Strictly positive definite spike train kernels for point process divergences

Il Memming Park[1], **Sohan Seth**[2], **Murali Rao**[3], **José C. Príncipe**[1, 2]
[1]Department of Biomedical Engineering, University of Florida, Gainesville, FL.
[2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL.
[3]Department of Mathematics, University of Florida, Gainesville, FL.

Keywords: Point process, spike trains, divergence, strictly positive definite kernel, dissimilarity, hypothesis testing

Abstract

Exploratory tools that are sensitive to arbitrary statistical variations in spike train observations open up the possibility of novel neuroscientific discoveries. Developing such tools, however, is difficult due to the lack of Euclidean structure of the spike train space, and an experimenter usually prefers simpler tools that only capture limited statistical features of the spike train such as mean spike count or mean firing rate. We explore strictly positive definite kernels on the space of spike trains to offer both a structural representation of this space and a platform for developing statistical measures that explore features beyond count or rate. We apply these kernels to construct measures of divergence between two point processes, and use them for hypothesis testing, that is, to observe if two sets of spike trains originate from the same underlying probability law. Although there exist positive definite spike train kernels in the literature, we establish that these kernels are not strictly definite, and thus, do not induce measures of divergence. We discuss the properties of both these existing non-strict kernels and the novel strict kernels in terms of their computational complexity, choice of free parameters, and performance on both synthetic and real data through kernel principal component analysis and hypothesis testing.

1 Introduction

Detecting changes in spike train observations is fundamental to the study of neuroscience at both neuronal and systems levels. For example, when studying neural coding, it is critical to ensure that the target neuron is encoding the stimulus (or behavior) of interest, which may be reflected in arbitrary statistical changes in its spiking pattern. This requires deciding if two

sets of spike trains originate from the same probability law (point process). However, this seemingly benign task is technically daunting since the space of spike trains lacks Euclidean structure, making direct application of standard tests for equality of distribution such as χ^2-test, Kolmogorov-Smirnov test or Kullback-Leibler divergence unusable [Park and Príncipe, 2010, Seth et al., 2010a, Naud et al., 2011]. Therefore, most analyses on spike trains are done by observing changes in simpler statistics such as mean spike count [Hubel and Wiesel, 1959], mean firing rate [Perkel et al., 1967], time to first spike [Johansson and Birznieks, 2004], interspike interval distribution [Kang and Amari, 2008, Churchland et al., 2010], or average pairwise spike train distance [Naud et al., 2011]. Such partial quantification blinds the experimenter to other quantifiable statistical structures, which results at least in poor data utilization but can even have rather dire consequences in experimental science. For instance, searching for neurons *in vivo* using a typical statistical feature associated with a neural code results in selection bias in favor of that specific neural code. In this paper, we follow an alternate approach of designing a two-sample test on point processes through the use of strictly positive definite kernels on spike trains, since such test does not limit itself to a particular statistical feature [Gretton et al., 2007, Sriperumbudur et al., 2008].

The popularity of positive definite kernels stems from the observations that, first, they are inner products in some feature space that is often nonlinearly related to the input space, second, they can be defined on rather abstract spaces such as non-Euclidean space, and third, they implicitly impose a structure of the space. These simple observations allow the extension of many standard linear algorithms designed on the Euclidean space to nonlinear algorithms on abstract spaces by merely substituting the linear inner product with a suitable kernel, known as the kernel trick. This approach has been widely exploited in the context of text mining, bioinformatics and information retrieval where kernels have been defined on strings, bag of words, and graphs [Schölkopf and Smola, 2002]. The usefulness of kernel methods has not gone unnoticed in the neuroscience literature, where several successful attempts have been made on designing positive definite kernels on spike trains [Shpigelman et al., 2005, Eichhorn et al., 2004, Schrauwen and Campenhout, 2007, Paiva et al., 2009], and applying them in practical problems such as classification for brain machine interface [Shpigelman et al., 2005, Eichhorn et al., 2004], and somatosensory stimulation [Li et al., 2011]. Positive definite kernels, however, may fall short in the context of representing an arbitrary probability law—kernels with such a property has recently been explored and are called *characteristic* kernels [Sriperumbudur et al., 2008]—and therefore, they cannot be readily applied for testing equality of distributions. As we will demonstrate in detail later, the available spike train kernels in the literature are not characteristic. It has, however, been shown that *strictly positive definite*[1] (spd) kernels (a subset of positive definite kernels) are characteristic [Sriperumbudur et al., 2010]. Given this recent development, and other interesting aspects of positive definite kernels, we focus on exploring such kernels for spike trains.

We consider two separate approaches for designing spd kernels, where both approaches follow two simple steps; first, we injectively map the spike trains to a more structured space, and second, we define spd kernels on the latter spaces that induces a kernel in the spike train space. The proposed approaches differ in their respective representation of the spike trains:

[1] Strict in a measure theoretic sense which is stronger than the usual countable summation sense (see definition 1).

(i) a *stratified space* which is formed by a disjoint union (direct sum) of Euclidean spaces, and (ii) a *functional space* where they are represented as functions over time in L_2. In the stratified space approach, we utilize a novel family of kernels in the Euclidean space to form spd kernels on spike trains, while in the functional space we exploit novel kernels in L_2. We submit that these kernels should also, (i) involve as few parameters as possible, since, from a practical standpoint, a user is forced to perform an exhaustive search (evaluation) over all possible combinations of these parameters to achieve the best (statistically correct) result, and (ii) be computationally efficient. Finally, we also explore what attributes of the point process these kernels are sensitive to.

