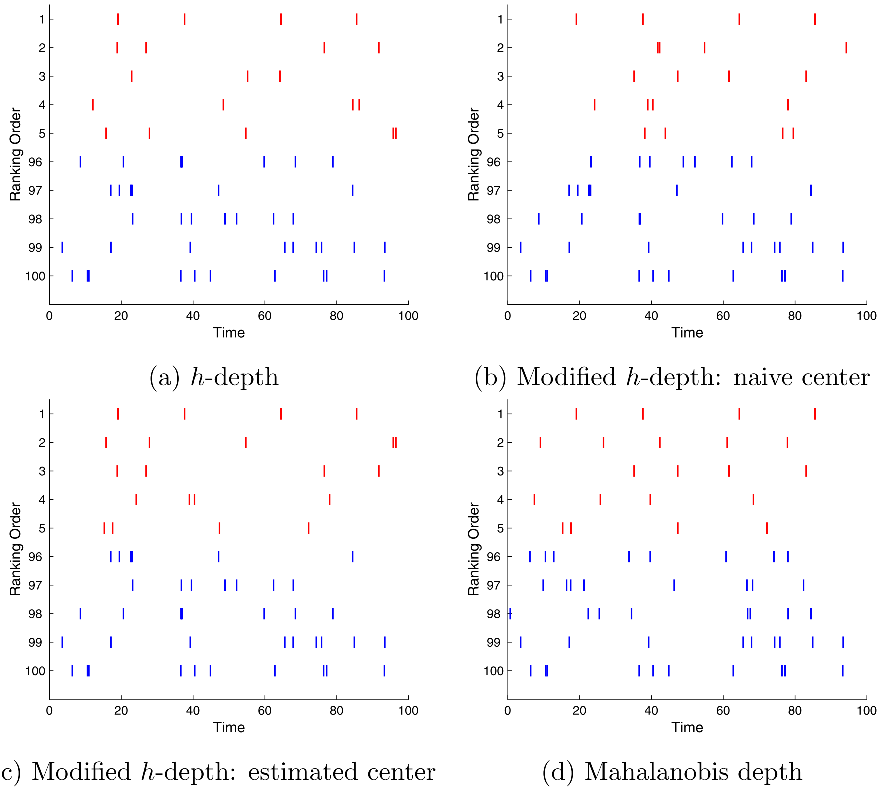
Based on the h-depth and modified h-depth in section 2, we compute the sample depth value for each given process where the constant h is set to $T = 100$. For the modified h-depth, the center is determined in four ways: set by prior information, estimated by RJMCMC annealing, line search and combined method. As the Poisson process is homogeneous, we may naively set the center as $(20, 40, 60, 80)$. The output of the center estimation is shown in Table 1. We can see that all the estimated centers have lower SSD than the naive one. In particular, the combined method has the superior performance, with the smallest SSD value and the smallest time cost. More analysis to compare the 3 center estimation methods for this simulation can be seen in Appendix N. Because of this result, we will only use this method for the modified h-depth in comparison with other depth methods.

The depth values for all 100 processes are computed and the observations are then ranked. For better illustration, instead of showing all the ranked processes, we display only the top 5 and bottom 5 in Fig. 2. For comparison, we also adopt the generalized Mahalanobis depth with the power weight $r = 1$, which properly balance the importances of the number of events and their distribution, and show its ranking result in Fig. 2. It can be seen that the outputs from the h-depth and modified h-depth are similar to the output from the Mahalanobis depth. That is, observations with more uniformly distributed events turn to have larger depth values. In addition, the processes with the number of events around the mean value of 4.5 have larger depth values.

We then display the corresponding smoothed processes in Fig. 3. Note that the generalized Mahalanobis depth is not included as it is not based on smoothing functions. All three plots demonstrate a center-outward decreasing depth value

Table 1
Center estimation output for the HPP simulation, where the time cost indicates mean (±standard deviation) over 5 repetitions.

Method	Estimated center	SSD	Time cost (s)
Naive	(20.00, 40.00, 60.00, 80.00)	14217	–
RJMCMC Annealing	(19.85, 30.10, 67.36, 79.44)	13864	231.78 (±13.38)
Line Search	(17.07, 34.13, 63.84, 81.79)	13877	274.34 (±41.32)
Combined	(20.04, 30.59, 67.26, 78.44)	13857	70.75 (±6.24)



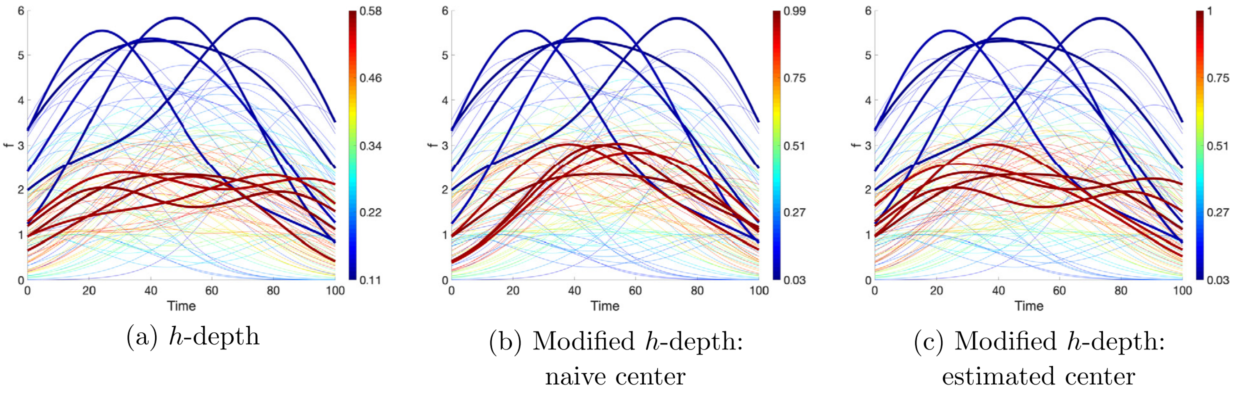
(a) h-depth

(b) Modified h-depth: naive center

(c) Modified h-depth: estimated center

(d) Mahalanobis depth

Fig. 2. Top 5 (red) and bottom 5 (blue) processes ranked by the depth values in the HPP simulation. (a) By the h-depth. (b) By the modified h-depth with naive center in Table 1. (c) By the modified h-depth with center estimated using the combined method in Table 1. (d) By the Mahalanobis depth. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

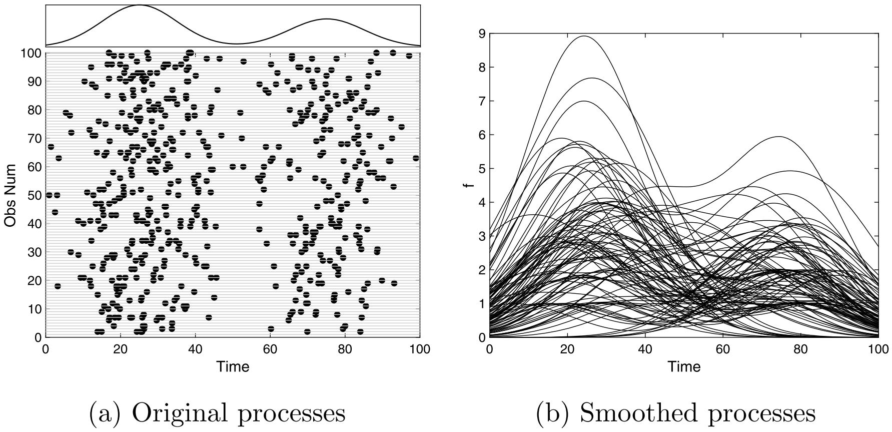


(a) h-depth

(b) Modified h-depth: naive center

(c) Modified h-depth: estimated center

Fig. 3. Color-mapped smoothed processes based on depth values for the HPP simulation, where top 5 (red) and bottom 5 (blue) are marked with thick lines. (a) By the h-depth. (b) By the modified h-depth with the naive center. (c) By the modified h-depth with the estimated center using the combined method.



(a) Original processes

(b) Smoothed processes

Fig. 4. An IPP simulation with a dual-peak intensity function. (a) 100 realizations from IPP[$\lambda(u)$] on interval $[0, 100]$. Each row represents one realization, where each dot is one event. The intensity $\lambda(u)$ is shown on the top of all realizations. (b) Smoothed processes of the 100 realizations in (a) using the modified Gaussian kernel.

Table 2
Center estimation output for the IPP simulation, where the time cost indicates mean (±standard deviation) over 5 repetitions.

Method	Estimated center	SSD	Time cost (s)
RJMCMC Annealing	$(13.73, 27.94, 32.69, 62.85, 79.38)$	14152	372.32 (±144.80)
Line Search (SGD)	$(18.72, 24.49, 37.57, 65.67, 82.67)$	14064	314.67 (±13.43)
Combined	$(19.01, 24.17, 37.27, 65.42, 82.32)$	14062	68.24 (±18.05)

structure that observations whose smoothing curves are closer to the middle of all the curves will have larger depth values. Overall, these simulation results are reasonable and consistent with the basic notion of center-outward ranks.

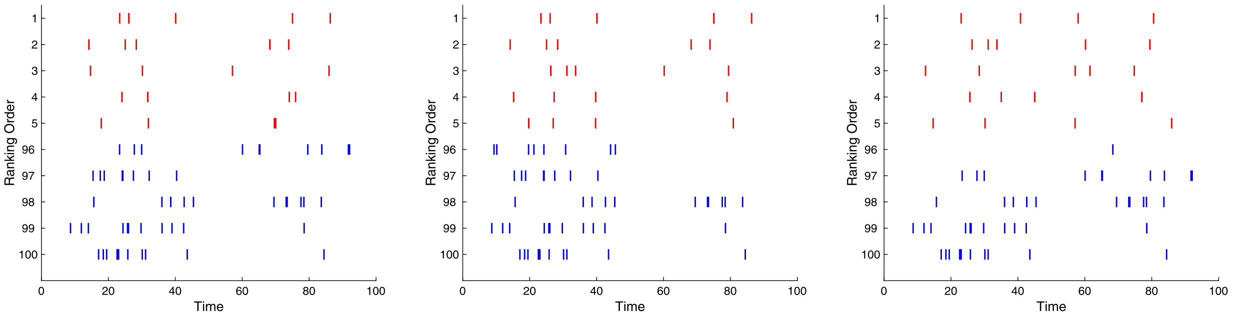
4.1.2. *Inhomogeneous Poisson process (IPP)*

We will further apply the depth on point process using observations from IPP$\big[\lambda(u)\big]$ to evaluate the performance. The rate function $\lambda(u)$ is chosen as $\lambda(u) = 3\phi(u; 25, 10) + 2\phi(u; 75, 10)$ and the event time interval is still $[0, 100]$, where $\phi(\cdot; \mu, \sigma)$ is the density for normal distribution with mean μ and standard deviation σ. As $\lambda(u)$ contains two peaks, the simulated event times will mainly distribute around 25 and 75. The kernel function is also the modified version of Gaussian kernel with $c_1 = 1, c_2 = 25$, so that the output smoothing curves have proper function values (within [1,100]) and also proper smoothness (number of peaks less than 5). 100 samples are generated. The point process plot and smoothing curve plot are shown in Fig. 4.

Similar to section 4.1.1, we let the constant h equal $T = 100$ in the h-depth and modified h-depth. Because there is no naive center in this case, only the estimated centers are used in the modified h-depth. We again use the RJMCMC annealing, line search and combined method. The output is shown in Table 2. It can be seen that the three estimated centers are very similar. The line search method and the combined method both have the lowest SSD, whereas the latter one is more efficient. That is, the combined method results in the superior performance, and therefore we will only use this method for the modified h-depth in comparison with other depth methods. More analysis to compare these 3 center estimation methods for this simulation can be seen in Appendix O.

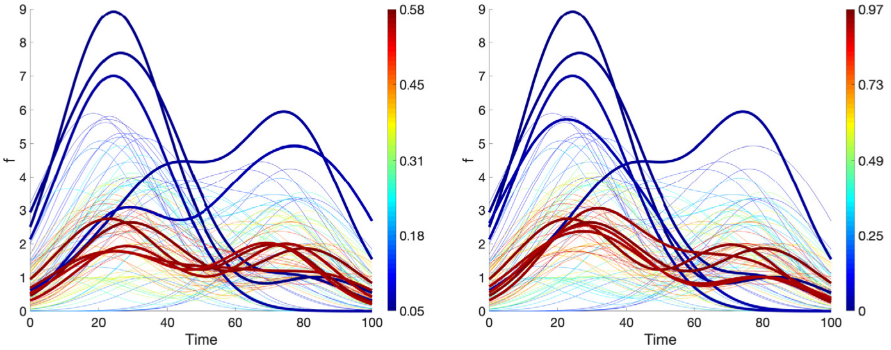
After computing the depth values for all observations, we obtain the ranking result. Fig. 5 shows the top 5 and bottom 5 ranked observations. Similar to the HPP case, we include the result from the generalized Mahalanobis depth with $r = 1$ for comparison. It can be seen that the outputs from h-depth and modified h-depth are similar – observations with event times around 25 and 75 and number of events around 5 have larger depth. In contrast, for the Mahalanobis depth, we can also observe that processes with number of events around 5 have larger depth. However, rather than concentrating at two peak time locations, the top 5 processes have event time more uniformly distributed. It indicates that h-depth and modified h-depth can better capture the pattern inside the sample. If further taking into account the true $\lambda(u)$ that the peak at 25 is higher than the peak at 75, we should expect to see more event times concentrated around 25 than around 75. Comparing Fig. 5 (a) and (b), we can see that for the modified h-depth, all the top-5 observations have more events around 25 than around 75, whereas this is not clearly shown for the h-depth. The smoothed processes ranked by the h-depth and modified h-depth are shown in Fig. 6. They both exhibit a clear center-outward structure. That is, observations whose smoothing curves are peaked at around 25 and 75 have relatively larger depth values.

In summary, based on the result from the HPP and IPP simulations, we can see that the proposed h-depth and modified h-depth both can properly build a center-outward rank on the given point process data.



(a) h-depth (b) Modified h-depth (c) Mahalanobis depth

Fig. 5. Top 5 (red) and bottom 5 (blue) processes ranked by the depth values for the IPP simulation. (a) By the h-depth. (b) By the modified h-depth with center estimated using the combined method in Table 2. (c) By the Mahalanobis depth.



(a) h-depth (b) Modified h-depth

Fig. 6. Color-mapped smoothed processes based on depth values for the IPP simulation, where top 5 (red) and bottom 5 (blue) are marked with thick lines. (a) By the h-depth. (b) By the modified h-depth with center estimated using the combined method.

4.2. Experimental data application

We will examine the proposed depth framework using a real spike train recording, where the data were included in the Quantitative Single-Neuron Modeling Competition 2009 (Naud et al., 2009) and were accessible at http://dx.doi.org/10.6080/K0PN93H3. The detailed experiment description can be found in (Carandini et al., 2007; Sincich et al., 2007). Briefly, the experiment was performed on rhesus monkeys that retinal input (visual stimulus) was applied and extracellular potentials were recorded for both the retinal (pre-synaptic) and the geniculate (post-synaptic) simultaneously. There were 10 seconds stimulus, where the first 5 seconds were the same for all trials and the last 5 seconds were unique for each trial. In total, 76 trials, including 76 pre-synaptic observations and 76 post-synaptic observations, were performed. Since the task in the competition was to predict the post-synaptic spikes given the pre-synaptic spikes, only 38 pairs of (post-synaptic, pre-synaptic) were fully given. We will use these spike train data to test the smoothing depth framework through a classification task: In the training set, we have labeled observations for pre-synaptic (abbreviated as “pre-group”) and post-synaptic (abbreviated as “post-group”). In the testing set, we compute the depth for each observation in the two groups. The group with a larger depth value will be the predicted label.

Before the classification, an exploratory data analysis is performed. The spike times are within the 10 seconds range and the point process plot is shown in Fig. 7. 10 example pairs of pre-group, post-group observations are shown. As stated in the experiment description, the stimuli were the same for the first 5 seconds and different for the last 5 seconds. This well explains why the signals have very similar pattern in the first 5 seconds. Therefore, it is reasonable to separate the trial into two parts, first 5 seconds and last 5 seconds, and we can smooth them in different ways.

After separation, we apply the modified Gaussian kernel to smooth the observations with $c_1 = 1$, $c_2 = 100$ for first 5 seconds and $c_1 = 1$, $c_2 = 50$ for last 5 seconds, so that the output smoothing curves have proper function values (within $[1, 100]$) and also proper smoothness (number of peaks less than 5). As the last 5 second data is much more noisy than the first 5 second data, a smaller c_2 is selected for stronger smoothness. The smoothing curve plot is given in Fig. 8. Using the combined method in section 2.4.2, we estimate the Karcher mean for the two groups and the corresponding smoothing curves are shown as thick lines. It can be seen that the Karcher mean curves are able to catch the shape patterns in the first 5 seconds as the observations are nearly aligned, while in the last 5 seconds, the Karcher mean is nearly flat since the observations are very noisy.

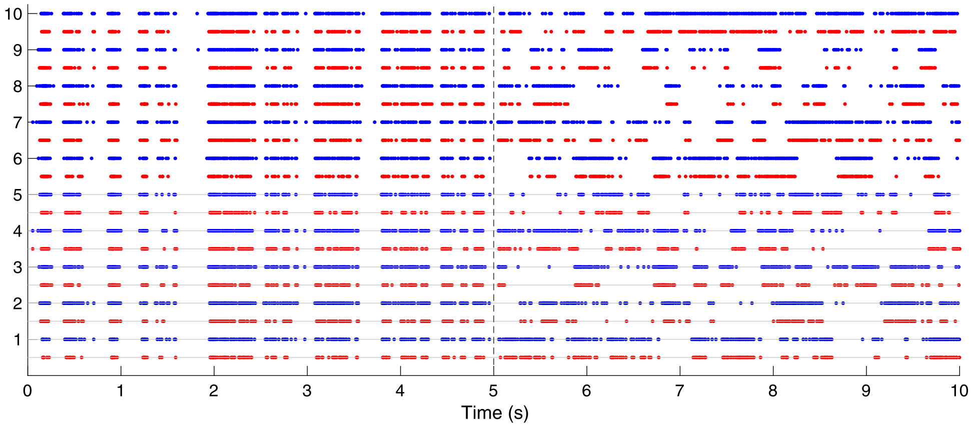


Fig. 7. The point process plot for the spike train observations. Blue: pre-group; Red: post-group. 10 example pairs were selected and displayed.

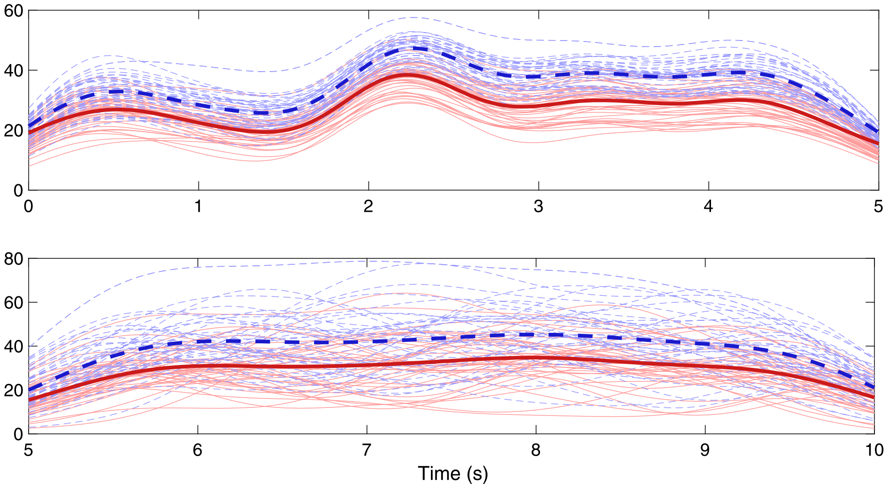


Fig. 8. Smoothed observations for pre-group and post-group in the first 5 seconds (top panel) and last (bottom panel) 5 seconds. Blue dashed lines: pre-group; Red solid lines: post-group. The thick lines denote the estimated Karcher means in the two groups, respectively.

The classification is done by a 4-fold cross validation on the pairs of (pre-synaptic, post-synaptic) observations. The 38 pairs are randomly shuffled and separated into 4 folds. For each fold, the labels of the observations are predicted, using the remaining 3 folds as training set. After iterating all the 4 folds, the predicted labels of all the 76 observations can be obtained. Then the accuracy and F1 score (van Rijsbergen, 1979), which is the harmonic mean of the precision (positive predictive value) and the recall (true positive rate), are computed based on all the 76 predicted labels and true labels to measure the classification performance. In addition to the classical h-depth and modified h-depth with center given, we conduct the classification through the generalized Mahalanobis depth on point process (Liu et al., 2017), as well as the band depth and modified band depth (López-Pintado and Romo, 2009) on the smoothed observations with the number of curves to form band (J) to be 3 and 2 respectively. A brief definition of the band depth and modified band depth is given as the following: Suppose $\eta_1(u), \cdots, \eta_n(u)$ are observed functions with $u \in \Theta$, then for any of these curves η, the Band depth of η is: $BD_{n,J}(\eta) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1\le i_1<\cdots<i_j\le n} I\big\{\forall u \in \Theta,\ \min_{r=1,\cdots,j} \eta_{i_r}(u) \le \eta(u) \le \max_{r=1,\cdots,j} \eta_{i_r}(u)\big\}$, and the Modified Band depth of η is: $MBD_{n,J}(\eta) = \frac{1}{\zeta(\Theta)} \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1\le i_1<\cdots<i_j\le n} \zeta\Big(\big\{u \in \Theta \mid \min_{r=1,\cdots,j} \eta_{i_r}(u) \le \eta(u) \le \max_{r=1,\cdots,j} \eta_{i_r}(u)\big\}\Big)$, where $I\{\cdot\}$ is the indicator function that $I\{True\} = 1$ and $I\{False\} = 0$, and ζ is the Lebesgue measure.

The classification result is shown in Table 3. All methods are able to provide reasonable classification power on the spike train data. The overall results on the first 5 seconds are better than the last 5 seconds. For the last 5 seconds, it can be seen that the modified h-depth has the best accuracy (76.32%) and F1 scores (75.68%, 76.92%). For the first 5 seconds, the modified h-depth, modified band depth and Mahalanobis depth all have the best accuracy (81.58%), where the modified h-depth is better on the post group (F1 score 82.05%) and the other two are better on the pre-group (F1 score 81.58%). In summary, we can see that the proposed smoothing depth framework provides relatively accurate classification in the practical point process observations.

Table 3
The classification result. F1 score (pre) and F1 score (post) are computed according to the recall and precision on pre-group classification and post-group classification. The boldface indicates best result in each column.

Method	Accuracy		F1 score (pre)		F1 score (post)	
	first 5 s	last 5 s	first 5 s	last 5 s	first 5 s	last 5 s
Classical h-depth	78.95%	65.79%	79.49%	61.76%	78.38%	69.05%
Modified h-depth	**81.58%**	**76.32%**	81.08%	**75.68%**	**82.05%**	**76.92%**
Band depth	78.95%	61.84%	77.14%	53.97%	80.49%	67.42%
Modified band depth	**81.58%**	71.05%	**81.58%**	72.50%	81.58%	69.44%
Mahalanobis depth	**81.58%**	73.68%	**81.58%**	72.97%	81.58%	74.36%

5. Summary and future work

In this paper, we at first proposed a kernel smoothing representation for point process observations. We found that the smoothing procedure builds a bijective mapping between them. We then defined a proper metric distance between the smoothing curves and transformed the problems from point processes to smooth functions. Based on the notion of h-depth on functions, we proposed two methods to form the depth structure on point process observations, i.e. the classical h-depth and the modified h-depth with center given. We then proposed a center-estimation method for the modified h-depth, with the idea of finding the Karcher mean through a combination of the RJMCMC annealing and line search. During the tests on the simulated data, both classical h-depth and modified h-depth resulted in reasonable outcomes: a center-outward decreasing depth structure that observations with smoothing curves closed to the center will have higher depth. In the real neuronal spike train data, we tested the new depth structures with a depth-based classification task and both methods resulted in accurate classifications.

Our investigation is only the starting point of the smoothing curve exploration and there is much to be further studied in the future. We have proposed the modified Gaussian kernel that satisfies the four basic requirements. We will explore other kernel functions that also satisfy the requirements. In the modified h-depth, a combination of RJMCMC annealing and line search has been given as the method to estimate the Karcher mean, which serves as the “center”. There should be other ways to define the “center” and other estimation methods. In this paper, two simulations using Poisson process have been done. More simulation studies based on other types of point processes such as Cox process will be explored in the future so that the proposed method can be demonstrated more comprehensively. These are all very good topics to further develop depth function on point process. In addition, the modified h-depth in the paper is just a simple way of using the Karcher mean. It only takes into consideration the distance between the input and the “center”, which is a kind of “mean”. More depth structures based on the idea of Karcher mean can be developed in the future to include more information about the data cloud, such as the “shape”. In Propositions 1 and 2, we have proven the continuity and inverse continuity. More in-depth study on the error bound with respect to number of events and event times would be an interesting direction to pursue. Finally, more functional depth methods, other than the h-depth, can be adopted to smoothed point processes, such as the band depth and modified band depth. Their ranking performance needs thorough investigations.

Alternatively to the smoothing procedure, we can explore other methods to define depth on point process data. For example, we may directly define a metric on point processes, irrespective to their cardinalities, and then define a new depth function using the metric. However, more careful work is needed to examine the computational efficiencies and mathematical properties of the new depth. A recent study on point process depth has utilized the equivalent representation by using inter-event times, and then defined a depth on the simplex domain of the inter-event times using Dirichlet distribution (Qi et al., 2021). We note that Dirichlet distribution may not properly characterize data points in simplex and will explore other distributions to better characterize the center-outward rank in the data.

Depth is a quantitative measure of how close a given observation is to the center of a data cloud. It is a generalization of the ranking process from univariate real numbers to more complex spaces. A data point with high depth can be seen as close to the “median” and vice versa. Therefore, any ranking-related task may be achieved through depth. In this paper, an example application of classification is included. There are also other possible applications for the h-depth and modified h-depth, such as outlier detection, quantile computation and nonparametric testing on point process observation space Ω. These applications will be further developed and studied in the future.

Acknowledgement

The authors would like to thank Robert Serfling for his motivating discussions and suggestion of the proposed study in this manuscript.

Appendix A. Proof of Lemma 1

The proof will go over the 4 requirements one-by-one:

1. Continuous and non-negative: this is clearly true since exponential function is positive and $c_1 > 0$. Also, polynomial and exponential function are continuous. As a composition of them, $K_G(x; T)$ is continuous and positive.
2. Positive at zero: $K_G(0; T) = c_1 > 0$.
3. Linearly independent with shifting:

- Proof of "$\Longrightarrow$"

By definition, $0 = \sum_{i=1}^{n} \alpha_i K_G(x - t_i; T) \doteq c_1 \sum_{i=1}^{n} \alpha_i e^{-\frac{c_2}{T^2}(x-t_i)^2}$. Multiplying $\frac{1}{c_1} e^{\frac{c_2}{T^2}(x-t_1)^2}$ on both sides, we have:

$$
\begin{aligned}
0 &\doteq \sum_{i=2}^{n} \alpha_i e^{\frac{c_2}{T^2}[(x-t_1)^2-(x-t_i)^2]} + \alpha_1 \\
&= \sum_{i=2}^{n} \alpha_i e^{\frac{c_2}{T^2}[t_1^2-t_i^2+2(t_i-t_1)x]} + \alpha_1 \\
&= \sum_{i=2}^{n} [\alpha_i e^{\frac{c_2}{T^2}(t_1^2-t_i^2)}][e^{\frac{2c_2}{T^2}(t_i-t_1)}]^x + \alpha_1 \\
&= \sum_{i=2}^{n} p_i q_i^x + p_1
\end{aligned}
$$

where $p_i = \alpha_i e^{\frac{c_2}{T^2}(t_1^2-t_i^2)}, q_i = e^{\frac{2c_2}{T^2}(t_i-t_1)}$ for $i = 2, 3, \cdots, n$ and $p_1 = \alpha_1$. Denote $g(x) = p_1 + \sum_{i=2}^{n} p_i q_i^x$, then as $g(x) \doteq 0$, the k-th order derivative, $g^{(k)}(x) \doteq 0$, which gives:

$$
\begin{aligned}
g(x) &= p_1 + p_2 q_2^x + p_3 q_3^x + \cdots + p_n q_n^x = 0 \\
g^{(1)}(x) &= p_2 q_2^x \ln q_2 + p_3 q_3^x \ln q_3 + \cdots + p_n q_n^x \ln q_n = 0 \\
g^{(2)}(x) &= p_2 q_2^x (\ln q_2)^2 + p_3 q_3^x (\ln q_3)^2 + \cdots + p_n q_n^x (\ln q_n)^2 = 0 \\
\vdots &\quad \vdots \\
g^{(n-1)}(x) &= p_2 q_2^x (\ln q_2)^{n-1} + p_3 q_3^x (\ln q_3)^{n-1} + \cdots + p_n q_n^x (\ln q_n)^{n-1} = 0
\end{aligned}
$$

In the form of matrix, these equations are equivalent to:

$$
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
0 & \ln q_2 & \ln q_3 & \cdots & \ln q_n \\
0 & (\ln q_2)^2 & (\ln q_3)^2 & \cdots & (\ln q_n)^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & (\ln q_2)^{n-1} & (\ln q_3)^{n-1} & \cdots & (\ln q_n)^{n-1}
\end{bmatrix}
\begin{bmatrix}
1 & & & & \\
& q_2^x & & & \\
& & q_3^x & & \\
& & & \ddots & \\
& & & & q_n^x
\end{bmatrix}
\begin{bmatrix}
p_1 \\ p_2 \\ \vdots \\ p_n
\end{bmatrix}
= 0
$$

As t_i is increasing, we have $1 < q_2 < \cdots < q_n$, so:

(a) $\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
0 & \ln q_2 & \ln q_3 & \cdots & \ln q_n \\
0 & (\ln q_2)^2 & (\ln q_3)^2 & \cdots & (\ln q_n)^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & (\ln q_2)^{n-1} & (\ln q_3)^{n-1} & \cdots & (\ln q_n)^{n-1}
\end{bmatrix}$ is full-rank, because $0 < \ln q_2 < \cdots < \ln q_n$ and a polynomial with order $n-1$ can have at most $n-1$ different roots.

(b) $\begin{bmatrix}
1 & & & & \\
& q_2^x & & & \\
& & q_3^x & & \\
& & & \ddots & \\
& & & & q_n^x
\end{bmatrix}$ is also full-rank, because $q_i^x > 0$ for all $i = 2, 3, \cdots, n$.

Therefore, $\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & \ln q_2 & \ln q_3 & \cdots & \ln q_n \\ 0 & (\ln q_2)^2 & (\ln q_3)^2 & \cdots & (\ln q_n)^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (\ln q_2)^{n-1} & (\ln q_3)^{n-1} & \cdots & (\ln q_n)^{n-1} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ & q_2^x & & & \\ & & q_3^x & & \\ & & & \ddots & \\ & & & & q_n^x \end{bmatrix}$ is full-rank and invertible. In this way, $\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = 0$, which means $p_i = 0$ for $i = 1, 2, \cdots, n$. As $e^{\frac{c_2}{T^2}(t_1^2 - t_i^2)} > 0$, we have $\alpha_1 = \alpha_2 = \cdots = 0$.

- Proof of "$\Longleftarrow$"

  If we have $\alpha_1 = \alpha_2 = \cdots = 0$, then $\sum_{i=1}^{n} \alpha_i K_G(x - t_i; T) \doteq \sum_{i=1}^{n} 0 \doteq 0$ for any x.

4. Scale invariance: By definition, the event time interval length T is a parameter of K_G. In addition, we have:

$$K_G(\alpha x; \alpha T) = c_1 e^{-\frac{c_2}{(\alpha T)^2}(\alpha x)^2} = c_1 e^{-\frac{c_2}{T^2}x^2} = K_G(x; T)$$

Since the modified Gaussian kernel satisfies all the 4 requirements, it is a proper kernel function. □

Appendix B. Proof of Lemma 2

The proof will show that the relation is both surjection and injection. In this way we know that it is a bijection.

- Surjection (Onto)

  According to the definition of the smoothing curve and the smoothing curve space, for any element f in $\mathbb{F}$, since $\mathbb{F} = \cup_{l=1}^{\infty} \mathbb{F}_l$, there exists $l \in \mathbb{N}^+$ such that $f \in \mathbb{F}_l = \{g_x(t) = \sum_{i=1}^{l} K(t - x_i; T) \quad | \quad x = (x_1, x_2, \cdots, x_l) \in \Omega_l\}$, so there exists one $x \in \Omega_l \subset \Omega$ such that $f = g_x$, the smoothing curve of x. Therefore, for any element f in $\mathbb{F}$, there exists one element x in Ω that f is the smoothing function of x, which means that the smoothing function is a surjection from Ω to $\mathbb{F}$.
- Injection (One-to-one)

  Now suppose any two observed point processes $x, y \in \Omega$ satisfy: $f_x(t) \doteq f_y(t)$ for any $t \in [0, T]$, where $f_x(t) = \sum_{i=1}^{n} K(t - x_i; T)$, $f_y(t) = \sum_{j=1}^{m} K(t - y_j; T) \in \mathbb{F}$. Taking the unique event time from x and y, denoted as $x_{0,1} < x_{0,2} < \cdots < x_{0,n_0}$ and $y_{0,1} < y_{0,2} < \cdots < y_{0,m_0}$, we can write $f_x(t)$, $f_y(t)$ as

$$f_x(t) = \sum_{i=1}^{n_0} a_i K(t - x_{0,i}; T) \quad f_y(t) = \sum_{j=1}^{m_0} b_j K(t - y_{0,j}; T)$$

  where $a_i, b_i \in \mathbb{N}^+$ represent the number of the occurrence for even time $x_{0,i}, y_{0,j}$ in $\{x_i\}_{i=1}^{n}, \{y_j\}_{j=1}^{m}$ respectively. Then we have:

$$\begin{aligned} 0 &\doteq f_x(t) - f_y(t) \\ &= \sum_{i=1}^{n_0} a_i K(t - x_{0,i}; T) - \sum_{j=1}^{m_0} b_j K(t - y_{0,j}; T) \\ &= \sum_{\substack{i \in \{1,2,\cdots,n_0\} \text{ s.t.} \\ \forall t = 1,2,\cdots,m, x_{0,i} \neq y_{0,t}}} a_i K(t - x_{0,i}; T) - \sum_{\substack{j \in \{1,2,\cdots,m_0\} \text{ s.t.} \\ \forall t = 1,2,\cdots,n, y_{0,j} \neq x_{0,t}}} b_j K(t - y_{0,j}; T) \\ &\quad + \sum_{\substack{i \in \{1,2,\cdots,n_0\} \\ j \in \{1,2,\cdots,m_0\} \\ \text{s.t.} x_{0,i} = y_{0,j}}} (a_i - b_j) K(t - x_{0,i}; T) \end{aligned} \tag{B.1}$$

  As shown in Eqn. (B.1), the set of event time in the three summations are disjoint. According to the requirement 3 for the kernel function K, Eqn. (B.1) means that all the coefficients in the three summations should be 0. Because $a_i, b_j > 0$, no event time $x_{0,i}, y_{0,j}$ is put into the first two summations. Therefore, we have $n_0 = m_0$ and $a_1 = b_1, a_2 = b_2, \cdots, a_{n_0} = b_{m_0}$, $x_{0,1} = y_{0,1}, x_{0,2} = y_{0,2}, \cdots, x_{0,n_0} = y_{0,m_0}$, so $\{x_i\}_{i=1}^{n} = \{y_j\}_{j=1}^{m}$ and $x = y$. In this way, for any two elements $f_x, f_y \in \mathbb{F}$ such that $f_x = f_y$, their corresponding elements in Ω should be the same, which means that the smoothing function is an injection from Ω to $\mathbb{F}$.

Therefore, the smoothing function is both surjection and injection, thus it is a bijective mapping from Ω to $\mathbb{F}$. □

Appendix C. Proof of Theorem 1

The proof will go over all the requirements for a proper metric.