The rest of the paper is organized as follows. In Section 2, we discuss the notion of kernel based divergence and dissimilarity measures, and briefly explore the existing positive definite kernels in the literature based on binning. In Section 3, we introduce a family of novel strictly positive definite kernels by representing the space of spike trains as a direct sum of Euclidean spaces. Although these kernels are efficient to compute, they suffer from estimation issues in practice. In order to alleviate this drawback, we introduce another family of strictly positive kernels in Section 4, by representing a spike train as a function in an appropriate L_2 space. We also examine a few existing kernels on this representation and discuss their non-strictness. In Section 5 we explore the characteristics of the existing and proposed kernels using kernel principal component analysis, and compare their performance on simulated data in the context of hypothesis testing. In Section 6, we apply these kernels on real neural recordings, and in Section 7, we conclude the paper with discussions on the practical aspects of these kernels and their applicability, as well as other uses of such kernels. A MATLAB implementation of the kernels and methods introduced in this paper can be found online[2].

2 Background

2.1 Kernel based divergence and dissimilarity

We start our discussion in an abstract topological space for technical completeness. Later we will introduce a topological space for spike trains. Let $(\mathfrak{X}, \tau)$ be a topological space and $\mathcal{F} = \mathfrak{B}(\tau)$ be the Borel σ-algebra. We will focus on measures on the measurable space $(\mathfrak{X}, \mathcal{F})$.

Definition 1 (Strictly positive definite kernel) *Let $(\mathfrak{X}, \mathcal{F})$ be a measurable space, and $K : \mathfrak{X} \times \mathfrak{X} \to \mathbb{C}$ be a $(\mathcal{F} \otimes \mathcal{F}/\mathfrak{B}(\mathbb{C}))$-measurable function. The kernel K is called* positive definite *if and only if for any finite non-zero complex Borel measure $\mu : \mathcal{F} \to \mathbb{C}$, $\iint K(x, y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu^*(y) \geq 0$, where $(\cdot)^*$ denotes complex conjugation. Furthermore, if the inequality is strict, then the kernel K is called* strictly positive definite*.*

Notice that this definition of (integrally) strict positive definiteness [Sriperumbudur et al., 2009] is a generalization of the usual definition that involves countable summation rather than integration [Pinkus, 2004].

[2] `http://code.google.com/p/spiketrainlib`

Let $\mathcal{M}_+$ denote the set of probability measures on $(\mathfrak{X}, \mathcal{F})$. When $\mathfrak{X}$ is the space of spike trains, $\mathcal{M}_+$ is the space of point processes (random process that generates spike trains). Given a positive definite kernel $K : \mathfrak{X} \times \mathfrak{X} \to \mathbb{C}$, we define a measure of dissimilarity $\mathcal{D}_K : \mathcal{M}_+ \times \mathcal{M}_+ \to \mathbb{R}_0^+$ as,

$$\mathcal{D}_K(P, Q) = \iint K(x, y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y), \tag{1}$$

where $\mu = P - Q$. Due to the positive definiteness of K, $\mathcal{D}_K(P, Q)$ is non-negative, and $P = Q$ implies $\mathcal{D}_K(P, Q) = 0$. A divergence measure $\mathbb{D}_K(P, Q) : \mathcal{M}_+ \times \mathcal{M}_+ \to \mathbb{R}_0^+$, on the other hand, can be constructed by incorporating a spd kernel instead of a pd kernel in (1) [Gretton et al., 2007]. Due to the strict positive definiteness of K, $\mathbb{D}_K(P, Q)$ is non-negative, and zero if and only if $P = Q$. As a matter of fact, one can show that the resulting divergence (dissimilarity) is a squared (pseudo-)metric induced from the distance in the reproducing kernel Hilbert space (RKHS) [Diks and Panchenko, 2007, Sriperumbudur et al., 2009].

A dissimilarity measure is a nonnegative statistic that assumes zero value if two probability laws are identical [Pekalska and Duin, 2005]. For example, if $\mathfrak{X} = \mathbb{R}$ and $K(x, y) = x \cdot y$, then $\mathcal{D}_K(P, Q) = (\mathrm{E}_{X \sim P}[\mathbf{X}] - \mathrm{E}_{Y \sim Q}[\mathbf{Y}])^2$, i.e. the pd kernel $K(x, y) = x \cdot y$ only compares the means of the random variables $X \sim P$ and $Y \sim Q$. A dissimilarity statistic, thus, creates an equivalence class where two probability laws are indistinguishable even if they are not the same, and the statistical inference only operates on the quotient space. A divergence measure, on the other hand, is a nonnegative statistic that assumes zero value if and *only if* two probability laws are identical [Sriperumbudur et al., 2009], thus it compares the entire probability law. Therefore, divergence is a stricter and a stronger statistic than dissimilarity.

This feature of an spd kernel can be better explained in terms of the characteristic kernel [Sriperumbudur et al., 2008]. Mathematically, a kernel K is called characteristic if the map $P \mapsto \mathrm{E}_{X \sim P}[K(\cdot, X)]$ is injective. This condition implies that a characteristic kernel generates a unique functional representation of the probability law in the reproducing kernel Hilbert space, which cannot be achieved by a pd kernel as shown in the case of $K(x, y) = x \cdot y$. In terms of characteristic kernel, therefore, a measure of divergence can be derived by measuring the distance between these unique functional representations of the corresponding probability laws. This distance is exactly given by the square root of eq (1). It has been recently shown that spd kernels are characteristic in nature, and that explains the advantage of spd kernels over pd kernels in the context of hypothesis testing [Sriperumbudur et al., 2009].