- Non-negativity: $\|g\|_p \geq 0$ for any function g, so $d_{K,p}(t,s) = \|f_s - f_t\|_p \geq 0$;
- Identity of indiscernible: Lemma 2 indicates that for each curve f_t, there is one and only one t corresponding to it and also the inverse. Since $d_{K,p}(t,s) = 0$ is equivalent to $f_s = f_t$, which means $s = t$, we have $d_{K,p}(t,s) = 0 \iff t = s$;
- Symmetry: $d_{K,p}(t,s) = \|f_s - f_t\|_p = \|f_t - f_s\|_p = d_{K,p}(s,t)$;
- Triangle inequality: if u, s and t are 3 observed point processes from Ω, then $d_{K,p}(u,s) + d_{K,p}(t,s) = \|f_u - f_s\|_p + \|f_t - f_s\|_p \geq \|f_u - f_s + f_s - f_t\|_p = \|f_u - f_t\|_p = d_{K,p}(u,t)$;

Because all the requirements for a proper metric are satisfied, $d_{K,p}$ is a proper metric in Ω. □

Appendix D. Proof of Proposition 1

According to the definition of smoothing function, $f_{x^{(n)}}(t) = \sum_{i=1}^{k} K(t - x_i^{(n)}; T)$. k is a finite positive integer constant and K is continuous, so for any $t \in \mathbb{R}$, since $\lim_{n\to\infty} x_i^{(n)} = y_i$ for any $i = 1, 2, \cdots, k$, we have

$$\begin{aligned}\lim_{n\to\infty} f_{x^{(n)}}(t) &= \sum_{i=1}^{k} \lim_{n\to\infty} K(t - x_i^{(n)}; T) = \sum_{i=1}^{k} K(t - \lim_{n\to\infty} x_i^{(n)}; T) \\ &= \sum_{i=1}^{k} K(t - y_i; T) = f_y(t)\end{aligned}$$

which means $f_{x^{(n)}}$ converges to f_y point-wise as n goes to ∞.

Based on definition, K is continuous and non-negative, so on the closed bounded interval $[-T, T]$, by the boundedness theorem, $K(x; T)$ must be bounded and attains its bounds, which means $\exists M > 0$, such that $0 \leq K(x; T) \leq M$ for any $x \in [-T, T]$. Then for any term $n \in \mathbb{Z}^+$, each event time has $x_i^{(n)} \in [0, T]$, $t - x_i^{(n)} \in [-T, T]$ for any $t \in [0, T]$, so $0 \leq K(t - x_i^{(n)}; T) \leq M$ for any $t \in [0, T]$. Since k is finite, this indicates that for any $t \in [0, T]$ and any $n \in \mathbb{Z}^+$, $0 \leq f_{x^{(n)}}(t) = \sum_{i=1}^{k} K(t - x_i^{(n)}; T) \leq kM$, which means $f_{x^{(n)}}(t)$ is uniformly bounded by kM on $[0, T]$. Similarly, as $z = (z_1, z_2, \cdots, z_l)' \in \Omega$, for any $i = 1, 2, \cdots, l$, $z_i \in [0, T]$, thus $0 \leq K(t - z_i; T) \leq M$ and $0 \leq f_z(t) = \sum_{i=1}^{l} K(t - z_i; T) \leq lM$ for any $t \in [0, T]$. Therefore we have

$$0 \leq |f_{x^{(n)}}(t) - f_z(t)|^p \leq [|f_{x^{(n)}}(t)| + |f_z(t)|]^p \leq (k + l)^p M^p$$

for any $t \in [0, T]$ and any $n \in \mathbb{Z}^+$, by triangle inequality, which means $|f_{x^{(n)}}(t) - f_z(t)|^p$ is uniformly bounded. Then since $f_{x^{(n)}}$ converges to f_y point-wise as n goes to ∞, $|f_{x^{(n)}} - f_z|^p$ converges to $|f_y(t) - f_z(t)|^p$ point-wise as n goes to ∞ due to the continuity of the function $|.|^p$. By the dominated convergence theorem,

$$\lim_{n\to\infty} \int_0^T |f_{x^{(n)}}(t) - f_z(t)|^p dt = \int_0^T |f_y(t) - f_z(t)|^p dt$$

As function $(.)^{\frac{1}{p}}$ is continuous on $[0, +\infty)$,

$$\begin{aligned}\lim_{n\to\infty} d_{K,p}(x^{(n)}, z) = \lim_{n\to\infty} \|f_{x^{(n)}} - f_z\|_p = \lim_{n\to\infty} [\int_0^T |f_{x^{(n)}}(t) - f_z(t)|^p dt]^{\frac{1}{p}} \\ = [\lim_{n\to\infty} \int_0^T |f_{x^{(n)}}(t) - f_z(t)|^p dt]^{\frac{1}{p}} = [\int_0^T |f_y(t) - f_z(t)|^p dt]^{\frac{1}{p}} \\ = \|f_y - f_z\|_p = d_{K,p}(y, z)\end{aligned}$$

which finishes the proof. □

Appendix E. Proof of Proposition 2

The proof will be provided in 3 steps:

- Step (i): We will show by contradiction that the dimension k_n is bounded when n is large, i.e. $\exists M > 0$ and $N > 0$, such that $\forall n \geq N$, $k_n \leq M$.

  Suppose the opposite statement is true, in other words:

$$\forall M > 0 \quad \text{and} \quad N > 0, \quad \exists n_0 \geq N, \quad \text{such that} \quad k_n > M. \tag{E.1}$$

  Because K is continuous, non-negative and $K(0;T) > 0$, we have:

$$\min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\} = J > 0$$

  because there exists a neighborhood of zero that K is positive due to continuity. As the event time interval is $[0.T]$, by Holder inequality we have

$$\|f_{x^{(n)}} \cdot 1\|_1 \leq \|f_{x^{(n)}}\|_p \|1\|_{\frac{p}{p-1}} = \|f_{x^{(n)}}\|_p T^{\frac{p-1}{p}}$$

  which gives:

$$\|f_{x^{(n)}}\|_p \geq T^{-\frac{p-1}{p}} \|f_{x^{(n)}}\|_1 = T^{-\frac{p-1}{p}} \int_0^T |\sum_{i=1}^{k_n} K(t - x_i^{(n)};T)|dt$$

$$= T^{-\frac{p-1}{p}} \sum_{i=1}^{k_n} \int_{-x_i^{(n)}}^{T-x_i^{(n)}} K(t;T)dt$$

  For any $i = 1, 2, \cdots, n$, $x_i^{(n)}$ is in $[0,T]$, so $\max\{x_i^{(n)}, T - x_i^{(n)}\} \geq (x_i^{(n)} + T - x_i^{(n)})/2 = T/2$, which means at least one of $x_i^{(n)}$ and $T - x_i^{(n)}$ is larger or equal to $T/2$. Based on this we have:

$$\int_{-x_i^{(n)}}^{T-x_i^{(n)}} K(x;T)dx = \int_{-x_i^{(n)}}^{0} K(x;T)dx + \int_0^{T-x_i^{(n)}} K(x;T)dx$$

  - If $T - x_i^{(n)} \geq T/2$, then $\int_0^{T-x_i^{(n)}} K(x;T)dx \geq \int_0^{T/2} K(x;T)dx \geq \min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^0 K(x;T)dx\}$;
  - If $x_i^{(n)} \geq T/2$, then $\int_{-x_i^{(n)}}^0 K(x;T)dx \geq \int_{-T/2}^0 K(x;T)dx \geq \min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^0 K(x;T)dx\}$;

  Both $\int_{-x_i^{(n)}}^0 K(x;T)dx$ and $\int_0^{T-x_i^{(n)}} K(x;T)dx$ are non-negative since K is non-negative, so:

$$\int_{-x_i^{(n)}}^{T-x_i^{(n)}} K(x;T)dx \geq \min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\} = J > 0$$

  for any $i = 1, 2, \cdots, n$. Thus,

$$\|f_{x^{(n)}}\|_p \geq T^{-\frac{p-1}{p}} \sum_{i=1}^{k_n} \int_{-x_i^{(n)}}^{T-x_i^{(n)}} K(t;T)dt \geq k_n J T^{-\frac{p-1}{p}} > 0$$

  Then by triangle inequality,

$$d_{K,p}(x^{(n)}, y) = \|f_{x^{(n)}} - f_y\|_p \geq \|f_{x^{(n)}}\|_p - \|f_y\|_p \geq k_n J T^{-\frac{p-1}{p}} - \|f_y\|_p$$

So $\lim_{n\to\infty} d_{K,p}(x^{(n)}, y) = 0$ gives: $\forall \epsilon > 0$, $\exists N_0 > 0$ such that $\forall n \geq N_0$, $d_{K,p}(x^{(n)}, y) < \epsilon$, but by (E.1), assigning $N = N_0$, we have $\forall M > 0$, $\exists n_0 \geq N_0$, such that $k_{n_0} > M$, thus $d_{K,p}(x^{(n_0)}, y) \geq k_{n_0} J T^{-\frac{p-1}{p}} - \|f_y\|_p > M J T^{-\frac{p-1}{p}} - \|f_y\|_p$. These two relations result in: $\forall \epsilon, M > 0$, we have $\epsilon > d_{K,p}(x^{(n_0)}, y) > M J T^{-\frac{p-1}{p}} - \|f_y\|_p$, which is clearly false that a counterexample can be $M = J^{-1} T^{\frac{p-1}{p}} (1 + \epsilon + \|f_y\|_p) > 0$, so we come to a contradiction. Therefore, we have $\exists M > 0$ and $N > 0$, such that $\forall n \geq N$, $k_n \leq M$.

- Step (ii): After we see that the dimension of $x^{(n)}$ will be bounded by M as n becomes larger, we will consider only the sequence after the $[N] + 1$ term, i.e. $\{x^{(n)}\}_{n=[N]+1}^{\infty}$. In this way, the dimension of each term, k_n, is non-larger than M, so k_n can only take $[M]$ many different values: $1, 2, \cdots, [M]$. Then there must exist a sub-sequence $\{x^{(n_r)}\}_{r=1}^{\infty}$ such that the dimension for each term is the same k_0, i.e. $x^{(n_r)} = (x_1^{(n_r)}, x_2^{(n_r)}, \cdots, x_{k_0}^{(n_r)})'$ for $r = 1, 2, \cdots$. Note that so far there might be more than one k_0 having an infinitely long sub-sequence. We will now show that the only possible value for k_0 is k, i.e. $k_0 = k$.
  For any infinite-long sub-sequence of $\{x^{(n)}\}_{n=1}^{\infty}$ such that each term has the same dimension k_0, we denote it as $\{x^{(n_r)}\}_{r=1}^{\infty}$. Since $x_i^{(n_r)}$ is within $[0, T]$ for any $i = 1, 2, \cdots, k_0$, $\{x^{(n_r)}\}_{r=1}^{\infty}$ is a infinite sequence defined on a closed and bounded set, so there exists a sub-sequence of $\{x^{(n_r)}\}_{r=1}^{\infty}$, which is also a sub-sequence of $\{x^{(n)}\}_{n=1}^{\infty}$, that converges to a vector in $[0, T]^{k_0}$. Denote the sub-sequence to be $\{x^{(n_{r_j})}\}_{j=1}^{\infty}$ and the vector it converges to be $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_{k_0}^{(0)})'$, so $\lim_{j\to\infty} x^{(n_{r_j})} = x^{(0)}$ i.e. $\lim_{j\to\infty} x_i^{(n_{r_j})} = x_i^{(0)}$ for any $i = 1, 2, \cdots, k_0$. Because x to f_x is a bijection, $d_{K,p}(x, y)$ is a function of x and thus $\{d_{K,p}(x^{(n_{r_j})}, y)\}_{j=1}^{\infty}$ is a sub-sequence of $\{d_{K,p}(x^{(n)}, y)\}_{n=1}^{\infty}$, so

$$\lim_{j\to\infty} d_{K,p}(x^{(n_{r_j})}, y) = \lim_{n\to\infty} d_{K,p}(x^{(n)}, y) = 0$$

  Then according to Proposition 1, since $\lim_{j\to\infty} x^{(n_{r_j})} = x^{(0)}$ with the same dimension k_0, we have:

$$0 = \lim_{j\to\infty} d_{K,p}(x^{(n_{r_j})}, y) = d_{K,p}(x^{(0)}, y) = \|f_{x^{(0)}} - f_y\|_p$$

  Since $\|g\|_p = 0 \iff g \doteq 0$ almost everywhere and $\forall s \in \Omega$, its smoothing function $f_s(t)$ is continuous for $\forall t \in \mathbb{R}$, we have $f_{x^{(0)}}(t) - f_y(t) \doteq 0$, i.e. $f_{x^{(0)}}(t) \doteq f_y(t)$ for any t. Because x to f_x is a bijection, this means $x^{(0)} = y$, thus $\dim(x^{(0)}) = k_0 = k = \dim(y)$. In this way, we have shown that any infinite long sub-sequence of $\{x^{(n)}\}_{n=1}^{\infty}$, where the dimension of each term is the same, must have the dimension to be k.
  As mentioned before, for $\{x^{(n)}\}_{n=[N]+1}^{\infty}$, the dimension of each term can only be one of $1, 2, \cdots, [M]$. By the conclusion we have just obtained, for any $k_* \in \{1, 2, \cdots, [M]\}$ and $k_* \neq k$, there are only finitely many $x^{(n)}$s that have dimension to be k_*. Therefore, $\exists N_* > 0$, such that $\forall n \geq N_*$, $x^{(n)}$ has dimension to be k. Denote $\tilde{N} = [\max\{N, N_*\}] + 1$, then $\{x^{(n)}\}_{n=\tilde{N}}^{\infty}$ is a k-dimensional vector sequence, with each term to be $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \cdots, x_k^{(n)})'$.
- Step (iii): For the last step, we will show by contradiction that $\{x^{(n)}\}_{n=1}^{\infty}$ converges to y.
  Suppose it does not converge to y, then by definition, we have: $\exists A > 0$, such that $\forall L > 0$, $\exists n_0 > L$, $\|x^{(n_0)} - y\| \geq A$ where $\|.\|$ is any vector norm. Setting $L \geq \tilde{N}$, we can find a sub-sequence of $\{x^{(n)}\}_{n=1}^{\infty}$, denoted as $\{x^{(n_z)}\}_{z=1}^{\infty}$, that each term has dimension to be k and satisfies $\|x^{(n_z)} - y\| \geq A$. By definition, each term of $\{x^{(n_z)}\}_{z=1}^{\infty}$ is defined on a closed and bounded set $\Psi = \{(x_1^{(n_z)}, x_2^{(n_z)}, \cdots, x_k^{(n_z)})' \in [0, T]^k \;\mid\; \|x^{(n_z)} - y\| \geq A\}$, so it must have a sub-sequence $\{x^{(n_{z_j})}\}_{j=1}^{\infty}$ that converges in Ψ. We denote the vector that $x^{(n_{z_j})}$ converges to be $x^{(C)}$, i.e. $\lim_{j\to\infty} x^{(n_{z_j})} = x^{(C)} = (x_1^{(C)}, x_2^{(C)}, \cdots, x_k^{(C)})' \in \Psi$.
  As discussed above, since x to f_x is a bijection, $\{d_{K,p}(x^{(n_{z_j})}, y)\}_{j=1}^{\infty}$ is a sub-sequence of $\{d_{K,p}(x^{(n)}, y)\}_{n=1}^{\infty}$, so we have:

$$0 = \lim_{n\to\infty} d_{K,p}(x^{(n)}, y) = \lim_{j\to\infty} d_{K,p}(x^{(n_{z_j})}, y)$$

  As $\lim_{j\to\infty} x^{(n_{z_j})} = x^{(C)}$ with consistent dimension k, by Proposition 1, we have

$$0 = \lim_{j\to\infty} d_{K,p}(x^{(n_{z_j})}, y) = d_{K,p}(x^{(C)}, y) = \|f_{x^{(C)}} - f_y\|_p$$

  Therefore, we have $\|f_{x^{(C)}} - f_y\|_p = 0$ which means $f_{x^{(C)}}(t) \doteq f_y(t)$ for any t, i.e. $x^{(C)} = y$ since x to f_x is a bijection. Then $\|x^{(C)} - y\| = 0 < A$, contradicts with $x^{(C)} \in \Psi$, so we conclude that $\{x^{(n)}\}_{n=1}^{\infty}$ converges to y.
  In this way, both the convergence in dimension: $k_n = k$ for n sufficiently large and convergence in each element: $\lim_{n\to\infty} x_i^{(n)} = y_i, \forall i = 1, 2, \cdots, k$ are proved. □

Appendix F. Proof of Lemma 3

For any functions $x, y \in \Lambda$, for the continuity of $HD(.)$, we need to prove $|HD(x) - HD(y)| \to 0$ if $x \to y$.

$$|HD(x) - HD(y)| = |\mathbb{E}(\exp\{-\frac{\|x - X\|^2}{2h}\} - \exp\{-\frac{\|y - X\|^2}{2h}\})|$$

At the right side of the equation, since h is a constant which is greater than 0, we just need to consider $\lim_{x\to y} |\mathbb{E}(\exp\{-\|x - X\|^2\} - \exp\{-\|y - X\|^2\})|$. For function $f(x) = e^{-x}$, based on mean value theorem, there is a δ between x and y s.t. $|f(x) - f(y)| = |f'(\delta)(x - y)| = |-e^{-\delta}(x - y)|$. Since in this situation $\delta \geq 0$, $|f(x) - f(y)| \leq |x - y|$.

$$\begin{aligned}
&|\mathbb{E}(\exp\{-\|x - X\|^2\} - \exp\{-\|y - X\|^2\})| \\
\leq\ & \mathbb{E}|\exp\{-\|x - X\|^2\} - \exp\{-\|y - X\|^2\}| \\
\leq\ & \mathbb{E}|\|x - X\|^2 - \|y - X\|^2| \\
=\ & \mathbb{E}|\|x\|^2 - \|y\|^2 - 2\langle x - y, X\rangle| \\
\leq\ & |\|x\|^2 - \|y\|^2| + 2\mathbb{E}|\langle x - y, X\rangle|
\end{aligned}$$

Because $x \to y$, $|\|x\|^2 - \|y\|^2| = 0$. By Cauchy inequality,

$$\mathbb{E}|\langle x - y, X\rangle| \leq \mathbb{E}(\|x - y\| * \|X\|) = \|x - y\|\mathbb{E}\|X\|$$

It equals to 0 when $x \to y$ and $\mathbb{E}\|X\| < \infty$. Therefore, $\lim_{x\to y} |HD(x) - HD(y)| = 0$ when $\mathbb{E}\|X\| < \infty$. □

Appendix G. Proof of Proposition 3

We prove these properties in the following three steps:

1. Continuity
   Based on the continuity of h-depth in Lemma 3, we only need to prove $\mathbb{E}\|f_S\| < \infty$. Let $|S|$ denote the number of events in S. Then $S = (S_1, \cdots, S_{|S|})$. Using the law of iterated expectation, we have

$$\mathbb{E}\|f_S\| = \mathbb{E}(\mathbb{E}(\|f_S\|\,||S|)),$$

   where the conditional expectation is given as:

$$\begin{aligned}
\mathbb{E}(\|f_S\|\,||S|) &= \mathbb{E}(\sqrt{\int_0^T [\sum_{i=1}^{|S|} K(t - S_i; T)]^2 dt}\quad ||S|) \\
&\leq \mathbb{E}(\sqrt{|S| \sum_{i=1}^{|S|} \int_0^T K(t - S_i; T)^2(t) dt}\quad ||S|) \\
&\leq c\mathbb{E}(|S|)
\end{aligned}$$

   where c is a positive constant, dependent on the kernel function K only. Now it is apparent that if $\mathbb{E}(|S|) < \infty$, then $E(\|z_S\|) \leq \mathbb{E}(c\mathbb{E}(|S|)) = c\mathbb{E}(|S|) < \infty$. Therefore, $D(s; P_S)$ is continuous with respect to s.
2. Linear Invariance
   For any $s = (s_1, \cdots, s_k) \in \Omega$, the smoothed process is $f_s(t) = \sum_{i=1}^k K(t - s_i; T)$. For scale coefficient $a > 0$ and translation $b \in \mathbb{R}$, the interval $[0, T]$ is transformed to $[b, aT + b]$ with length aT. Then,

$$\begin{aligned}
&\|f_{as+b} - f_{aS+b}\|^2 \\
=\ & \int_b^{aT+b} \left(\sum_{i=1}^k K(t - (as_i + b); aT) - \sum_{i=1}^{|S|} K(t - (aS_i + b); aT) \right)^2 dt \\
=\ & a\int_0^T \left(\sum_{i=1}^k K(au + b - (as_i + b); aT) - \sum_{i=1}^{|S|} K(au + b - (aS_i + b); aT) \right)^2 du
\end{aligned}$$

$$
\begin{aligned}
&= a\int_0^T \left(\sum_{i=1}^{k} K(a(u-s_i);aT) - \sum_{i=1}^{|S|} K(a(u-S_i);aT) \right)^2 du \\
&= a\int_0^T \left(\sum_{i=1}^{k} K(u-s_i;T) - \sum_{i=1}^{|S|} K(u-S_i;T) \right)^2 du \\
&= a\|f_s - f_S\|^2
\end{aligned}
$$

where the second last equality is due to the scale invariance of the smoothing kernel K. The hyper-parameter h is CT before the transformation, and aCT after the transformation, where $C > 0$ is a constant. Therefore,

$$
\begin{aligned}
D(as+b; P_{aS+b}) &= \mathbb{E}(G_{aT}(\|f_{as+b} - f_{aS+b}\|)) = \mathbb{E}(\exp(-\frac{\|f_{as+b} - f_{aS+b}\|^2}{2aCT})) \\
&= \mathbb{E}(\exp(-\frac{a\|f_s - f_S\|^2}{2aCT})) = \mathbb{E}(G_T(\|f_s - f_S\|)) \\
&= D(s; P_S).
\end{aligned}
$$

3. Vanishing at Infinity
For kernel function K, it is continuous, non-negative and $K(0;T) > 0$, we have:

$$
\min\{\int_0^{\frac{T}{2}} K(t;T)dt, \int_{-\frac{T}{2}}^{0} K(t;T)dt\} = J > 0
$$

because there exists a neighborhood of zero that K is positive due to continuity. Assume $s = (s_1, s_2, \cdots, s_k) \in \Omega$ on $[0, T]$, by Holder inequality we have

$$
\|f_s \cdot 1\|_1 \leq \|f_s\|_2 \|1\|_2 = \|f_s\|_2 T^{\frac{1}{2}}
$$

which gives:

$$
\|f_s\|_2 \geq T^{-\frac{1}{2}} \|f_s\|_1 = T^{-\frac{1}{2}} \int_0^T |\sum_{i=1}^{k} K(t-s_i;T)|dt = T^{-\frac{1}{2}} \sum_{i=1}^{k} \int_{-s_i}^{T-s_i} K(t;T)dt
$$

For any $i = 1, 2...k$, $\max\{s_i, T - s_i\} \geq (s_i + T - s_i)/2 = T/2$.

$$
\int_{-s_i}^{T-s_i} K(t;T)dt = \int_{-s_i}^{0} K(t;T)dt + \int_0^{T-s_i} K(t;T)dt
$$

- If $T - s_i \geq T/2$, then $\int_0^{T-s_i} K(t;T)dt \geq \int_0^{T/2} K(t;T)dt \geq \min\{\int_0^{\frac{T}{2}} K(t;T)dt, \int_{-\frac{T}{2}}^{0} K(t;T)dt\}$;
- If $s_i \geq T/2$, then $\int_{-s_i}^{0} K(t;T)dt \geq \int_{-T/2}^{0} K(t;T)dt \geq \min\{\int_0^{\frac{T}{2}} K(t;T)dt, \int_{-\frac{T}{2}}^{0} K(t;T)dt\}$;

Both $\int_{-s_i}^{0} K(t;T)dt$ and $\int_0^{T-s_i} K(t;T)dt$ are non-negative since K is non-negative, so:

$$
\int_{-s_i}^{T-s_i} K(t;T)dt \geq \min\{\int_0^{\frac{T}{2}} K(t;T)dt, \int_{-\frac{T}{2}}^{0} K(t;T)dt\} = J > 0
$$

Thus,

$$
\|f_s\|_2 \geq T^{-\frac{1}{2}} \sum_{i=1}^{k} \int_{-s_i}^{T-s_i} K(t;T)dt \geq kJT^{-\frac{1}{2}} > 0
$$

Now in h-depth, S is a random point process on $[0,T]$ in the probability space $(\Omega, \mathcal{F}, P_S)$. By triangle inequality,

$$\|f_s - f_S\|_2 \geq \|f_s\|_2 - \|f_S\|_2 \geq kJT^{-\frac{1}{2}} - \|f_S\|_2$$

$\|f_s - f_S\|_2 \to \infty$ if $k \to \infty$. Because $\exp\{-\frac{\|f_s - f_S\|^2}{2h}\} \leq 1$, fix S, $\exp\{-\frac{\|f_s - f_S\|^2}{2h}\} \to 0$ if $k \to \infty$, it converges pointwise. Based on dominant convergence theorem,

$$\lim_{k\to\infty} \int \exp\{-\frac{\|f_s - f_S\|^2}{2h}\} df_S = 0$$

Therefore, $\lim_{k\to\infty} \mathbb{E}(G_h(\|f_s - f_S\|)) = 0$, $D(s; P_S) \to 0$ when $|s| \to \infty$

To show why ***P4*** and ***P5*** do not hold for h-depth, consider the counter example below:

A process provides either 1 event or 2 events with $P(n=1) = \frac{2}{3}$ and $P(n=2) = \frac{1}{3}$ on event time interval $[0,1]$. Given that $n=1$, the event time takes value 0 or 1 equal likely: $P(S=(0)|n=1) = P(S=(1)|n=1) = \frac{1}{2}$ and given $n=2$, the event time takes the value $S=(0,1)'$ with probability 1. If we use the modified Gaussian kernel mentioned before with $\epsilon_1 = \epsilon_2 = 1$ to do the smoothing, then as the length of the time interval is $T=1$, we have $K(x;T) = e^{-x^2}$ and

$$\Omega = \{(0), (1), (0,1)'\} \quad \mathbb{F} = \{f_0(x) = e^{-x^2}, f_1(x) = e^{-(x-1)^2}, f_{0,1}(x) = e^{-x^2} + e^{-(x-1)^2}\}$$

Using the classical h-depth, choosing $h = T = 1$, we have $\forall s \in \Omega$, $D(s; P_S) = \mathbb{E}(e^{-\frac{1}{2}\|f_s - f_S\|^2})$. As $P(S=(0)) = P(S=(1)) = P(S=(0,1)') = \frac{1}{3}$, the expectation can be computed as:

$$D(s; P_S) = \frac{1}{3}[e^{-\frac{1}{2}\|f_s - f_0\|^2} + e^{-\frac{1}{2}\|f_s - f_1\|^2} + e^{-\frac{1}{2}\|f_s - f_{0,1}\|^2}]$$

Since Ω only contains 3 elements, computing every possible s numerically, we have:

- $D((0); P_S) = \frac{1}{3}[1 + e^{-\frac{1}{2}\|f_0 - f_1\|^2} + e^{-\frac{1}{2}\|f_0 - f_{0,1}\|^2}] = 0.8885$;
- $D((1); P_S) = \frac{1}{3}[1 + e^{-\frac{1}{2}\|f_1 - f_0\|^2} + e^{-\frac{1}{2}\|f_1 - f_{0,1}\|^2}] = 0.8885$;
- $D((0,1)'; P_S) = \frac{1}{3}[1 + e^{-\frac{1}{2}\|f_{0,1} - f_0\|^2} + e^{-\frac{1}{2}\|f_{0,1} - f_1\|^2}] = 0.8277$;

As shown, both (0) and (1) in Ω can maximize the depth function, which indicates that the center in this case is not unique. Therefore, h-depth may not satisfy ***P4*** and ***P5*** for all cases. □

Appendix H. Proof of Proposition 4

The proof for each property of the center-based h-depth is shown below:

1. Continuity
Suppose $s \in \Omega$ is a given observed event time vector and $s_c \in \Omega$ denote the pre-determined center. $t \in \Omega$ is the input event time vector, changing t such that $d_{K,2}(s,t) \to 0$, i.e. $\|f_s - f_t\| \to 0$ by definition of $d_{K,2}$. Then by triangle inequality,

$$\begin{aligned}
0 \leq |D(s; s_c) - D(t; s_c)| &= |\exp(-\frac{\|f_s - f_{s_c}\|^2}{2h}) - \exp(-\frac{\|f_t - f_{s_c}\|^2}{2h})| \\
&= e^{-\frac{\|f_s - f_{s_c}\|^2}{2h}} |1 - e^{-\frac{1}{2h}(\|f_t - f_{s_c}\| - \|f_s - f_{s_c}\|)(\|f_t - f_{s_c}\| + \|f_s - f_{s_c}\|)}| \\
&\leq e^{-\frac{\|f_s - f_{s_c}\|^2}{2h}} |1 - e^{-\frac{1}{2h}\|f_t - f_s\|(\|f_t - f_s + f_s - f_{s_c}\| + \|f_s - f_{s_c}\|)}| \\
&\leq e^{-\frac{\|f_s - f_{s_c}\|^2}{2h}} |1 - e^{-\frac{1}{2h}\|f_t - f_s\|(\|f_t - f_s\| + 2\|f_s - f_{s_c}\|)}|
\end{aligned}$$

As $\|f_s - f_t\| \to 0$ and $|s| < \infty$, $e^{-\frac{1}{2h}\|f_t - f_s\|(\|f_t - f_s\| + 2\|f_s - f_{s_c}\|)} \to 1$, so the right hand side will converge to 0. By squeeze theorem, we have $D(t; s_c) \to D(s; s_c)$ as $d_{K,2}(s,t) \to 0$, thus the continuity of $D(s, s_c)$ is shown.
2. Linear Invariance
This property is a consequence of the condition 4 on $K(\cdot; T)$ that $K(x; T) \doteq K(\alpha x; \alpha T)$ for any constant $\alpha \in \mathbb{R}^+$. Under the given information of $[0,T]$ being the event time interval, $s_c = (s_{c,1}, s_{c,2}, \cdots, s_{c,n_c})'$ being the center event time vector and $t = (t_1, t_2, \cdots, t_n)'$ being an observed event time vector, a linear transformation, with scaling parameter α and shifting parameter β, will change both the interval and vector, to $[\beta, \alpha T + \beta]$, $\tilde{s_c} = (\alpha s_{c,1} + \beta, \alpha s_{c,2} + \beta, \cdots, \alpha s_{c,n_c} + \beta)'$ and $\tilde{t} = (\alpha t_1 + \beta, \alpha t_2 + \beta, \cdots, \alpha t_n + \beta)'$. We have

$$\|f_{\tilde{t}} - f_{\tilde{s_c}}\|^2 = \int_{\beta}^{\alpha T+\beta} |f_{\tilde{t}}(x) - f_{\tilde{s_c}}(x)|^2 dx$$

$$= \int_{\beta}^{\alpha T+\beta} |f_{\tilde{t}}(\alpha y+\beta) - f_{\tilde{s_c}}(\alpha y+\beta)|^2 d(\alpha y+\beta)$$

$$= \alpha \int_{\beta}^{\alpha T+\beta} |\sum_{i=1}^{n} K(\alpha y+\beta-\alpha t_i-\beta;\alpha T) - \sum_{i=1}^{n_c} K(\alpha y+\beta-\alpha s_{c,i}-\beta;\alpha T)|^2 dy$$

$$= \alpha \int_{\beta}^{\alpha T+\beta} |\sum_{i=1}^{n} K(y-t_i;T) - \sum_{i=1}^{n_c} K(y-s_{c,i};T)|^2 dy$$

$$= \alpha \int_{\beta}^{\alpha T+\beta} |f_t - f_{s_c}|^2 dy$$

$$= \alpha \|f_t - f_{s_c}\|^2$$

Thus, when $h = CT$ with constant $C > 0$, the center-based h-depth will be

$$D(\tilde{t},\tilde{s_c}) = \exp(-\frac{\|f_{\tilde{t}} - f_{\tilde{s_c}}\|^2}{2\alpha CT}) = \exp(-\frac{\|f_t - f_{s_c}\|^2}{2CT}) = D(t,s_c)$$

In this way, applying a linear transformation will not change the values of the modified h-depth with center given, thus the linear invariance is proved.

3. Vanishing at Infinity
The proof is nearly the same as the first part of the proof for Proposition 2.
Since K is continuous and non-negative, together with $K(0;T) > 0$, we have:

$$\min\{\int_{0}^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\} = J > 0$$

because there exists a neighborhood of zero that K is positive, due to continuity. Then we have:

$$\int_{0}^{T} f_t(x)dx = \sum_{i=1}^{n} \int_{0}^{T} K(x-t_i;T)dx = \sum_{i=1}^{n} \int_{-t_i}^{T-t_i} K(x;T)dx$$

For any $i = 1, 2, \cdots, n$, t_i is in $[0, T]$, so:

$$\max\{t_i, T-t_i\} \geq (t_i + T - t_i)/2 = T/2$$

which means at least one of t_i and $T - t_i$ is larger or equal to $T/2$. As K is non-negative, we have:

$$\int_{-t_i}^{T-t_i} K(x;T)dx = \int_{-t_i}^{0} K(x;T)dx + \int_{0}^{T-t_i} K(x;T)dx$$

- If $T - t_i \geq T/2$, then $\int_0^{T-t_i} K(x;T)dx \geq \int_0^{T/2} K(x;T)dx \geq \min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\}$;
- If $t_i \geq T/2$, then $\int_{-t_i}^{0} K(x;T)dx \geq \int_{-T/2}^{0} K(x;T)dx \geq \min\{\int_0^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\}$;

Both $\int_{-t_i}^{0} K(x;T)dx$ and $\int_{0}^{T-t_i} K(x;T)dx$ are non-negative, so:

$$\int_{-t_i}^{T-t_i} K(x;T)dx \geq \min\{\int_{0}^{\frac{T}{2}} K(x;T)dx, \int_{-\frac{T}{2}}^{0} K(x;T)dx\} = J > 0$$

for any $i = 1, 2, \cdots, n$. Thus,

$$\int_{0}^{T} f_t(x)dx = \sum_{i=1}^{n} \int_{-t_i}^{T-t_i} K(x;T)dx \geq nJ > 0$$

which results in $\int_0^T f_t(x)dx \to \infty$ when $n \to \infty$. Because f_t is non-negative, it is equivalent to: as $n \to \infty$, $\int_0^T |f_t(x)|dx = \|f_t\|_1 \to \infty$, so based on Holder inequality, $\|f_t\| \geq \frac{1}{\sqrt{T}}\|f_t\|_1 \to \infty$. Therefore, by triangle inequality, $0 \leq D(t, s_c) \leq e^{-\frac{1}{2h}(\|f_t\| - \|f_{s_c}\|)^2} \to 0$ as $n \to \infty$. Then by squeeze theorem, $D(t, s_c) \to 0$ as $n \to \infty$. In this way, we have shown that the center-based h-depth will vanish at infinity.

4. Maximum at the Center
Based on the definition of $\mathbb{L}^2$ norm, $\|f_t - f_{s_c}\| \geq 0$ for any $f_t \in \mathbb{F}$. The equivalence can be achieved if and only if $f_t - f_{s_c} \doteq 0$, i.e. $f_t \doteq f_{s_c}$. Thus,

$$D(t, s_c) = \exp(-\frac{\|f_t - f_{s_c}\|^2}{2h}) \leq 1$$

if and only if $f_t = f_{s_c}$ the maximum can be achieved. According to Lemma 2, t to f_t is a bijective mapping. Therefore, the center s_c is the only one element in Ω that can achieve the maximum.