2.2 Binned kernel

One of the main difficulties of dealing with spike trains is the lack of a natural structural representation, such as a Euclidean structure [Seth et al., 2010a]. Among several existing approaches of inducing structure in this space, the simplest and most popular is *binning* [Victor, 2002]. In simple terms, binning counts the number of spikes that appear in prefixed segments of time referred to as *bins*, thus effectively representing a spike train on a grid in a Euclidean space where the dimensionality of the space is governed by the number of bins. Although this approach is widely used in practical applications for its simplicity and effectiveness, it has disadvantages that it only considers the number of spikes within a bin rather than their exact locations, and introduces discontinuity in time. The exact locations

can be, in principle, captured by reducing the size of these bins; however, it also increases the dimensionality of the Euclidean representation, increasing the data requirements for proper estimation.

Since the binned representation projects a spike train in a Euclidean space, a kernel on spike trains can be easily implemented using kernels on the Euclidean space such as l_2 inner product or the Gaussian kernel. However, this approach is not preferred in practice since the dimensionality of this space i.e. the number of bins is usually high, making the realizations relatively sparse. In addition, the kernels cannot easily incorporate the similarity of spikes in temporally neighboring bins – spikes from different bins do not interact. The count kernel K_{count} defined as the product of total number of spikes is an extreme case where there is only one bin. Note that the dissimilarity induced by the count kernel is simply the squared difference between the mean number of spikes.

Shpigelman et al. [2005] has considered a different approach of designing kernels on binned representation, where the resulting non-trivial kernel has been referred to as spikernel. Spikernel is, perhaps, the earliest kernel designed for spike trains that has been explored in the literature. The biologically motivated design of this kernel relies on measuring the similarity between arbitrary segments (of equal lengths) of the binned spike trains. Spikernel is positive definite by design since it explicitly constructs a mapping to the feature space, and computes the inner product between two mapped spike trains to evaluate the kernel. However, this kernel is not strictly positive definite. Consider two spike trains such that one is a *jittered* version of the other, but they both have the same spike counts given a certain bin resolution. Then the Gram matrix formed by these two spike trains has zero determinant. In fact, any kernel defined on the binned representation is only positive definite but not strictly positive definite due to this restriction.

It should be noted that spikernel involves *five* free parameters, including bin size, and it is computationally expensive. Although the appropriate value of the bin size can be estimated from the time resolution of the spike train, there are no formal method of choosing the other parameters. Therefore, it is suggested that a user perform an exhaustive search over the parameter space, making the kernel difficult to use in practice. We observe in our simulation that this kernel performs very well compared to other strictly positive definite kernels. A possible explanation of this observation is the higher *degrees of freedom*.

A similar approach has also been explored by Eichhorn et al. [2004] by using edit distance (alignment score) between two strings to binned spike trains to construct an explicit mapping to the feature space. This approach has very similar advantages and disadvantages.

3 Stratified kernel

Due to the lack of strict definiteness of the binned representation, it is essential to investigate other richer representations of the spike train space. In this section, we follow the stratified representation and introduce a set of strictly definite kernels on spike trains.

Stratified representation Let Ω be the set of all finite length spike trains, that is, each $\omega \in \Omega$ can be represented by a finite set of (distinct) action potential timings $\{t_1, t_2, \ldots, t_n\}$ where n is the number of spikes. Assuming *finite number of action potentials* within the time interval of interest $\mathcal{T} \subset \mathbb{R}$, one can partition the non-Euclidean spike train space Ω in disjoint

partitions $\Omega_0, \Omega_1, \cdots$ such that Ω_n contains all possible spike trains with exactly n action potentials within $\mathcal{T}$. We call this method stratification [Victor, 2002]. Notice that $\Omega = \bigcup_{n=0}^{\infty} \Omega_n$. Ω_0 has only one element representing the empty spike train (no action potential). For $n \neq 0$, Ω_n is essentially $\mathcal{T}^n$, since n action potentials can be fully described by n (ordered) time instances and vice versa. Without loss of generality, let $\Omega_n = \mathcal{T}^n$, hence obtaining Euclidean space representation of each Ω_n. Define a Borel algebra of Ω by the σ-algebra generated by the union of Borel sets defined on the interval; $\mathfrak{B}(\Omega) = \sigma\left(\bigcup_{n=0}^{\infty} \mathfrak{B}(\Omega_n)\right)$. Note that any measurable set $A \in \mathfrak{B}(\Omega)$ can be partitioned into $\{A_n = A \cap \Omega_n\}_{n=0}^{\infty}$, each measurable in corresponding measurable space $(\Omega_n, \mathfrak{B}(\Omega_n))$. Any finite point process can be defined as a probability measure P on the measurable space $(\Omega, \mathfrak{B}(\Omega))$ [Daley and Vere-Jones, 1988].

Generic kernel design Although the stratification approach has been used by [Victor, 2002] to estimate the mutual information, this approach has not been explicitly exploited to design kernels. However, it provides an opportunity to build kernels on Ω by combining kernels on $\mathbb{R}^n$ using the following theorem.

Theorem 1 (Stratified strictly positive definite kernel) *Let $\{K^{(n)}\}_{n=0}^{\infty}$ be a family of bounded strictly positive definite kernels $K^{(n)} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ for every $n \in \mathbb{N}$. Define,*

$$K_s(\omega_i, \omega_j) = \begin{cases} K^{(n)}(\omega_i, \omega_j) & \text{if both } \omega_i, \omega_j \in \Omega_n, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Then K_s is a strictly positive definite kernel on spike trains,

Proof 1 *For any non-zero Borel measure μ, we can decompose it as $\mu = \sum_n \mu|_n$.*

$$\begin{aligned} \int K_s(\omega_i, \omega_j)\, \mathrm{d}\mu(\omega_i)\, \mathrm{d}\mu(\omega_j) &= \int \sum_n K^{(n)}(\omega_i, \omega_j)\, \mathrm{d}\mu|_n(\omega_i)\, \mathrm{d}\mu|_n(\omega_j) \\ &= \sum_n \int K^{(n)}(\omega_i, \omega_j)\, \mathrm{d}\mu|_n(\omega_i)\, \mathrm{d}\mu|_n(\omega_j) > 0 \end{aligned}$$

where $\cdot|_n$ denotes the restriction of the measure to $\mathcal{F}_n$, and the interchange of sum and integral is justified by the boundedness of spd kernels.