5. Monotone Decreasing from the Center
Based on the definition of the metric $d_{K,2}$ in Ω, suppose s_1 and s_2 are any two observed event time vectors in Ω with $d_{K,2}(s_1, s_c) < d_{K,2}(s_2, s_c)$ where $s_c \in \Omega$ is the pre-defined center, then we have $\|f_{s_1} - f_{s_c}\| < \|f_{s_2} - f_{s_c}\|$. Because $D(t, s_c)$ is a monotone decreasing function with respect to $\|f_t - f_{s_c}\|$, we have $D(s_1, s_c) > D(s_2, s_c)$, which shows the monotone decreasing relation. □

Appendix I. Proof of Theorem 2

By definition, $\bar{S}_K^{(N)} = \arg\min_{t \in \Omega} \sum_{i=1}^{N} d_{K,2}^2(t, S_i)$ where $\bar{S}_K^{(N)}$ represents the set of all the minimization solutions. According to triangle inequality,

$$d_{K,2}(t, S_i) \geq d_{K,2}(t, \phi_0) - d_{K,2}(S_i, \phi_0)$$

where ϕ_0 is the event time vector with no event, as defined in Section 2.1.2. If the statement "for any i, $d_{K,2}(t, \phi_0) > 2d_{K,2}(S_i, \phi_0)$" is true, then we have $d_{K,2}(t, S_i) > d_{K,2}(S_i, \phi_0)$ for any i, which means t is not the minimum point since it results in larger SSD than ϕ_0, so we have the relation:

$$\forall i, d_{K,2}(t, \phi_0) > 2d_{K,2}(S_i, \phi_0) \implies t \notin \bar{S}_K^{(N)}$$

For the left hand side left part,

$$\begin{aligned} & d_{K,2}(t, \phi_0) \\ & = \|f_t - f_{\phi_0}\|_2 = \|\sum_{j=1}^{|t|} K(\cdot - t_j; T)\|_2 \\ & = \sqrt{\int_0^T [\sum_{j=1}^{|t|} K(x - t_j; T)]^2 dx} \geq \sqrt{\int_0^T \sum_{j=1}^{|t|} K^2(x - t_j; T) dx} \\ & = \sqrt{\sum_{j=1}^{|t|} \int_0^T K^2(x - t_j; T) dx} = \sqrt{\sum_{j=1}^{|t|} \|K(\cdot - t_j; T)\|_2^2} \\ & \geq \sqrt{|t|}\delta \end{aligned} \tag{I.1}$$

where $\delta = \min_{r\in[0,T]} \|K(\cdot - r; T)\|_2$ and (I.1) is due to $K \geq 0$. Since the kernel K is continuous and non-negative on $[0, T]$, δ must exist. $|t|$ is the dimension of t and t_j is the j-th event time in t.
For the left hand side right part,

$$
\begin{aligned}
2d_{K,2}(S_i, \phi_0) &= 2\|f_{S_i} - f_{\phi_0}\|_2 = 2\|\sum_{j=1}^{|S_i|} K(\cdot - S_{ij}; T)\|_2 \\
&= 2\sqrt{\int_0^T [\sum_{j=1}^{|S_i|} K(x - S_{ij}; T)]^2 dx} && \text{(I.2)} \\
&\leq 2\sqrt{\int_0^T |S_i|\sum_{j=1}^{|S_i|} K^2(x - S_{ij}; T) dx} && \text{(I.3)} \\
&= 2\sqrt{|S_i|\sum_{j=1}^{|S_i|} \int_0^T K^2(x - S_{ij}; T) dx} && \text{(I.4)} \\
&= 2\sqrt{|S_i|\sum_{j=1}^{|S_i|} \|K(\cdot - S_{ij}; T)\|_2^2} \\
&\leq 2|S_i|\Delta
\end{aligned}
$$

where (I.3) is obtained by Cauchy inequality and $\Delta = \max_{r\in[0,T]} \|K(\cdot - r; T)\|_2$. Since the kernel K is continuous and non-negative on $[0, T]$, Δ must exist. $|S_i|$ is the dimension of S_i and S_{ij} is the j-th event time in S_i.
Therefore, the left hand side has the relation:

$$
\begin{aligned}
\forall i, \sqrt{|t|}\delta > 2|S_i|\Delta &\iff |t| > 4(\frac{\Delta}{\delta}\max_{i=1,2,\cdots,N} |S_i|)^2 \\
&\implies \forall i, d_{K,2}(t, \phi_0) > 2d_{K,2}(S_i, \phi_0)
\end{aligned}
$$

so we have:

$$
|t| > 4(\frac{\Delta}{\delta}\max_{i=1,2,\cdots,N} |S_i|)^2 \implies t \notin \bar{S}_K^{(N)}
$$

The contrapositive statement is thus

$$
t \in \bar{S}_K^{(N)} \implies |t| \leq 4(\frac{\Delta}{\delta}\max_{i=1,2,\cdots,N} |S_i|)^2
$$

which means $\forall \bar{S} \in \bar{S}_K^{(N)}$, $|\bar{S}| \leq \lceil 4(\frac{\Delta}{\delta}\max_{i=1,2,\cdots,N} |S_i|)^2 \rceil = D_K^{(N)}$ where $\lceil x \rceil$ represents the smallest integer that is larger than or equal to x. As $\max_{i=1,2,\cdots,N} |S_i|)^2$, Δ and δ are all finite values, $D_K^{(N)}$ is finite, thus the dimension for any element in $\bar{S}_K^{(N)}$ is bounded. □

Appendix J. An example RJMCMC annealing algorithm

As discussed in section 2.4.2, we can use the RJMCMC method together with simulation annealing method to compute $\arg\min_{t\in\Omega} SSD(t; S_1, S_2, \cdots, S_N)$ and find the empirical Karcher mean $\bar{S}_K^{(N)}$. One way to construct the RJMCMC annealing algorithm, as an example, is provided below.
The pre-determined parameters and functions for RJMCMC method:

1. $J(k'|k)$: Only the adjacent dimensions are possible to jump to.

$$
J(k'|k) = \begin{cases} 1/3 & k = 2, 3, \cdots, D_K^{(N)} - 1 \text{ and } k' = k-1, k \text{ or } k+1 \\ 0 & k = 2, 3, \cdots, D_K^{(N)} - 1 \text{ and } k' \neq k-1, k \text{ or } k+1 \end{cases}
$$

$$
J(k'|1) = \begin{cases} 1/3 & k' = 2 \\ 2/3 & k' = 1 \\ 0 & \text{otherwise} \end{cases} \qquad J(k'|D_K^{(N)}) = \begin{cases} 1/3 & k' = D_K^{(N)} - 1 \\ 2/3 & k' = D_K^{(N)} \\ 0 & \text{otherwise} \end{cases}
$$

2. $h_{k'|k}$: Since the order of the event time has no impact on the smoothing function, we can release the restriction that $x_1 \leq x_2 \leq \cdots \leq x_k$ is not necessary, because the jump is restricted to adjacent dimensions, only $k' = k \pm 1$ needs to be specified, so

   - For $k = 2, 3, \cdots, D_K^{(N)}$, $h_{k-1|k}([x_1, x_2, \cdots, x_k]') = (x_1, [x_2, x_3, \cdots, x_k]')$, which means $(w_{k-1}, y) = h_{k-1|k}(x) \iff w_{k-1} = x_1$ and $y = (x_2, x_3, \cdots, x_k)$. No auxiliary variable is needed for x.
   - For $k = 1, 2, \cdots, D_K^{(N)} - 1$, $h_{k+1|k}(w_k, [x_1, x_2, \cdots, x_k]') = ([w_k, x_1, x_2, \cdots, x_k]')$, which means $(y) = h_{k-1|k}(w_k, x) \iff y = (w_k, x_1, x_2, \cdots, x_k)$. No auxiliary variable is needed for y. Without the restriction on the order, whether w_k is the smallest among $x_1, x_2, \cdots, x_k$ does not matter.

   Under this setting, it is easy to see that the function $h_{k'|k}$ is invertible with $h_{k'|k} = h_{k|k'}^{-1}$ and $h_{k'|k}$ is smooth. The absolute determinant of the Jacobin matrix is

$$|\frac{\partial(w_{k-1}, y)}{\partial(x_1, [x_2, \cdots, x_n]')}| = \begin{vmatrix} 1 & 0 \\ 0 & I_{k-1} \end{vmatrix} = 1 \text{ and } |\frac{\partial(y_1, [y_2, \cdots, y_{k+1}]')}{\partial(w_k, x)}| = \begin{vmatrix} 1 & 0 \\ 0 & I_k \end{vmatrix} = 1$$

3. $g_k(w_k)$: Without the restriction on the order, there is no need to make w_k to be the smallest, so we can simply sample from a Uniform distribution on $[0, T]$ to be w_k, i.e. $g_k(w_k) = \frac{1}{T} I_{[0,T]}(w_k)$.
4. $q_k(y|x)$: The simplest way to do same dimension update is by independent Metropolis Hasting algorithm that the candidate y is sampled directly from k i.i.d. Uniform distribution Unif$[0, T]$, i.e. $q_k(y|x) \doteq q_k(y) = \frac{1}{T^k} \Pi_{i=1}^k I_{[0,T]}(y_i)$.

After we have the pre-determined functions and parameters, the corresponding algorithm is then shown in Algorithm 2.

Appendix K. Proof of Proposition 5

Since we will search every possible dimensions $k = 1, 2, \cdots, D_K^{(N)}$, our goal is then to find the solution for

$$\hat{t}_k = \underset{t \in \Omega_k}{\arg\min} \sum_{i=1}^{N} d_{K,2}^2(t, S_i) = \underset{t \in \Omega_k}{\arg\min} \quad SSD(t; S_1, S_2, \cdots, S_N)$$

where SSD is the "Sum of Squared Distance" - the objective function to optimize and $d_{K,2}(x, y) = \| f_x - f_y \|_2$ is the distance between two event time vectors x, y with $\mathbb{L}^2$ norm.
The objective is in details:

$$\begin{aligned} SSD(t; S_1, S_2, \cdots, S_N) &= \sum_{i=1}^{N} \int_a^b [f_t(x) - f_{S_i}(x)]^2 dx \\ &= \int_a^b \sum_{i=1}^{N} [\sum_{j=1}^{k} K(x - t_j; T) - \sum_{j=1}^{|S_i|} K(x - S_{i,j}; T)]^2 dx \end{aligned}$$

where $|S_i|$ means the dimension of vector S_i and $t = [t_1, t_2, \cdots, t_k]^T, S_i = [S_{i,1}, S_{i,2}, \cdots, S_{i,|S_i|}]^T$. The time interval is set to be $[a, b]$ to be the general case, where $b - a = T$. In the paper for simplicity, we just considered $a = 0$ and $b = T$. The gradient is then:

$$\begin{aligned} &\frac{\partial SSD}{\partial t_r}(t; S_1, S_2, \cdots, S_N) \\ &= \int_a^b \sum_{i=1}^{N} -2K'(x - t_r)[\sum_{j=1}^{k} K(x - t_j; T) - \sum_{j=1}^{|S_i|} K(x - S_{i,j}; T)] dx \\ &= -2 \int_a^b K'(x - t_r)[N \sum_{j=1}^{k} K(x - t_j; T) - \sum_{i=1}^{N} \sum_{j=1}^{|S_i|} K(x - S_{i,j}; T)] dx \end{aligned} \tag{K.1}$$

After vectorization, Eqn. (K.1) gives:

$$\frac{\partial SSD}{\partial t}(t; S_1, S_2, \cdots, S_N) = -2 \int_a^b K'(x - t)[N \vec{1}_k^T K(x - t; T) - \sum_{i=1}^{N} \vec{1}_{|S_i|}^T K(x - S_i; T)] dx$$

Algorithm 2 RJMCMC Annealing Karcher Mean Estimation.

```
Input: The observed event time vectors: $S_1, S_2, \cdots, S_N$; the maximum number of iterations: $n_{max}$; initial value: $x_0$; initial dimension (dimension of $x_0$): $k_0$; the upper bound of dimension to search: $D_K^{(N)}$; The temperature at each step: $\{T_n\}_{n=1}^{n_{max}}$; pre-determined parameters and functions: $J(k'|k)$, $h_{k'|k}$, $g_k(w_k)$ and $q_k(y|x)$ as defined above.
for each $n = 0, 1, 2, \cdots, n_{max} - 1$ do
    Generate the dimension of the candidate $y$: $k' \sim J(k'|k_n)$;
    if $k' = k_n$ then
        • Sample $y \sim$Unif$[0, T]^{k_n}$;
        • Update $x_{n+1} = y$ with probability $\alpha(x_n, y)$ and $x_{n+1} = x_n$ with probability $1 - \alpha(x_n, y)$, where:
              $\alpha(x_n, y) = \min\{1, e^{[SSD(x_n; S_1, S_2, \cdots, S_N) - SSD(y; S_1, S_2, \cdots, S_N)]/T_{n+1}}\}$
        • Update $k_{n+1} = k_n$.
    else if $k' = k_n - 1$ then
        • Compute candidate $y = (x_{n,2}, x_{n,3}, \cdots, x_{n,k_n})$;
        • Update $x_{n+1} = y$ with probability $\alpha(x_n, y)$ and $x_{n+1} = x_n$ with probability $1 - \alpha(x_n, y)$, where:
              $\alpha(x_n, y) = \min\{1, \frac{1}{T} e^{[SSD(x_n; S_1, S_2, \cdots, S_N) - SSD(y; S_1, S_2, \cdots, S_N)]/T_{n+1}}\}$
        • If $x_{n+1}$ is updated to be $y$, then $k_{n+1} = k_n - 1$, otherwise $k_{n+1} = k_n$.
    else
        • Sample $w_{k_n} \sim$Unif$[0, T]$;
        • Compute candidate $y = (w_{k_n}, x_1, x_2, \cdots, x_{k_n})$;
        • Update $x_{n+1} = y$ with probability $\alpha(x_n, y)$ and $x_{n+1} = x_n$ with probability $1 - \alpha(x_n, y)$, where:
              $\alpha(x_n, y) = \min\{1, T e^{[SSD(x_n; S_1, S_2, \cdots, S_N) - SSD(y; S_1, S_2, \cdots, S_N)]/T_{n+1}}\}$
        • If $x_{n+1}$ is updated to be $y$, then $k_{n+1} = k_n + 1$, otherwise $k_{n+1} = k_n$.
    end if
    Record $x_{n+1}$ and $k_{n+1}$.
end for
With the recorded $\{x_n\}_{n=0}^{n_{max}}$ and $\{k_n\}_{n=0}^{n_{max}}$. Compute:
              $\hat{n} = \arg\min_{n=0,1,\cdots,n_{max}} SSD(x_n; S_1, S_2, \cdots, S_N)$
where $\hat{n}$ represents the set of all optimization solutions. Then calculate:
              $\bar{S}_K^{(N)} = \{sort(x_i) \quad | \quad i \in \hat{n}\}$
where $sort(x)$ means sort the vector $x$ to make the elements in an ascending order.
Output: $\bar{S}_K^{(N)}$ is the empirical Karcher mean.
```

where all operations are done element-wise. In this way, the analytical formula of the gradient has been derived.

Now taking into consideration the special case of the modified Gaussian kernel: $K_G(x|T) = c_1 e^{-\frac{c_2}{T^2}x^2}$. The derivative is $K'_G(x; T) = -\frac{2c_1c_2}{T^2} x e^{-\frac{c_2}{T^2}x^2}$. Bring into (K.1),

$$\begin{aligned}
&\frac{\partial SSD}{\partial t_r}(t; S_1, S_2, \cdots, S_N) \\
&= \frac{4c_1c_2}{T^2} \int_a^b (x - t_r) e^{-\frac{c_2}{T^2}(x-t_r)^2} [N \sum_{j=1}^{k} c_1 e^{-\frac{c_2}{T^2}(x-t_j)^2} - \sum_{i=1}^{N} \sum_{j=1}^{|S_i|} c_1 e^{-\frac{c_2}{T^2}(x-S_{i,j})^2}] dx \\
&= \frac{4c_1^2c_2}{T^2} \int_a^b [N \sum_{j=1}^{k} (x - t_r) e^{-\frac{c_2}{T^2}[(x-t_r)^2 + (x-t_j)^2]} - \sum_{i=1}^{N} \sum_{j=1}^{|S_i|} (x - t_r) e^{-\frac{c_2}{T^2}[(x-t_r)^2 + (x-S_{i,j})^2]}] dx \\
&= \frac{4c_1^2c_2}{T^2} [N \sum_{j=1}^{k} \int_a^b (x - t_r) e^{-\frac{c_2}{T^2}[(x-t_r)^2 + (x-t_j)^2]} dx - \sum_{i=1}^{N} \sum_{j=1}^{|S_i|} \int_a^b (x - t_r) e^{-\frac{c_2}{T^2}[(x-t_r)^2 + (x-S_{i,j})^2]} dx] \\
&= \frac{4c_1^2c_2}{T^2} [N \sum_{j=1}^{k} g(t_r, t_j) - \sum_{i=1}^{N} \sum_{j=1}^{|S_i|} g(t_r, S_{i,j})]
\end{aligned}$$

where $g(x, y) = \int_a^b (z-x)e^{-\frac{c_2}{T^2}[(z-x)^2+(z-y)^2]}dz$. To further simplify, replacing z by $z = u + x$, we have:

$$
\begin{aligned}
g(x, y) &= \int_{a-x}^{b-x} ue^{-\frac{c_2}{T^2}[u^2+(u+x-y)^2]}du = \int_{a-x}^{b-x} ue^{-\frac{c_2}{T^2}[2u^2+2(x-y)u+(x-y)^2]}du \\
&= \int_{a-x}^{b-x} ue^{-\frac{2c_2}{T^2}[(u+\frac{x-y}{2})^2+\frac{(x-y)^2}{4}]}du = e^{-\frac{c_2}{2T^2}(x-y)^2}\int_{a-x}^{b-x} ue^{-\frac{2c_2}{T^2}(u+\frac{x-y}{2})^2}du
\end{aligned}
$$

Replacing u by $u = w - \frac{x-y}{2}$, we have:

$$
\begin{aligned}
& g(x, y) \\
&= e^{-\frac{c_2}{2T^2}(x-y)^2}\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} (w - \frac{x-y}{2})e^{-\frac{2c_2}{T^2}w^2}dw \\
&= e^{-\frac{c_2}{2T^2}(x-y)^2}[\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} we^{-\frac{2c_2}{T^2}w^2}dw - \frac{x-y}{2}\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} e^{-\frac{2c_2}{T^2}w^2}dw]
\end{aligned}
\tag{K.2}
$$

Then for the two integrals inside:

$$
\begin{aligned}
\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} we^{-\frac{2c_2}{T^2}w^2}dw &= \frac{1}{2}\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} e^{-\frac{2c_2}{T^2}w^2}d(w^2) = -\frac{T^2}{4c_2}e^{-\frac{2c_2}{T^2}w^2}\Big|_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} \\
&= \frac{T^2}{4c_2}[e^{-\frac{2c_2}{T^2}(a-\frac{x+y}{2})^2} - e^{-\frac{2c_2}{T^2}(b-\frac{x+y}{2})^2}]
\end{aligned}
$$

$$
\begin{aligned}
\int_{a-\frac{x+y}{2}}^{b-\frac{x+y}{2}} e^{-\frac{2c_2}{T^2}w^2}dw &= \frac{T}{2\sqrt{c_2}}\int_{\frac{2\sqrt{c_2}}{T}(a-\frac{x+y}{2})}^{\frac{2\sqrt{c_2}}{T}(b-\frac{x+y}{2})} e^{-\frac{1}{2}u^2}du \\
&= \frac{T\sqrt{\pi}}{\sqrt{2c_2}}\int_{\frac{2\sqrt{c_2}}{T}(a-\frac{x+y}{2})}^{\frac{2\sqrt{c_2}}{T}(b-\frac{x+y}{2})} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du \\
&= \frac{T\sqrt{\pi}}{\sqrt{2c_2}}[\Phi(\frac{2\sqrt{c_2}}{T}(b - \frac{x+y}{2})) - \Phi(\frac{2\sqrt{c_2}}{T}(a - \frac{x+y}{2}))]
\end{aligned}
$$

where $u = \frac{2\sqrt{c_2}}{T}w$ and Φ is the cumulative distribution function for the standard Normal distribution $N(0, 1)$.