The corresponding divergence measure can be simplified as

$$\mathbb{D}_s(P, Q) = \sum_{n=0}^{\infty} \iint K^{(n)}(\omega_i, \omega_j)\, \mathrm{d}\mu|_n(\omega_i)\, \mathrm{d}\mu|_n(\omega_j) \tag{3}$$

where $\mu = P - Q$.

Class of kernels The individual kernels on each strata can be chosen from a generic kernel as follows.

Definition 2 ($\mathfrak{S}$-admissible functions) *Let $\mathfrak{S}$ be the Schwarz space of rapidly decreasing functions. $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is $\mathfrak{S}$-admissible if for every $h \in \mathfrak{S}$, the following integral equation has a solution f, $\int g(x, u) f(u)\, \mathrm{d}u = h(x)$.*

Theorem 2 (Composition kernels) *If $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ is a $\mathfrak{S}$-admissible or a strictly positive definite function, ξ is a measure, and $\operatorname{supp}(\xi) = \mathbb{R}^n$, then the following kernel is strictly positive definite on $\mathbb{R}^n$, $K(x, y) = \int g(x, u)g^*(y, u)\,\mathrm{d}\xi(u)$.*

Proof 2 *The proof of this theorem can be found in [Seth et al., 2010b, Theorem 1].*

Examples of basis functions $g(x, u)$ used for composition kernels are e^{ixu}, $\mathbb{I}(x \leq u)$, and $e^{-(x-u)^2}$ for $\mathbb{R}$. In $\mathbb{R}^n$, we can use the tensor product kernels: $\prod_i g(x_i, u_i)$.

There are a few notable special cases of this type of kernels. Notice that, if $K_1^{(n)} = 1$ for all $n \in \mathbb{N}$, which induces a positive definite kernel, the dissimilarity becomes, $\mathcal{D}_1(P, Q) = \sum_n (P(\Omega_n) - Q(\Omega_n))^2$, which corresponds to the Cramér–von-Mises (C-M) statistic for the count distributions $[P(\Omega_n)]_n$ and $[Q(\Omega_n)]_n$. Intuitively this kernel only compares if two spike trains have the same number of spikes, and therefore, it is not spd. On the other hand, Gaussian kernels on $\mathbb{R}^n$ can be used to construct spd kernels on the space of spike trains. However, this process requires choosing kernel sizes for each strata.

An interesting kernel is found by using the following composite kernels on $\mathbb{R}^n$,

$$K_s^{(n)}(\omega_i, \omega_j) = \int \mathbb{I}(\omega_i \leq \omega)\mathbb{I}(\omega_j \leq \omega)\,\mathrm{d}(\omega)$$

where $\mathbb{I}(\omega_i \leq \omega_j) = \prod_d \mathbb{I}(\omega_i^d \leq \omega_j^d)$, and $\omega = [\omega^1, \ldots, \omega^d]$. The corresponding divergence is given by,

$$\mathbb{D}_s = \sum_n \int \left[P(\Omega_n)F_P(\omega^{(n)}) - Q(\Omega_n)F_Q(\omega^{(n)})\right]^2 \mathrm{d}(\omega^{(n)}) \tag{4}$$

where F denotes the cumulative distribution function in $\mathbb{R}^n$. Intuitively, this kernel not only compares if two spike trains have the same spike counts, but it also compares the positions of the spikes. Notice that it is very similar to the Cramér–von-Mises type divergence proposed in [Seth et al., 2010a]. The advantage of this type of kernel is that it is parameter free, unlike the Gaussian kernel.

Computational issues The kernels defined on the stratification space are efficient to compute, and easy to design. However, this type of kernels suffer from estimation issues for small sample size, especially when the count distribution is flat (not concentrated). In such circumstances the samples become scattered in different strata Ω_n, and the divergence become difficult to evaluate for a particular strata. In addition, the stratification approach suffers from the curse of dimensionality, when the number of spikes in a spike train is large. Hence, it may be necessary to reduce the time window of observation to reduce this dimensionality.

4 Functional kernel

It is somewhat obvious that the estimation issues of stratified kernel are a direct consequence of the fact that for such kernels two spike trains with different spike counts do not interact. To alleviate this issue we follow a different representation.

Functional representation Let $\mathfrak{F}$ be the space of all L_2 integrable functions over $\mathcal{T}$ i.e. $\mathfrak{F} = L_2(\mathcal{T})$. Given $\omega = \{t_1, \ldots, t_n\}$, define a mapping $G : \Omega \to \mathfrak{F}$ as $G(\omega)(t) = \sum_{i=1}^n g(t, t_i)|_{\mathcal{T}}$

such that G is injective. When g is a translation invariant function that decays at ∞, $G(\omega)$ can be considered a smoothed spike train [van Rossum, 2001].