Therefore, bring inside (K.2), we have:

$$
\begin{aligned}
g(x, y) = e^{-\frac{c_2}{2T^2}(x-y)^2}\{&\frac{T^2}{4c_2}[e^{-\frac{2c_2}{T^2}(a-\frac{x+y}{2})^2} - e^{-\frac{2c_2}{T^2}(b-\frac{x+y}{2})^2}] \\
&-\sqrt{\frac{\pi}{8c_2}}T(x-y)[\Phi(\frac{2\sqrt{c_2}}{T}(b - \frac{x+y}{2})) - \Phi(\frac{2\sqrt{c_2}}{T}(a - \frac{x+y}{2}))]\}
\end{aligned}
$$

We will use g in element-wise computation, in other words, we have:

- x, y are same-size matrices:

$$
g(\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}, \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \cdots & \cdots & \cdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}) = \begin{bmatrix} g(x_{11}, y_{11}) & \cdots & g(x_{1n}, y_{1n}) \\ \cdots & \cdots & \cdots \\ g(x_{m1}, y_{m1}) & \cdots & g(x_{mn}, y_{mn}) \end{bmatrix}
\tag{K.3}
$$

- x is univariate and y is matrix:

$$g(x, \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \cdots & \cdots & \cdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}) = \begin{bmatrix} g(x, y_{11}) & \cdots & g(x, y_{1n}) \\ \cdots & \cdots & \cdots \\ g(x, y_{m1}) & \cdots & g(x, y_{mn}) \end{bmatrix}$$

- x is matrix and y is univariate:

$$g(\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}, y) = \begin{bmatrix} g(x_{11}, y) & \cdots & g(x_{1n}, y) \\ \cdots & \cdots & \cdots \\ g(x_{m1}, y) & \cdots & g(x_{mn}, y) \end{bmatrix} \tag{K.4}$$

In this way, we can do vectorization on the gradient and get the following:

$$\frac{\partial SSD}{\partial t_r}(t; S_1, S_2, \cdots, S_N) = \frac{4c_1^2 c_2}{T^2}[N\sum_{j=1}^{k} g(t_r, t_j) - \sum_{i=1}^{N}\sum_{j=1}^{|S_i|} g(t_r, S_{i,j})]$$

$$\frac{\partial SSD}{\partial t}(t; S_1, S_2, \cdots, S_N)$$

$$= \frac{4c_1^2 c_2}{T^2}[N\sum_{j=1}^{k} g(t, t_j) - \sum_{i=1}^{N}\sum_{j=1}^{|S_i|} g(t, S_{i,j})] \tag{K.5}$$

$$= \frac{4c_1^2 c_2}{T^2}[Ng(t\vec{1}_k^T, \vec{1}_k t^T)\vec{1}_k - \sum_{i=1}^{N} g(t\vec{1}_k^T, \vec{1}_{|S_i|}S_i^T)\vec{1}_{|S_i|}] \tag{K.6}$$

where Eqn. (K.5) is derived based on element-wise computation on function g, i.e. Eqn. (K.4), and due to Eqn. (K.3), it is further simplified to Eqn. (K.6). In this way, the gradient for the modified Gaussian kernel case has been derived. □

Appendix L. An example line search algorithm

As discussed in section 2.4.2, a gradient-based optimization method can be applied within each dimension and then a comparison across all the output can return the solution of the minimum SSD. A commonly used gradient-based method, stochastic gradient descent (SGD), will iteratively apply gradient descent within a mini-batch of the overall dataset. As an example, one way to construct the line search algorithm through SGD method is provided in Algorithm 3.

Appendix M. Proof of Theorem 4

- For point 1 and 2, the convergence of the minimum and solution to the minimum:

Since the number of events is bounded by $D > 0$, suppose $D_I = \lceil D \rceil$ where $\lceil x \rceil$ means the smallest integer that is larger than or equal to x, then the space of the event time vector is actually: $\Omega = \cup_{i=0}^{D_I}\Omega_i = \cup_{i=1}^{D_I}\{x = (x_1, x_2, \cdots, x_i) \in \mathbb{R}^i \mid 0 \le x_1 \le x_2 \le \cdots \le x_i \le T\} \cup \{\phi_0\}$. We can show that the metric space $(\Omega, d_{K,p})$ is separable and bounded:

– Separable:

For set $\Omega_i^{(q)} = \{x = (x_1, x_2, \cdots, x_i) \in \mathbb{Q}^i \mid 0 \le x_1 \le x_2 \le \cdots \le x_i \le T\}$, with $\mathbb{Q}^i$ representing the space of rational vectors with dimension i, it is a subset of $\Omega_i = \{x = (x_1, x_2, \cdots, x_i) \in \mathbb{R}^i \mid 0 \le x_1 \le x_2 \le \cdots \le x_i \le T\}$ and $\forall x \in \Omega_i$, $\exists y \in \Omega_i^{(q)}$, such that x can be arbitrarily well-approximated by y, since $\mathbb{Q}^i$ is dense in $\mathbb{R}^i$. Thus, $\Omega_i^{(q)}$ is dense in Ω_i. Because $\mathbb{Q}^i$ is countable, as a subset of $\mathbb{Q}^i$, $\Omega_i^{(q)}$ is also countable. Therefore, $\Omega_i^{(q)}$ is dense and countable.

Let $\Omega^{(q)} = \cup_{i=0}^{D_I}\Omega_i^{(q)}$. Because $\Omega_i^{(q)}$ is dense in Ω_i for every i, $\Omega^{(q)}$ is dense in Ω ($\forall x \in \Omega$, $\exists i$ such that $x \in \Omega_i$, so $\exists y \in \Omega_i^{(q)} \subset \Omega^{(q)}$ such that x can be arbitrarily well-approximated by y). Finite union of countable set will also be countable, so $\Omega^{(q)}$ is countable. Therefore, $\Omega^{(q)}$ is dense in Ω and countable. By definition, we have Ω to be separable.

– Bounded:

For any two elements $x, y \in \Omega$, we have: $d_{K,p}(x, y) = \|f_x - f_y\|_p \le \|f_x\|_p + \|f_y\|_p$ where, by definition, f_x, f_y are the smoothing curve: $f_x(t) = \sum_{i=1}^{|x|} K(t - x_i; T)$ and $f_y(t) = \sum_{j=1}^{|y|} K(t - y_i; T)$ with $|x|$ to be the dimension of x.

Let $\Delta_p = \max_{r\in[0,T]} \|K(\cdot - r; T)\|_p$. Since the kernel K is continuous and non-negative on $[0, T]$, Δ_p must exist. Then since the number of events is bounded by D_I, we have $|x|, |y| \le D_I$ and thus

$$\|f_x\|_p = \|\sum_{i=1}^{|x|} K(t - x_i; T)\|_p \le \sum_{i=1}^{|x|} \|K(t - x_i; T)\|_p \le |x|\Delta_p \le D_I\Delta_p$$

Algorithm 3 Line Search (SGD) Karcher Mean Estimation.

```
Input: the observed event time vectors: $S_1, S_2, \cdots, S_N$; the set of dimension to search: $D_{search}$; the initial values for each dimension in $D_{search}$: $t_k^{(init)}$ for $k \in D_{search}$; The batch size: $B$; the learning rate: $r$; the maximum number of epochs: $ep_{max}$; convergence indicator: $\epsilon$.
for each dimension $k \in D_{search}$ do
    Take the initial value: $t_0 = t_k^{(init)}$ and $ep = 1$;
    while $ep \le ep_{max}$ do
            • Take the initial value: $t_{ep,0} = t_{ep-1}$;
            • Shuffle the dataset randomly and separate to $\lceil \frac{N}{B} \rceil$ mini-batches (partitions) with each containing at most $B$ observations, where $\lceil x \rceil$ means the smallest integer that is larger than or equal to $x$;
        for mini-batch $m = 1, 2, \cdots, \lceil \frac{N}{B} \rceil$ do
            – Access the observations in the $m$-th mini-batch: $\{S_{m,j}\}_{j=1}^{l_m}$, where $l_m$ is the number of observations in this mini-batch. We have:
                (*) If $\frac{N}{B} \notin \mathbb{N}$, then $l_m = B$ for $m = 1, 2, \cdots, \lfloor \frac{N}{B} \rfloor$ and $l_{\lceil \frac{N}{B} \rceil} = N - \lfloor \frac{N}{B} \rfloor B$, where $\lfloor x \rfloor$ is the largest integer that is smaller than or equal to $x$.
                (*) If $\frac{N}{B} \in \mathbb{N}$, then $l_m = B$ for any $m = 1, 2, \cdots, \frac{N}{B}$.
            – Compute the gradient for the current mini-batch $\frac{\partial SSD}{\partial t}(t;\ S_{m,1},\ S_{m,2},\ \cdots,\ S_{m,l_m})$ following Proposition 5, where the dimension of $t$ is $k$.
            – Update: $t_{ep,m} = t_{ep,m-1} - r \frac{\partial SSD}{\partial t}(t; S_{m,1}, S_{m,2}, \cdots, S_{m,l_m})$.
            – Correct the result: for any element $t_{ep,m,j}$ in $t_{ep,m}$, where $j = 1, 2, \cdots, k$, $t_{ep,m,j} = \min\{\max\{t_{ep,m,j}, 0\}, T\}$
        end for
            • Record the result: $t_{ep} = t_{ep,\lceil \frac{N}{B} \rceil}$.
        if $|SSD(t_{ep}; S_1, S_2, \cdots, S_N) - SSD(t_{ep-1}; S_1, S_2, \cdots, S_N)| < \epsilon$ then
            Stop and record $\hat{t}_k = t_{ep}$;
        else
            Set $ep = ep + 1$.
        end if
    end while
end for
With the record $\{\hat{t}_k\}_{k \in D_{search}}$, compute:
        $\hat{k} = \arg\min_{k \in D_{search}} SSD(\hat{t}_k; S_1, S_2, \cdots, S_N)$
where $\hat{k}$ represents the set of all optimization results. Then calculate:
        $\bar{S}_K^{(N)} = \{sort(\hat{t}_i) \mid i \in \hat{k}\}$
where $sort(x)$ means sort the vector $x$ to make the elements in an ascending order.
Output: $\bar{S}_K^{(N)}$ is the empirical Karcher mean.
```

Similarly, we have $\|f_y\|_p \le D_I \Delta_p$. Therefore, $d_{K,p}(x, y) \le \|f_x\|_p + \|f_y\|_p \le 2D_I\Delta_p$. As D_I and Δ_p are both fixed and positive, by definition, Ω is bounded.

Since Ω is separable and bounded, we can apply **Theorem 1** in the paper given by Ginestet (2013): $(\Omega, \mathbb{F}, P_S)$ is a probability space; $(\Omega, d_{K,p})$ is a metric space; $S_1, S_2, \cdots, S_N$ are i.i.d. samples in Ω; Then

$$\min_{t \in \Omega} \frac{1}{N} \sum_{i=1}^{N} d_{K,p}^2(t, S_i) \to \min_{t \in \Omega} \mathbb{E}[d_{K,p}^2(t, S)] \quad \text{a.s.} \tag{M.1}$$

and

$$\operatorname*{Lim\,sup}_{N \to \infty}[\arg\min_{t \in \Omega} \frac{1}{N} \sum_{i=1}^{N} d_{K,p}^2(t, S_i)] \subset \arg\min_{t \in \Omega} \mathbb{E}[d_{K,p}^2(t, S)] \tag{M.2}$$

Since $S_K^{(N)} = \arg\min_{t \in \Omega} \frac{1}{N} \sum_{i=1}^{N} d_{K,p}^2(t, S_i)$ and $\mu_K = \arg\min_{t \in \Omega} \mathbb{E}[d_{K,p}^2(t, S)]$, Eqn. (M.1) and (M.2) will be the same as Eqn. (13) and (14), which proves the theorem.

- For point 3, the convergence of the depth value:

Due to Remark 1 which is proved below, $\bar{S}_K^{(N)}$ will be non-empty, so $\hat{s}_c^{(N)}$ exists for any $N \in \mathbb{N}^+$. For the sequence $\{\hat{s}_c^{(N)}\}_{N=1}^{\infty}$, by Theorem 2, all elements have dimension within $[0, D_K^{(N)}]$ where $D_K^{(N)} = \lceil 4(\frac{\Delta}{\delta} \max_{i=1,2,\cdots,N} |S_i|)^2 \rceil$, $\delta = \min_{r \in [0,T]} \|K(\cdot - r; T)\|_2$ and $\Delta = \max_{r \in [0,T]} \|K(\cdot - r; T)\|_2$. According to the theorem statement, the number of events for the point process is upper bounded by a constant $D > 0$. As a result, because $S_1, \cdots, S_N$ are realizations of the point process, $\max_{i=1,2,\cdots,N} |S_i| \le D$, which means $D_K^{(N)} \le D_K^* = \lceil 4(\frac{\Delta}{\delta} D)^2 \rceil$ with D_K^* independent of N. Thus, $\hat{s}_c^{(N)}$ has dimension within $\{0, 1, \cdots, D_K^*\}$ for all N. There exists a subsequence $\{\hat{s}_c^{(N_i)}\}_{i=1}^{\infty}$ such that all terms have the same dimension k.

Because $\hat{s}_c^{(N_i)}$ is in a closed bounded set $\Omega_k = \{x = (x_1, \cdots, x_k)' | 0 \le x_1 \le \cdots \le x_k \le T\}$, there exists a convergent subsequence $\{\hat{s}_c^{(N_{i_j})}\}_{j=1}^{\infty}$, which is also a subsequence of $\{\hat{s}_c^{(N)}\}_{N=1}^{\infty}$. Let $\lim_{j\to\infty} \hat{s}_c^{(N_{i_j})} = s_c^*$, then by Proposition 1, $\lim_{j\to\infty} d_{K,p}(\hat{s}_c^{(N_{i_j})}, s_c^*) = 0$. Because $\hat{s}_c^{(N_{i_j})}$ is an element in $\bar{S}_K^{(N_{i_j})}$, for any $j = 1, 2, \cdots$,

$$0 \le d_{K,p}(s_c^*, \bar{S}_K^{(N_{i_j})}) = \inf_{\bar{S} \in \bar{S}_K^{(N_{i_j})}} d_{K,p}(s_c^*, \bar{S}) \le d_{K,p}(s_c^*, \hat{s}_c^{(N_{i_j})})$$

so by squeeze theorem, we have:

$$0 = \liminf_{j\to\infty} d_{K,p}(s_c^*, \bar{S}_K^{(N_{i_j})}) = \lim_{j\to\infty} d_{K,p}(s_c^*, \hat{s}_c^{(N_{i_j})}) = 0$$

In this way, $s_c^* \in \operatorname{Lim\,sup} \bar{S}_K^{(N)}$ by definition, which implies that $\operatorname{Lim\,sup} \bar{S}_K^{(N)}$ is non-empty. From point 2, we know $\operatorname{Lim\,sup} \bar{S}_K^{(N)} \subset \mu_K$ a.s. and $\mu_K = \{s_c^{(p)}\}$. Therefore, $\operatorname{Lim\,sup} \bar{S}_K^{(N)} = \{s_c^{(p)}\}$ and $s_c^* = s_c^{(p)}$ almost surely. Then we have the dimension $k = \dim(s_c^{(p)})$ a.s., which means, any subsequence of $\{\hat{s}_c^{(N)}\}_{N=1}^{\infty}$ that all terms have the same dimension will have the dimension to be $\dim(s_c^{(p)})$ almost surely. In this way, for any $k \in \{0, 1, \cdots, D_K^*\}$ that $k \ne \dim(s_c^{(p)})$, there are only finitely many elements in $\{\hat{s}_c^{(N)}\}_{N=1}^{\infty}$ that have dimensions to be k with probability one, i.e.

$$\exists N_0 \in \mathbb{N} \text{ such that } \forall N \ge N_0,\ \dim(\hat{s}_c^{(N)}) = \dim(s_c^{(p)}) \text{ a.s.}$$

Considering $\{\hat{s}_c^{(N)}\}_{N=N_0}^{\infty}$, with probability one, all terms have dimension to be $\dim(s^{(p)})$. Suppose the sequence does not converge to $s_c^{(p)}$, then we have

$$\exists e_0 > 0 \text{ such that } \forall N \ge N_0,\ \exists N_* > N \text{ such that } \|\hat{s}_c^{(N_*)} - s_c^{(p)}\| \ge e_0 \text{ a.s.}$$

Iteratively taking $N = N_*$ with initial value $N = N_0$, we can obtain an infinite subsequence $\{\hat{s}_c^{(N_i^*)}\}_{i=1}^{\infty}$ that $\|\hat{s}_c^{(N_i^*)} - s_c^{(p)}\| \ge e_0$ a.s. for all $i = 1, 2, \cdots$. $\{\hat{s}_c^{(N_i^*)}\}_{i=1}^{\infty}$ is in closed bounded set $\Omega_{\dim(s_c^{(p)})}$, so similar to $\hat{s}_c^{(N_i)}$ that has been discussed above, there exists a subsequence $\{\hat{s}_c^{(N_{i_j}^*)}\}_{j=1}^{\infty}$ that converges. Denote the limiting vector to be $\lim_{j\to\infty} \hat{s}_c^{(N_{i_j}^*)} = s_c^{**}$, we have

$$\begin{aligned} & 0 \le d_{K,p}(s_c^{**}, \bar{S}_K^{(N_{i_j}^*)}) = \inf_{\bar{S} \in \bar{S}_K^{(N_{i_j}^*)}} d_{K,p}(s_c^{**}, \bar{S}) \le d_{K,p}(s_c^{**}, \hat{s}_c^{(N_{i_j}^*)}) \\ \Longrightarrow\ & 0 \le \liminf_{j\to\infty} d_{K,p}(s_c^{**}, \bar{S}_K^{(N_{i_j}^*)}) \le \lim_{j\to\infty} d_{K,p}(s_c^{**}, \hat{s}_c^{(N_{i_j}^*)}) = 0 \\ \Longrightarrow\ & s_c^{**} \in \operatorname{Lim\,sup} \bar{S}_K^{(N)} \end{aligned}$$

Thus $s_c^{**} = s_c^{(p)}$ a.s. However, $\lim_{j\to\infty} \hat{s}_c^{(N_{i_j}^*)} = s_c^{**}$ implies that $\forall \epsilon > 0$, $\exists j_0 > 0$ such that $\forall j > j_0$, $\|\hat{s}_c^{(N_{i_j}^*)} - s_c^{**}\| < \epsilon$. Taking $\epsilon = e_0/2$, $\forall j > j_0$, $\|\hat{s}_c^{(N_{i_j}^*)} - s_c^{**}\| < e_0/2$. As a subsequence, we have $\|\hat{s}_c^{(N_{i_j}^*)} - s_c^{(p)}\| \ge e_0, \forall j > 0$ a.s. Then by triangle inequality,

$$\|s_c^{**} - s_c^{(p)}\| \ge \|\hat{s}_c^{(N_{i_j}^*)} - s_c^{(p)}\| - \|\hat{s}_c^{(N_{i_j}^*)} - s_c^{**}\| > e_0/2 > 0 \text{ a.s.}$$

which contradicts to $s_c^{**} = s_c^{(p)}$ a.s. Therefore, the original statement is false, the sequence $\hat{s}_c^{(N)}$ converges to $s_c^{(p)}$ a.s., i.e. $\dim(\hat{s}_c^{(N)}) \to \dim(s_c^{(p)})$ and $\hat{s}_{c_i}^{(N)} \to s_{c_i}^{(p)}$ almost surely for $\forall i = 1, 2, \cdots, \dim(s_c^{(p)})$, as $N \to \infty$.