There are many different g's that make the mapping G injective. For example, when $g(x, y)$ is a bounded strictly positive definite function. To see this, consider two spike trains $\omega_1 = \{t_1^1, \ldots, t_n^1\}$ and $\omega_2 = \{t_1^2, \ldots, t_m^2\}$ such that all spike timings are distinct i.e. $\omega_1 \cap \omega_2 = \emptyset$, and assume that $G(\omega_1) = G(\omega_2)$, which implies that $|G(\omega_1) - G(\omega_2)|_{L_2}^2 = \mathbf{e}\mathbf{K}\mathbf{e}^\top = 0$ where $\mathbf{e} = [\underbrace{1, \ldots, 1}_{n}, \underbrace{-1, \ldots, -1}_{m}]$ and $[\mathbf{K}]_{ij} = \tilde{g}(t_i, t_j)$ where $t_i, t_j \in \omega_1 \cup \omega_2$ and $\tilde{g}(x, y) = \int g(u, x) g(u, y) \mathrm{d}u$. Now since $\mathbf{K}$ is full rank (by the virtue of strict positive definiteness of $\tilde{g}$), $G(\omega_1)(t) = G(\omega_2)(t)$ for all $t \in \omega_1 \cup \omega_2$ implies that $\omega_1 = \omega_2$. Popular continuous-time spike train smoothing by linear filtering with rectangular function or alpha-function [Dayan and Abbott, 2001] are also injective, given the width of the rectangular function is not more than length of $\mathcal{T}$.

We consider the σ-algebra of $L_2(\mathcal{T})$ to be the one that makes the map $G : \Omega \to L_2(\mathcal{T})$ measurable. Since G is measurable, we can define an induced probability measure U such that $U(L_2(\mathcal{T}) \backslash G(\Omega)) = 0$ and $U(G(A)) = P(A)$ for $A \in \sigma(\Omega)$. A (strictly) positive definite kernel $K(\omega_1, \omega_2)$ on Ω can be defined using a (strictly) positive definite kernel $\tilde{K}(\tilde{\lambda}_1, \tilde{\lambda}_2)$ on the L_2 space since,

$$\int K(\omega_1, \omega_2) \, \mathrm{d}P(\omega_1) \, \mathrm{d}Q(\omega_2) = \int \tilde{K}(\tilde{\lambda}_1, \tilde{\lambda}_2) \, \mathrm{d}U(\tilde{\lambda}_1) \, \mathrm{d}V(\tilde{\lambda}_2)$$

where $\tilde{\lambda} = G(\omega)$.

Existing pd kernels Paiva et al. [2009] has proposed the following kernel using exponential smoothing function,

$$K_{\mathrm{mCI}}(\omega_1, \omega_2) = \int_{\mathcal{T}} \tilde{\lambda}_1(t) \tilde{\lambda}_2(t) \, \mathrm{d}t. \tag{5}$$

If the smoothing function is non-negative and locally centered around the spike, the conditional mean $\int K_{\mathrm{mCI}}(\omega, \cdot) \, \mathrm{d}P(\omega)$ can be interpreted as an estimate of the marginal firing rate (intensity function). The kernel K_{mCI} is generated by the trivial inner product kernel in the L_2 space, thus the name memoryless cross intensity (mCI). Notice that several smoothing functions can also be used in this context. For example, the use of exponential smoothing function is biologically motivated [van Rossum, 2001], and it is computationally simpler due to the characteristic of Laplacian function [Rao et al., 2011, Section 4.1]. The Laplacian function is found by taking the inner product of two exponential functions [Paiva et al., 2009], while the inner product of two Gaussian functions results in a Gaussian function. Notice that, K_{mCI} induces a spike train RKHS where a van Rossum like distance [van Rossum, 2001] is induced by the inner product, and this kernel involves only one free parameter i.e. the size of the smoothing function. Similar kernels have been proposed in [Schrauwen and Campenhout, 2007, Houghton, 2009, Paiva et al., 2010]. However, it is easy to see that these kernels are not strictly positive definite. For example, consider the spike trains $\{1, 2\}$, $\{1\}$, and $\{2\}$. Then Gram matrices generated by these kernels are singular. Therefore, the dissimilarity statistic $\mathcal{D}_{K_{\mathrm{mCI}}}$ is not a divergence.

Paiva et al. [2009] has also proposed a nonlinear, and a richer, extension of this kernel referred to as the nonlinear cross intensity (nCI). Direct sum of positive definite kernels de-

fines a positive definite kernel, and this has motivated the following construction [Paiva et al., 2009],

$$K_{\text{nCI}}(\omega_1, \omega_2) = \int_{\mathcal{T}} \exp\left\{-\frac{1}{\sigma}\left(\tilde{\lambda}_1(t) - \tilde{\lambda}_2(t)\right)^2\right\} \mathrm{d}t. \tag{6}$$

This kernel performs better than K_{mCI}, a possible explanation being the nonlinear nature of the kernel. However, it should be noted that it involves one more free parameter i.e. the kernel width of the outer Gaussian kernel. Also, this kernel is difficult to compute, and a rectangular smoothing function is used to reduce the computational complexity Paiva et al. [2009]. However, this kernel is again not strictly positive definite, and that can be again shown by a counter example. Consider the following spike trains $\{1, 2, 3\}$, $\{1, 3\}$, $\{1, 2\}$, and $\{1\}$, then the resulting Gram matrix is singular.

Schoenberg kernel In order to establish a strictly positive definite kernel on this representation, we generalize the radial basis type kernel on Euclidean space.

Definition 3 (Completely monotone functions [Schaback and Wendland, 2001]) *A function $\phi : (0, \infty) \to \mathbb{R}$ is said to be* completely monotone on $[0, \infty)$ *if $\phi \in \mathbf{C}^{\infty}(0, \infty)$, continuous at zero, and $(-1)^l \phi^{(l)}(r) \geq 0, \quad l \in \mathbb{N}, r > 0$ where $\phi^{(l)}$ denotes the l-th derivative.*

Examples of completely monotone functions on $[0, \infty)$ are $\alpha, e^{-\alpha x}, \frac{1}{(x+\alpha^2)^{\beta}}$ where α and β are constants [Berg et al., 1984].

Theorem 3 (Strictly positive definite Schoenberg kernels on L_2) *If a function $\phi : [0, \infty) \to \mathbb{R}$ is completely monotone on $[0, \infty)$ but not a constant function, then $K(x, y) = \phi(\|x - y\|^2)$ is strictly positive definite on a separable and compact subset of L_2 where $\|\cdot\|$ denotes the L_2 norm.*

Proof 3 *(Sketch) Since ϕ is a completely monotone function, it has a Laplace transform representation, $\phi(z) = \int_0^{\infty} e^{-zt} \mathrm{d}m(t)$ where m is a finite positive Borel measure on $\mathbb{R}$. Following [Christmann and Steinwart, 2010, Theorem 2.2] $\exp(-t\|x - y\|^2)$ is universal on a separable compact subset of L_2 for $t > 0$, and it implies strictly positive definiteness [Sriperumbudur et al., 2010].*

Note that proof relies on the restriction of maximum number of spikes within a compact interval $\mathcal{T}$, which is a biological and experimental restriction, the space of spike trains Ω is compact (and separable), ane hence, the image of Ω in $\mathfrak{F}$ is also compact (and separable).

Instances of Schoenberg kernel A typical instance of Schoenberg kernel considers both ϕ and the smoothing function g to be exponential function i.e.

$$K^e_{\text{Sch}}(\omega_1, \omega_2) = \exp\left\{-\int_{\mathcal{T}} \frac{1}{\sigma}\left(\tilde{\lambda}_1(t) - \tilde{\lambda}_2(t)\right)^2 \mathrm{d}t\right\}. \tag{7}$$

We refer to this kernel as Schoenberg (e). Notice that this kernel has been mentioned by [Paiva et al., 2009], but it has not been explored further, and its strict definiteness has not been established. On the other hand, Paiva et al. [2009] has suggested the kernel K_{nCI} as a substitute and simplification of K^e_{Sch}. Notice that if the smoothing process is a linear filter,

the bounded norm of the smoothing process holds when the smoothing filter is bounded and the total number of spikes is bounded. This condition virtually holds for bounded duration spike trains.

K^e_{Sch} also involves an extra free parameter besides the smoothing parameter. However, the proposed approach allows other alternatives which might not involve any parameter. One such option is the Heaviside function. It is easy to see that the smoothed representation of a spike train using a Heaviside function is simply the realization of the corresponding counting process. To differentiate the use of Heaviside function, we refer to the resulting kernel as Schoenberg (i).

$$K^i_{\text{Sch}}(\omega_1, \omega_2) = \exp\left(-\frac{1}{\sigma}\int_{\mathcal{T}} (I_{\omega_1}(t) - I_{\omega_2}(t))^2 \,\mathrm{d}t\right)$$

where $I_\omega(t) = \sum_j \mathbb{I}(t > t^\omega_j)$. Notice that using Heaviside function is also computationally simpler since the inner product of two smoothed functions simply depends on the locations of the spikes in an ordered set. Conceptually, however, due to the cumulative temporal counting nature of this approach, the resulting kernels depend on the initial spikes more than the latter ones. However, we observe that this kernel nonetheless work well in practice with the advantage of using one less free parameter.

Notice that compared to stratified kernel, functional kernels do not suffer from estimation issues in extreme conditions of high counts and low sample size, since they can compare two spike trains involving different spike number of spikes. But this is achieved at the cost of higher computational complexity. We provide detailed description of these kernels in Table 1. This table also includes a recently proposed kernel K_{reef} [Fisher and Banerjee, 2010], specialized for spike response model 0 with reciprocal exponential exponential function (REEF) activation. This kernel integrates out the unknown parameters, hence eliminates the need to estimate any free parameters. However, the strict definiteness of the kernel has not been pursued.

5 Features

So far we have discussed kernels and induced dissimilarities and divergences, but we have not addressed what they are actually sensitive to when applied to data. In principle dissimilarities are only sensitive to certain statistical differences, while divergences are sensitive to arbitrary statistical differences, however, in practice their sensitivity depends on the choice of kernel and is limited by data. We explore the issue of sensitivity in two aspects: (1) decomposition of divergence in orthonormal projections using KPCA, and (2) sensitivity analysis for explicit features using hypothesis testing.

5.1 Kernel PCA based decomposition

Principal component analysis (PCA) is widely used to extract "features", linear projections of data in a finite dimensional Euclidean space obtained by finding directions of greatest variances. These features, or principal components (PCs), are simply obtained by the eigen-decomposition of the data covariance matrix. Likewise, in kernel PCA (KPCA), one decom-