For the smoothed process, the kernel function K is continuous, so $f_s(t) = \sum_{i=1}^n K(t - s_i; T)$ is continuous with respect to both s and t. In this way, $\lim_{N\to\infty} f_{\hat{s}_c^{(N)}}(t) = f_{s_c^{(p)}}(t)$ a.s. for any $t \in [0, T]$ and $f_s(t)$ can attain its maximum and minimum on $[0, T]$. Let $A = \max_{w\in[0,T]} [f_s(w) - f_{\hat{s}_c^{(N)}}(w)]^2$, then for $\forall t \in [0, T]$, $[f_s(t) - f_{\hat{s}_c^{(N)}}(t)]^2 \le A < \infty$, $[f_s(t) - f_{\hat{s}_c^{(N)}}(t)]^2 \to [f_s(t) - f_{s_c^{(p)}}(t)]^2$ a.s. as $N \to \infty$ and $\int_0^T A dw = AT < \infty$, so by dominant convergence theorem,

$$\lim_{N\to\infty} \|f_s - f_{\hat{s}_c^{(N)}}\|^2 = \lim_{N\to\infty} \int_0^T [f_s(t) - f_{\hat{s}_c^{(N)}}(t)]^2 dt$$

$$\begin{aligned}&= \int_0^T [f_s(t) - f_{s_c^{(p)}}(t)]^2 dt \text{ a.s.}\\ &= \| f_s - f_{s_c^{(p)}} \|^2 \text{ a.s.}\end{aligned}$$

Therefore, as exponential function is continuous, we have:

$$\begin{aligned}\lim_{N\to\infty} D(s;\hat{s}_c^{(N)}) &= \lim_{N\to\infty} \exp[\frac{1}{2h}\| f_s - f_{\hat{s}_c^{(N)}}\|^2]\\ &= \exp[\frac{1}{2h}\lim_{N\to\infty} \| f_s - f_{\hat{s}_c^{(N)}}\|^2]\\ &= \exp[\frac{1}{2h}\| f_s - f_{s_c^{(p)}}\|^2] \text{ a.s.}\\ &= D(s;s_c^{(p)}) \text{ a.s.}\end{aligned}$$

- For Remark 1, the existence of the minimum and thus the solution to the minimum:

To show that the minimum can be achieved, we will separate the discussion to the empirical version and the population version.

- Empirical version: $\min_{t\in\Omega} \frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$.

According to Proposition 1, we know that for any $i \in \mathbb{N}^+$ and any $y \in \Omega_i$, if $x \in \Omega_i$ converges to y, then for any $z \in \Omega$, $d_{K,p}(x,z)$ will converge to $d_{K,p}(y,z)$, i.e.

$$\forall z \in \Omega, \quad \lim_{x\to y \text{ in } \Omega_i} d_{K,p}(x,z) = d_{K,p}(y,z)$$

Thus, $d_{K,p}(\cdot,z)$ for any $z \in \Omega$ is continuous. As quadratic function and sum function are both continuous, we have: $\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$ is continuous for $t \in \Omega_i$, where $i \in \mathbb{N}^+$.

As shown in the above discussion for Ω to be bounded, we have: $0 \le d_{K,p}(x,y) \le 2D_I\Delta_p$ for any $x,y \in \Omega$. Thus, $\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i) \in [0, 4D_I^2\Delta_p^2]$ is bounded for $t \in \Omega_i$, where $i \in \mathbb{N}^+$.

Therefore, for any integer $i > 0$, $\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$ is a continuous and bounded function for $t \in \Omega_i$. Because $\Omega_i = \{x = (x_1,x_2,\cdots,x_i) \in \mathbb{R}^i \mid 0 \le x_1 \le x_2 \le \cdots \le x_i \le T\}$ is a closed and bounded set, $\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$ must achieve its minimum and maximum in Ω_i. Thus $\min_{t\in\Omega_i} \frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$ can be reached. In this way, since the number of events is bounded by $D < D_I \in \mathbb{N}^+$, we have:

$$\min_{i=1,2,\cdots,D_I}\min_{t\in\Omega_i}\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i) = \min_{t\in\Omega}\frac{1}{N}\sum_{i=1}^N d_{K,p}^2(t,S_i)$$

can be achieved. Therefore, the minimum of the empirical version can be achieved and the empirical Karcher mean exists.

- Population version: $\min_{t\in\Omega} \mathbb{E}[d_{K,p}^2(t,S)]$.

Suppose x goes to y in Ω_i for any $i \in \mathbb{N}^+$, we have:

$$\begin{aligned}0 &\le |\mathbb{E}[d_{K,p}^2(x,S)] - \mathbb{E}[d_{K,p}^2(y,S)]|\\ &= |\mathbb{E}[d_{K,p}^2(x,S) - d_{K,p}^2(y,S)]|\\ &\le \mathbb{E}|d_{K,p}^2(x,S) - d_{K,p}^2(y,S)|\\ &= \mathbb{E}[|d_{K,p}(x,S) + d_{K,p}(y,S)||d_{K,p}(x,S) - d_{K,p}(y,S)|]\end{aligned}$$

Since $0 \le d_{K,p}(x,y) \le 2D_I\Delta_p$ for any $x,y \in \Omega$, $|d_{K,p}(x,S) + d_{K,p}(y,S)| = d_{K,p}(x,S) + d_{K,p}(y,S) \le 4D_I\Delta_p$. By triangle inequality, $|d_{K,p}(x,S) - d_{K,p}(y,S)| \le d_{K,p}(x,y)$. So,

$$\begin{aligned}&\mathbb{E}[|d_{K,p}(x,S) + d_{K,p}(y,S)||d_{K,p}(x,S) - d_{K,p}(y,S)|]\\ &\le \mathbb{E}[4D_I\Delta_p d_{K,p}(x,y)] \le 4D_I\Delta_p d_{K,p}(x,y)\end{aligned}$$

which gives:

$$0 \le |\mathbb{E}[d_{K,p}^2(x,S)] - \mathbb{E}[d_{K,p}^2(y,S)]| \le 4D_I\Delta_p d_{K,p}(x,y)$$

According to Proposition 1, as x goes to y in Ω_i, we have $\lim_{x\to y \text{ in } \Omega_i} d_{K,p}(x,y) = 0$. Therefore, by squeeze theorem, $|\mathbb{E}[d_{K,p}^2(x,S)] - \mathbb{E}[d_{K,p}^2(y,S)]|$ will go to 0 as x goes to y, i.e.

$$\lim_{x\to y \text{ in } \Omega_i} |\mathbb{E}[d^2_{K,p}(x,S)] - \mathbb{E}[d^2_{K,p}(y,S)]| = 0$$

which means $\mathbb{E}[d^2_{K,p}(\cdot,S)]$ is continuous in Ω_i with $\forall i \in \mathbb{N}^+$.

Moreover, since $0 \le d_{K,p}(x,y) \le 2D_I\Delta_p$ for any $x, y \in \Omega$, we have $0 \le \mathbb{E}[d^2_{K,p}(t,S)] \le \mathbb{E}[4D_I^2\Delta_p^2] = 4D_I^2\Delta_p^2$, so $\mathbb{E}[d^2_{K,p}(\cdot,S)]$ is bounded in Ω_i with $\forall i \in \mathbb{N}^+$.

Therefore, for any integer $i > 0$, $\mathbb{E}[d^2_{K,p}(t,S)]$ is a continuous and bounded function for $t \in \Omega_i$. Because $\Omega_i = \{x = (x_1, x_2, \cdots, x_i) \in \mathbb{R}^i \mid 0 \le x_1 \le x_2 \le \cdots \le x_i \le T\}$ is a closed and bounded set, $\mathbb{E}[d^2_{K,p}(t,S)]$ must achieve its minimum and maximum in Ω_i. Thus $\min_{t\in\Omega_i} \mathbb{E}[d^2_{K,p}(t,S)]$ can be reached. In this way, since the number of events is bounded by $D < D_I \in \mathbb{N}^+$, we have:

$$\min_{i=1,2,\cdots,D_I} \min_{t\in\Omega_i} \mathbb{E}[d^2_{K,p}(t,S)] = \min_{t\in\Omega} \mathbb{E}[d^2_{K,p}(t,S)]$$

can be achieved. Therefore, the minimum of the population version can be achieved and the Karcher mean on population exists.

- For Remark 2, the feasibility of replacing $d^2_{K,p}$ by $d^r_{K,p}$:

For point 1 and 2, since the bounded and separable conditions for Ω are not related to the power of $d_{K,p}$, when we change the power from 2 to r, they will not be affected, thus the condition for **Theorem 1** in the paper given by Ginestet (2013) will still be feasible. According to the original theorem (Ginestet, 2013), the power is not limited to 2 but can be any r in $[1, \infty)$. In this way, the proof of point 1 and 2 will be totally the same and the conclusion holds.
For point 3, the proof only replies on point 2 that $\mathrm{Lim\,sup}\, \bar{S}_K^{(N)} \subset \mu_K$ almost surely. As point 2 holds after replacing the power of distance from 2 to r, point 3 will also hold.
For Remark 1:

- Empirical version:

As power function $f(x) = x^n$ is continuous with $n \ge 1$, the continuity of $\frac{1}{N}\sum_{i=1}^N d^r_{K,p}(t, S_i)$ for $t \in \Omega_i$, where $i \in \mathbb{N}^+$, can be shown in the same way. Based on $0 \le d_{K,p}(x,y) \le 2D_I\Delta_p$ for any $x, y \in \Omega$, we have $\frac{1}{N}\sum_{i=1}^N d^r_{K,p}(t,S_i) \in [0, 2^r pD_I^r\Delta_p^r]$ is bounded for $t \in \Omega_i$, where $i \in \mathbb{N}^+$. Thus, the empirical version $\frac{1}{N}\sum_{i=1}^N d^r_{K,p}(t,S_i)$ is still continuous and bounded for $t \in \Omega_i$. Then following the same statements, we can prove that the minimum of the empirical version can be achieved and the empirical Karcher mean exists.

- Population version:

Suppose x goes to y in Ω_i for any $i \in \mathbb{N}^+$, similarly we have:

$$\begin{aligned}
&|\mathbb{E}[d^r_{K,p}(x,S)] - \mathbb{E}[d^r_{K,p}(y,S)]| \\
&\le \mathbb{E}|d^r_{K,p}(x,S) - d^r_{K,p}(y,S)| \\
&= \mathbb{E}[|d_{K,p}(x,S) - d_{K,p}(y,S)|\,|\sum_{l=0}^{r} d^{r-l}_{K,p}(x,S) d^l_{K,p}(y,S)|]
\end{aligned}$$

By triangle inequality, $|d_{K,p}(x,S) - d_{K,p}(y,S)| \le d_{K,p}(x,y)$. In addition,

$$\begin{aligned}
|\sum_{l=0}^{r} d^{r-l}_{K,p}(x,S) d^l_{K,p}(y,S)| &= \sum_{l=0}^{r} d^{r-l}_{K,p}(x,S) d^l_{K,p}(y,S) \\
&\le \sum_{l=0}^{r} \max\{d_{K,p}(x,S), d_{K,p}(y,S)\}^r \\
&= (r+1)\max\{d_{K,p}(x,S), d_{K,p}(y,S)\}^r \\
&\le (r+1)2^r D_I^r \Delta_p^r
\end{aligned}$$

because $0 \le d_{K,p}(x,y) \le 2D_I\Delta_p$ for any $x, y \in \Omega$. Thus,

$$\begin{aligned}
0 &\le |\mathbb{E}[d^r_{K,p}(x,S)] - \mathbb{E}[d^r_{K,p}(y,S)]| \\
&\le \mathbb{E}|d_{K,p}(x,S) - d_{K,p}(y,S)|\,|\sum_{l=0}^{r} d^{r-l}_{K,p}(x,S) d^l_{K,p}(y,S)| \\
&\le \mathbb{E}[d_{K,p}(x,y)(r+1)2^r D_I^r \Delta_p^r] \\
&= (r+1)2^r D_I^r \Delta_p^r d_{K,p}(x,y)
\end{aligned}$$

Then similarly, as x goes to y in Ω_i, we have $\lim_{x \to y \text{ in } \Omega_i} d_{K,p}(x, y) = 0$. Therefore, by squeeze theorem, $|\mathbb{E}[d^r_{K,p}(x, S)] - \mathbb{E}[d^r_{K,p}(y, S)]|$ will go to 0 as x goes to y, which means $\mathbb{E}[d^r_{K,p}(\cdot, S)]$ is continuous in Ω_i with $\forall i \in \mathbb{N}^+$.

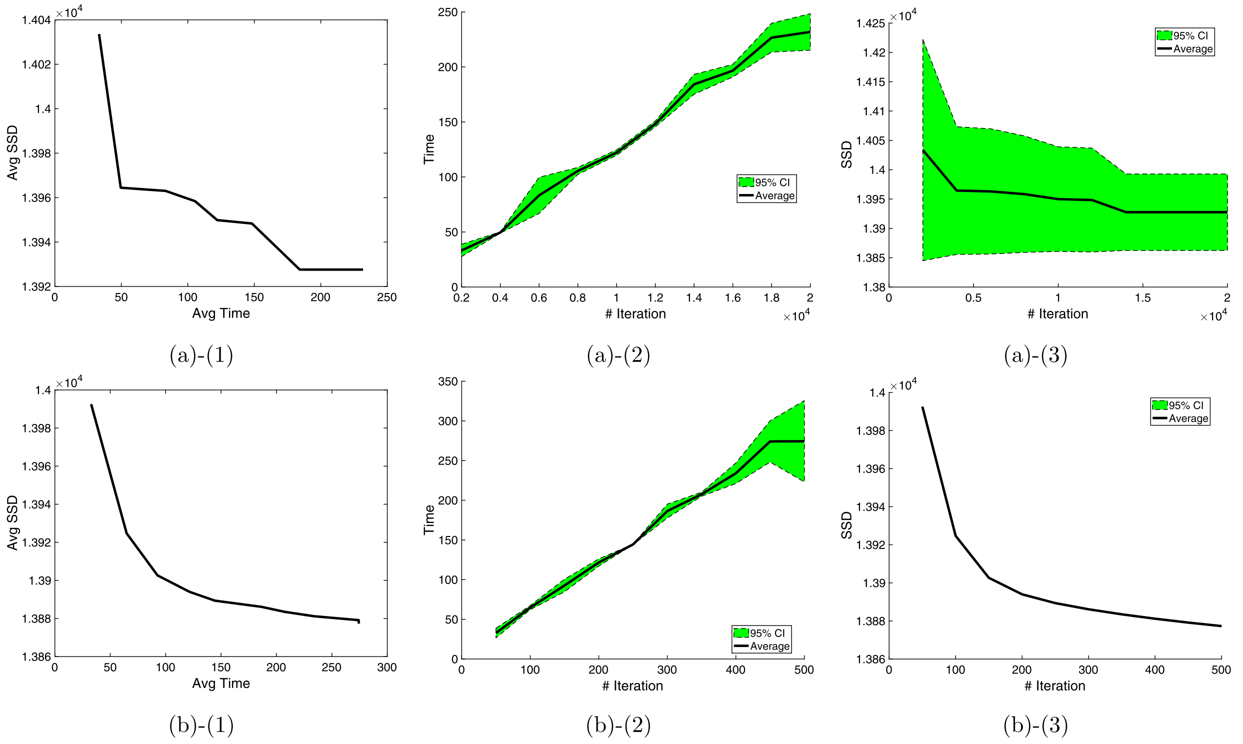
To show $\mathbb{E}[d^r_{K,p}(\cdot, S)]$ is bounded in Ω_i, the idea will be the same that as $0 \le d_{K,p}(x, y) \le 2D_I \Delta_p$, we have $0 \le \mathbb{E}[d^r_{K,p}(t, S)] \le 2^r D^r_I \Delta^r_p$. Thus, the population version $\mathbb{E}[d^r_{K,p}(t, S)]$ is still continuous and bounded for $t \in \Omega_i$. Then following the same statements, we can prove that the minimum of the population version can be achieved and the Karcher mean on population exists.

In this way, we can see that if we replace $d^2_{K,p}$ by $d^r_{K,p}$, 1, 2 and Remark 1 will not be influenced and will still be true. Therefore, the power can be generalized to any number $r \ge 1$. □

Appendix N. More analysis for modified *h*-depth in Section 4.1.1

To compare the 3 center estimation methods more comprehensively, we have done some further analysis through repeated experiments using different hyper-parameters on this simulated HPP dataset. Because each of the methods have the hyper-parameters that can control the number of iterations and the random seed, the center estimation results through changing these settings can reflect the relation between the time running the algorithm and the SSD of the result. For the RJMCMC annealing method, we adjust the maximum number of iterations and the random seed. For the line search method, we adjust the maximum number of epochs and the random seed. For the combined method, we adjust the maximum number of iterations in RJMCMC annealing part, the maximum number of epochs in line search part and the random seed.