name	Kernel	spd	time complexity	# parameters	reference
Count	$K_{\text{count}}(\omega_1, \omega_2) = \lvert\omega_1\rvert\lvert\omega_2\rvert$	no	$\mathcal{O}(N)$	0	
Spikernel[a]	$\phi_{\text{spikernel}}(\omega, \cdot) = \sum_{i \in I_{n,\lvert B(\omega)\rvert}} \mu^{d_{\mathbb{R}}(B_i(\omega),\cdot)} \lambda^{\lvert B(\omega)\rvert - i_1}$	no	$\mathcal{O}(\lvert B\rvert^2 n)$	5	[Shpigelman et al., 2005]
mCI	$K_{\text{mCI}}(\omega_1, \omega_2) = \sum_{i,j}^{N} \exp\left(-\lvert t_1^i - t_2^j\rvert/\tau\right)$	no	$\mathcal{O}(N \log N)$	1	[Paiva et al., 2007]
nCI	$K_{\text{nCI}}(\omega_1, \omega_2) = \frac{1}{T}\int_{\mathcal{T}} \exp\left(-\left(f_r(\omega_1)(t) - f_r(\omega_2)(t)\right)^2/\sigma\right) \mathrm{d}t$	no	$\mathcal{O}(N \log N)$	2	[Paiva et al., 2009]
Schoenberg (e)	$K^e_{\text{Sch}}(\omega_1, \omega_2) = \exp\left(-d^2_{\text{mCI}}(\omega_1, \omega_2)/\sigma\right)$	yes	$\mathcal{O}(N \log N)$	2	[Paiva et al., 2009]
Schoenberg (i)	$K^i_{\text{Sch}}(\omega_1, \omega_2) = \exp\left(-\frac{1}{\sigma}\int_{\mathcal{T}} (I_{\omega_1}(t) - I_{\omega_2}(t))^2 \mathrm{d}(t)\right)$	yes	$\mathcal{O}(N \log N)$	1	
Stratified	$K^{(n)}_{\text{strat}}(\omega_i, \omega_j) = \int K_{\mathbb{R}}(\omega_i \leq \omega) K_{\mathbb{R}}(\omega_j \leq \omega) \mathrm{d}(\omega)$	yes	$\mathcal{O}(N)$	1 ($K_{\mathbb{R}}$)	
REEK	$K_{\text{reef}}(\omega_1, \omega_2) = \sum_{i,j}^{N} \frac{(T - t_1^i)(T - t_2^j)}{\left(2T - t_1^i - t_2^j\right)^2}$	*unknown*	$\mathcal{O}(N^2)$	0	[Fisher and Banerjee, 2010]

Table 1: List of spike train kernels, existing ones and the ones proposed in the paper. N is the average number of spikes per spike train. f_r is a rectangular smoothing function, the time interval $\mathcal{T}$ is $[0, T)$. $K_{\mathbb{R}}$ is a Gaussian kernel on the Euclidean distance. d_{mCI} is the distance induced by mCI kernel, I_ω is a Heaviside function. $B(\cdot)$ denotes binned spike train, $\lvert\cdot\rvert$ is cardinality, $I_{n,m}$ is the set of segments of length n from total length m, B_i is a segment of B with first element at i_1. Unfortunately, we could not find available implementation nor was able to implement the kernel from Eichhorn et al. [2004] due to lack of details.

[a] ϕ is the mapping from the binned spike train to the feature space where the kernel is evaluated as the inner product.

poses the covariance operator of the Hilbert space to find PCs that are now "feature functionals", nonlinearly related to the input space [Schölkopf et al., 1998]. Each PC is orthogonal to each other, and depends on the data as well as the kernel. Each kernel gives a different set of features, and the set of features for the same kernel depends on the spike trains in hand. This makes the interpretation of "features" difficult, yet it provides a useful qualitative insight of interpreting kernel induced divergence.

Divergence and KPCA projections These KPCA features can be used to decompose the kernel induced divergences. Given a set of orthonormal basis $\{\Phi_i\}_{i=1}^{\infty}$ and eigenvalues λ_i provided by KPCA i.e.

$$\lambda_i \Phi_i(x) = \int K(x,y)\Phi_i(y)\,\mathrm{d}(P+Q)(y),$$

the dissimilarity or divergence (1) can be trivially decomposed as,

$$\begin{aligned}
\mathcal{D}_K(P,Q) &= \iint K(x,y)\,\mathrm{d}(P-Q)(x)\,\mathrm{d}(P-Q)(y) \\
&= \iint \sum_i \lambda_i \Phi_i(x)\Phi_i(y)\,\mathrm{d}(P-Q)(x)\,\mathrm{d}(P-Q)(y) \\
&= \sum_i \lambda_i \left(\int \Phi_i(x)\,\mathrm{d}P(x) - \int \Phi_i(y)\,\mathrm{d}Q(y) \right)^2 . \qquad (8)
\end{aligned}$$

Hence, the total divergence (dissimilarity) is a sum of squared distances between the means $\int \Phi_i(x)\,\mathrm{d}P(x)$ and $\int \Phi_i(x)\,\mathrm{d}Q(x)$ of each PC direction weighted by the eigenvalues. This allows us to visualize the differences of two point processes in each PC direction.

In practice KPCA and decomposition of divergence can be computed with simple matrix algebra. Let $X = \{x_1^{(1)} \ldots, x_m^{(1)}, x_1^{(2)}, \ldots, x_n^{(2)}\}$ be the set of spike trains originating from two point processes, and let $\mathbf{K}$ be the Gram matrix generated by these spike trains i.e. $\mathbf{K}_{ij} = \kappa(x_i, x_j)$ where κ is a pre-defined kernel as discussed in the previous section. Then it can be easily seen that the estimated divergence (dissimilarity) is,

$$\mathbb{D}_\kappa = \mathbf{e}^\top \mathbf{K} \mathbf{e} \quad \text{where } \mathbf{e} = [\underbrace{1,\ldots,1}_{m}, \underbrace{-1,\ldots,-1}_{n}].$$