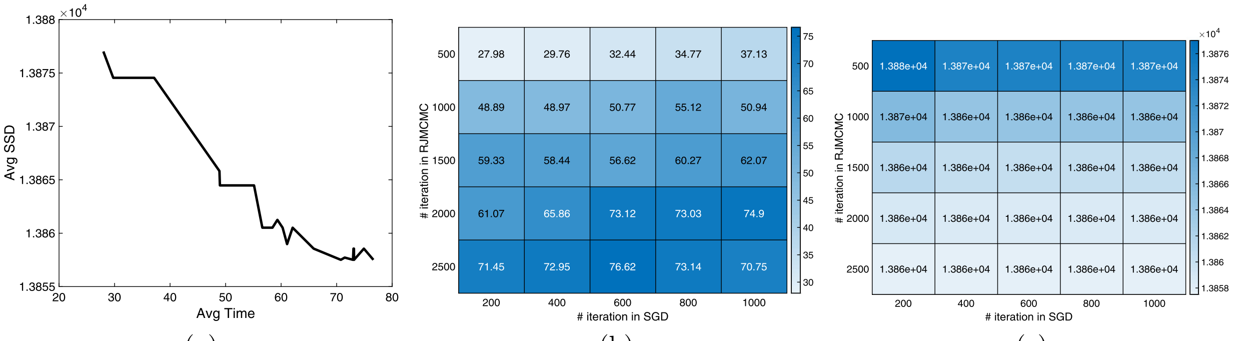
The results from the RJMCMC annealing method and line search method are given in Fig. N.9 (a) and (b) respectively, and the results from the combined method are shown in Fig. N.10. Note that the 95% confidence interval for SSD versus number of iteration is zero-length for the line search method, because the results coming from the iterative gradient-based optimization are very stable. As can be seen from the figures, for all the methods, the SSD decreases and the time cost increases, as the number of iterations increases. Based on the average SSD versus average time cost plots, all the methods have the SSD nearly converged (the curve becomes flat), when the time cost reaches the maximum. In this way, we can treat the result with the largest number of iterations to be the final output of the corresponding method, which provides



(a)-(1) (a)-(2) (a)-(3)

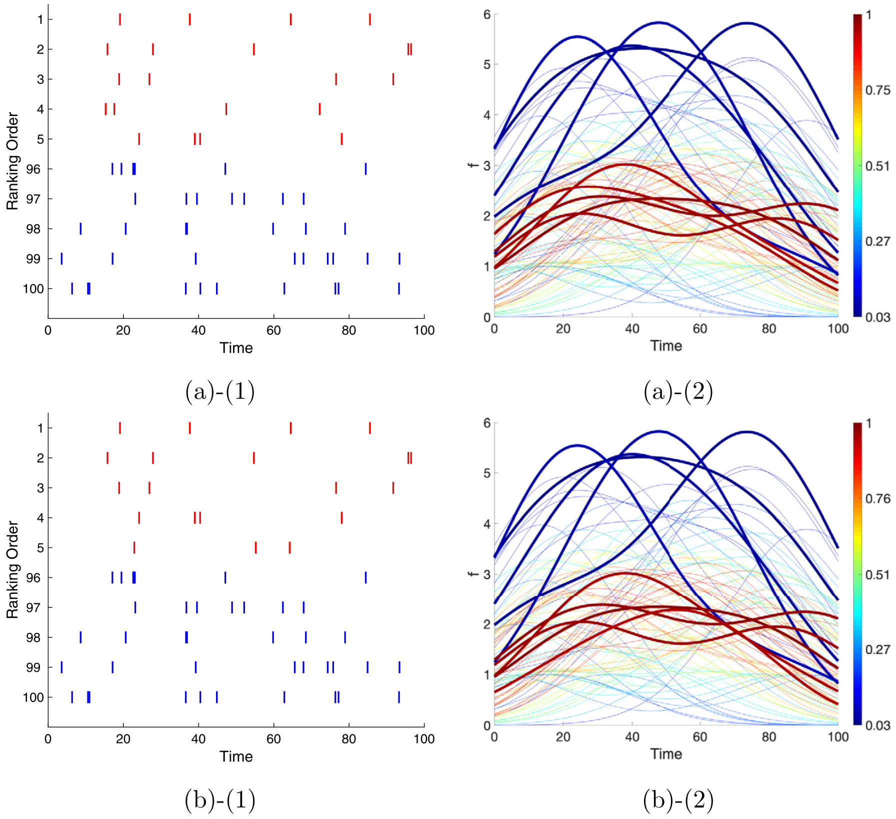
(b)-(1) (b)-(2) (b)-(3)

Fig. N.9. The repeated experiment results through adjusting number of iterations and random seed: RJMCMC annealing method and line search method. Row (a): RJMCMC annealing method. Row (b): Line search method. Column (1): The average SSD versus average time cost over different seeds. Column (2): The time cost versus number of iterations, average and 95% confidence interval. Column (3): The SSD versus number of iterations, average and 95% confidence interval.



(a) (b) (c)

Fig. N.10. The repeated experiment results through adjusting number of iterations and random seed: combined method. (a): The average SSD versus average time cost over different seeds. (b): The time cost heat-map, the darker the color, the larger the time cost. (c): The SSD heat-map, the darker the color, the larger the SSD.



(a)-(1) (a)-(2)

(b)-(1) (b)-(2)

Fig. N.11. The top 5 and bottom 5 plots and the color-mapped smoothed processes plot, using the estimated center from RJMCMC annealing method and line search method. Row (a): RJMCMC annealing method. Row (b): Line search method. Column (1): Top 5 (red) and bottom 5 (blue) processes ranked by the depth values in the HPP simulation. Column (2): Color-mapped smoothed processes based on depth values for the HPP simulation, where top 5 (red) and bottom 5 (blue) are marked with thick lines.

the numbers in Table 1. Therefore, the combined method has the converged output with the lowest SSD, and also it requires the smallest time cost to obtain the converged result.

In addition, the RJMCMC annealing method and line search method have the SSD reduced from around 14000 to around 13880 when increasing the number of iterations. In contrast, the combined method has the SSD changing from around 13880 to around 13860. The beginning SSD for the combined method is close to the ending SSD for the RJMCMC annealing method and line search method. This indicates that the combined method can output a result with a low SSD close to the converged value, even when the number of iterations is not large, which means the combined method is more efficient compared to the RJMCMC annealing method and the line search method.

In this analysis, the number of iterations is set to what can nearly achieve the convergence in SSD. We can observe that the optimal SSD for the 3 methods are close to each other, and the time cost for the RJMCMC annealing method and the line search method is much larger than the combined method. In principle, the 3 center estimation methods all should converge to the true solution of the minimum SSD. If a large enough number of iterations is provided, all the 3 methods should have the same results with the same SSD, but from this repeated experiment, we should expect that the time cost of the combined method will be the lowest.

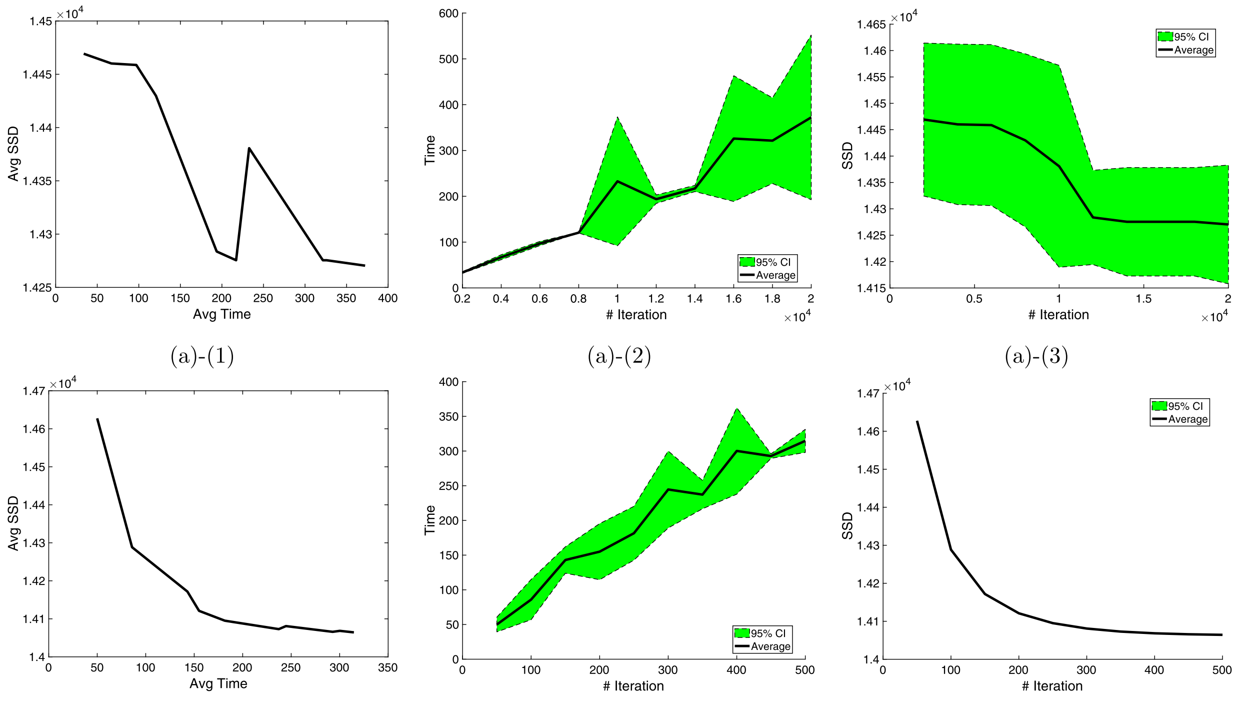
Similar to Fig. 2 and Fig. 3, we have plotted the top 5 and bottom 5 processes according to the depth value in Fig. N.11 column (1), and the color-mapped smoothed processes based on the depth value in Fig. N.11 column (2), for the RJMCMC annealing method and line search method. The results are very similar to the ones from the combined method, as the estimated centers are very similar.

Appendix O. More analysis for modified *h*-depth in Section 4.1.2

Similar to section Appendix N, to compare the 3 center estimation methods more comprehensively, we have done some further analysis through repeated experiments using different hyper-parameters on this simulated IPP dataset. We change the number of iterations and random seeds to learn the relation between the time running the algorithm and the SSD of the result.

The results from the RJMCMC annealing method and line search method are given in Fig. O.12 (a) and (b) respectively, and the results from the combined method are shown in Fig. O.13. For RJMCMC annealing method and combined method, there is a jump in the average SSD when the average time cost increases, which is caused by the unstable output of the RJMCMC annealing method. When the number of iterations is not large, the output of RJMCMC annealing contains both dimension-5 and dimension-6 centers given different random seeds. From the final converges output, we know that dimension-5 center has better SSD, which explains the jump as the output is not purely dimension-5. In addition, the 95% confidence interval for SSD versus number of iteration is zero-length for the line search method, because the results coming from the iterative gradient-based optimization are very stable.

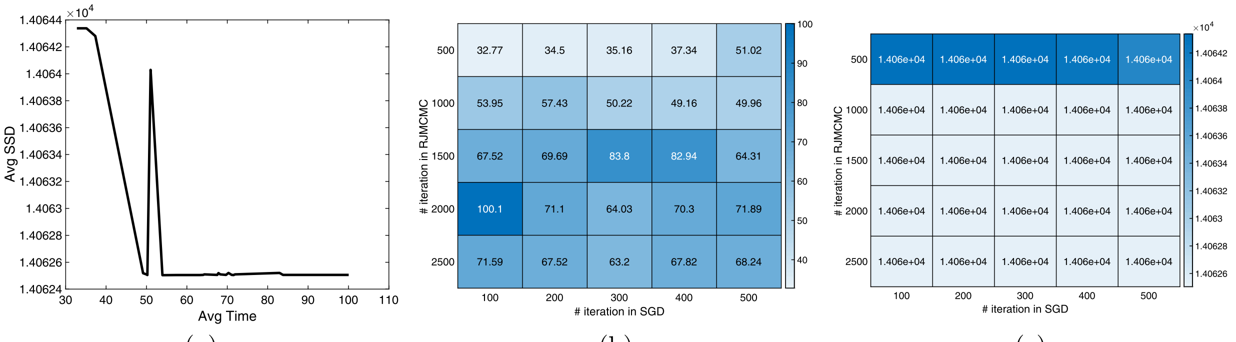
According to the figures, similar to the HPP simulation, all the methods have the SSD nearly converged when the time cost reaches the maximum, so the result with the largest number of iterations can be treated as the final output of the corresponding method, which provides the numbers in Table 2. Therefore, the combined method has the converged output with the lowest SSD, and also it requires the smallest time cost to obtain the converged result.



(a)-(1) (a)-(2) (a)-(3)

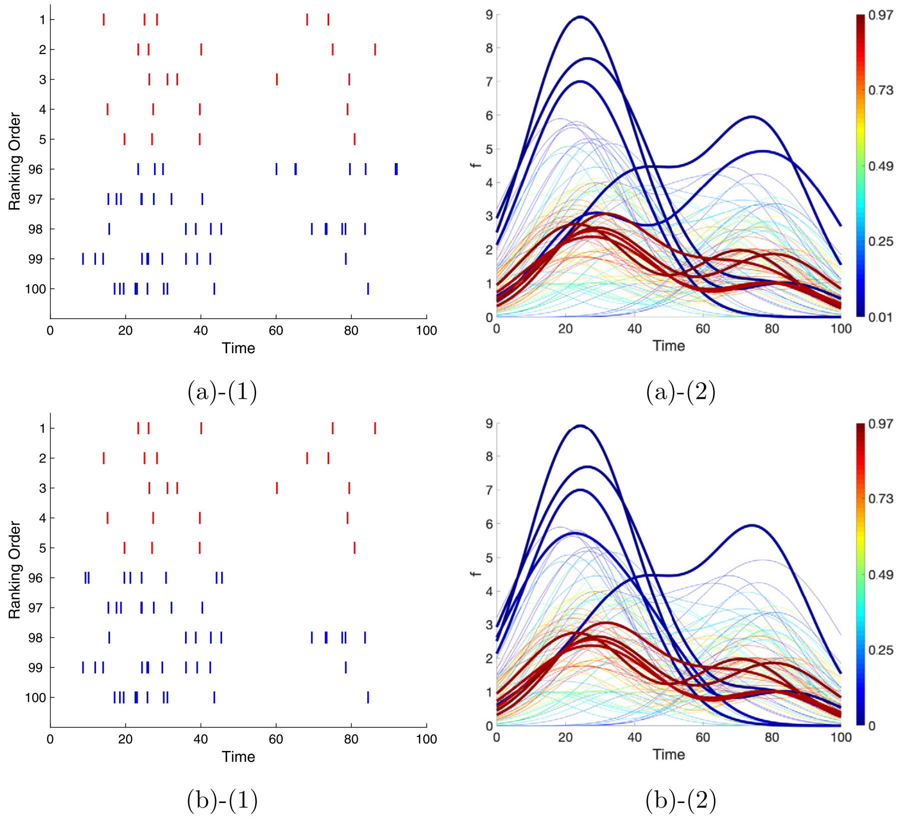
(b)-(1) (b)-(2) (b)-(3)

Fig. O.12. The repeated experiment results through adjusting number of iterations and random seed: RJMCMC annealing method and line search method. Row (a): RJMCMC annealing method. Row (b): Line search method. Column (1): The average SSD versus average time cost over different seeds. Column (2): The time cost versus number of iterations, average and 95% confidence interval. Column (3): The SSD versus number of iterations, average and 95% confidence interval.



(a) (b) (c)

Fig. O.13. The repeated experiment results through adjusting number of iterations and random seed: combined method. (a): The average SSD versus average time cost over different seeds. (b): The time cost heat-map, the darker the color, the larger the time cost. (c): The SSD heat-map, the darker the color, the larger the SSD.



(a)-(1) (a)-(2)

(b)-(1) (b)-(2)

Fig. O.14. The top 5 and bottom 5 plots and the color-mapped smoothed processes plot, using the estimated center from RJMCMC annealing method and line search method. Row (a): RJMCMC annealing method. Row (b): Line search method. Column (1): Top 5 (red) and bottom 5 (blue) processes ranked by the depth values in the HPP simulation. Column (2): Color-mapped smoothed processes based on depth values for the HPP simulation, where top 5 (red) and bottom 5 (blue) are marked with thick lines.

In addition, similar to the HPP simulation, if we consider the average SSD change as the average time cost increases, the RJMCMC annealing drops from 14500 to 14250 approximately, and line search drops from 14700 to 14000 approximately. Both have a large decrease in the average SSD. In contrast, the combined method has the SSD changing from around 14064 to around 14062, which is nearly unchanged compared to the other two methods. Therefore, we can see that the combined method can output a result with a low SSD even when the number of iterations is not large, which gives the same conclusion as the HPP case that the combined method is more efficient compared to the RJMCMC annealing method and the line search method.

Same as the HPP simulation, the number of iterations is set to what can nearly achieve the convergence in SSD. We can also observe that the optimal SSD for the 3 methods are close to each other, and the time cost for the RJMCMC annealing method and the line search method is much larger than the combined method. Theoretically, the 3 center estimation methods all should converge to the true solution of the minimum SSD. If a large enough number of iterations is provided, all the 3 methods should have the same results with the same SSD, but from this repeated experiment, we should expect that the time cost of the combined method will be the lowest.

Similar to Fig. 5 and Fig. 6, we have plotted the top 5 and bottom 5 processes according to the depth value in Fig. O.14 column (1), and the color-mapped smoothed processes based on the depth value in Fig. O.14 column (2), for the RJMCMC annealing method and line search method. The results are very similar to the ones from the combined method, as the estimated centers are very similar.

References

Barnett, V., 1976. The ordering of multivariate data. J. R. Stat. Soc. A 139, 318–344.

Carandini, M., Horton, J.C., Sincich, L.C., 2007. Thalamic filtering of retinal spike trains by postsynaptic summation. J. Vis. 7, 20.

Cuesta-Albertos, J.A., Nieto-Reyes, A., 2008. The random Tukey depth. Comput. Stat. Data Anal. 52, 4979–4988.

Cuevas, A., Febrero, M., Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. Comput. Stat. 22, 481–496.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6, 721–741.

Gijbels, I., Nagy, S., 2017. On a general definition of depth for functional data. Stat. Sci. 32, 630–639.

Ginestet, C.E., 2013. Strong consistency of set-valued Fréchet sample means in metric spaces. Preprint.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Grove, K., Karcher, H., 1973. How to conjugatec 1-close group actions. Math. Z. 132, 11–20.

Kuratowski, K., 2014. Topology, vol. 1. Elsevier.

Liu, R.Y., Singh, K., 1993. A quality index based on data depth and multivariate rank tests. J. Am. Stat. Assoc. 88, 252–260.

Liu, R.Y., et al., 1990. On a notion of data depth based on random simplices. Ann. Stat. 18, 405–414.

Liu, S., Wu, W., et al., 2017. Generalized Mahalanobis depth in point process and its application in neural coding. Ann. Appl. Stat. 11, 992–1010.

López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. J. Am. Stat. Assoc. 104, 718–734.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Naud, R., Berger, T., Bathellier, B., Carandini, M., Gerstner, W., 2009. Quantitative single-neuron modeling: competition 2009. In: Front. Neur. Conference Abstract: Neuroinformatics 2009, pp. 1–8.

Nieto-Reyes, A., Battey, H., et al., 2016. A topologically valid definition of depth for functional data. Stat. Sci. 31, 61–79.

Oja, H., 1983. Descriptive statistics for multivariate distributions. Stat. Probab. Lett. 1, 327–332.

Qi, K., Chen, Y., Wu, W., 2021. Dirichlet depths for point process. Electron. J. Stat. 15, 3574–3610.

van Rijsbergen, C., 1979. Information Retrieval, 2nd edn. Butterworths.

Sincich, L.C., Adams, D.L., Economides, J.R., Horton, J.C., 2007. Transmission of spike trains at the retinogeniculate synapse. J. Neurosci. 27, 2683–2692.

Tukey, J.W., 1975. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, vol. 2. Vancouver, 1975, pp. 523–531.

Van Laarhoven, P.J., Aarts, E.H., 1987. Simulated annealing. In: Simulated Annealing: Theory and Applications. Springer, pp. 7–15.

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. CRC Press.

Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. Ann. Stat., 461–482.