Now, let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{(m+n)}]$ be the matrix of eigenvectors obtained from the eigendecomposition related to KPCA i.e. $\mathbf{K}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$ where Λ is a diagonal matrix containing the eigenvalues of the respective eigenvectors, and corresponding eigenfunctions in the RKHS are simply $\tilde{\Phi}_k(\cdot) = \sum_i \mathbf{a}_k(i)\kappa(z_i, \cdot)$ where z_i's are elements of X. Next, let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{(m+n)}]$ where $\mathbf{b}_k$ is simply the projection of the spike trains in X on the k-th eigenfunction i.e. $\kappa(z_i, \cdot) = \sum_k \mathbf{b}_k(i)\tilde{\Phi}_k(\cdot)$, then $\mathbf{b}_k = \mathbf{K}\mathbf{a}_k/\lambda_k = \mathbf{a}_k$. Notice that the normalizing factor λ_k is essential since the eigenfunctions simply orthogonal but not orthonormal in this formulation, and $\langle \tilde{\Phi}_k(\cdot), \tilde{\Phi}_k(\cdot) \rangle = \lambda_k$. Then, sum of squared distances between the means of the projections is

$$\sum_{k=1}^{m+n} \langle \mathbf{e}^\top \mathbf{b}_k, \mathbf{e}^\top \mathbf{b}_k \rangle = \sum_{k=1}^{m+n} \lambda_k \mathbf{e}^\top \mathbf{b}_k \mathbf{b}_k^\top \mathbf{e} = \mathbf{e}^\top \mathbf{B}\mathbf{\Lambda}\mathbf{B}^\top \mathbf{e} = \mathbf{e}^\top \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top \mathbf{e} = \mathbf{e}^\top \mathbf{K}\mathbf{e} = \mathbb{D}_\kappa,$$

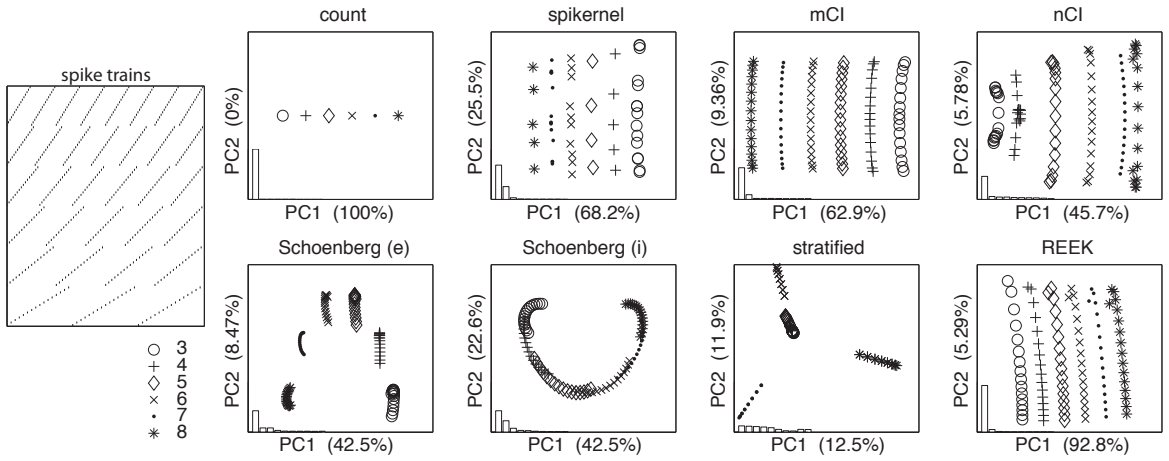


Figure 1: Two dimensional visualization of a periodic spike trains with a phase and period distribution using kernel PCA (KPCA). Horizontal axis corresponds to the projection on the eigenvector for the largest eigenvalue (PC1), and the vertical axis for the second largest (PC2). The inset bar graphs show the first ten normalized eigen-spectrum.

where the λ_ks appear again since the projections exist on a orthogonal basis and not orthonormal. Therefore, while discriminating two distributions through the KPCA projections we plot $\sqrt{\lambda_k}\mathbf{b}_k$ for the best two eigenvectors for which $\lambda_k \mathbf{e}^\top \mathbf{b}_k \mathbf{b}_k^\top \mathbf{e}$ are maximum since they explain away most of the divergence. Also, notice that KPCA is usually performed on the centered matrix $\mathbf{K}_c = \mathbf{HKH}$ where $\mathbf{H} = \mathbf{I}_{(m+n)} - \mathbf{1}_{(m+n)}\mathbf{1}_{(m+n)}^\top/(n+m)$. However, centering does not change the relation between divergence and the projections since $\mathbf{H}$ is simply absorbed in the eigenvector matrix $\mathbf{A}$.

Spike trains with count and location features In Figure 1, we visualize how a set of periodic spike trains with varying period and phase are embedded in the RKHS induced by each kernel. The dataset consiststs of 16 spike trains for each of 6 different periods each shifted such that the spike times uniformly spans the interval (total 96 spike trains). The two dimensional projection shows most kernels preserve the following "features", (i) count (number of spikes, or equivalently period) and (ii) phase. Notice that for each kernel we choose the free parameters from a grid of values that provides maximum value of the divergence (see the section 5.2 for details). The count kernel captures the first feature with PC1, and no other eigenvalue is significant, since the centered kernel matrix is rank 1. $K_{\text{spikernel}}$, K_{mCI}, K_{nCI}, K^e_{Sch}, and K_{reef} also seem to favor having the count feature as PC1, and in addition, PC2 mostly represents the phase feature i.e. the positions of the projections in this direction are ordered according to their phase. However, such interpretation should be made cautiously, since count and phase are not necessarily orthogonal in the RKHS. For example, K^i_{Sch} mixes the two features along many PCs, and stratified reveals strata as maximally separated clusters.

Next, we form two sets of spike trains by dividing the periodic spike trains of Figure 1 in two different ways to focus on each feature. Set A divides them in terms of count while maintaining the uniformity over phase within each condition. Set B divides the spike trains within each strata (of equal number of spikes) into early phase and late phase ones, thus maintaining count feature to be constant. Therefore, set A and B have the same KPCA

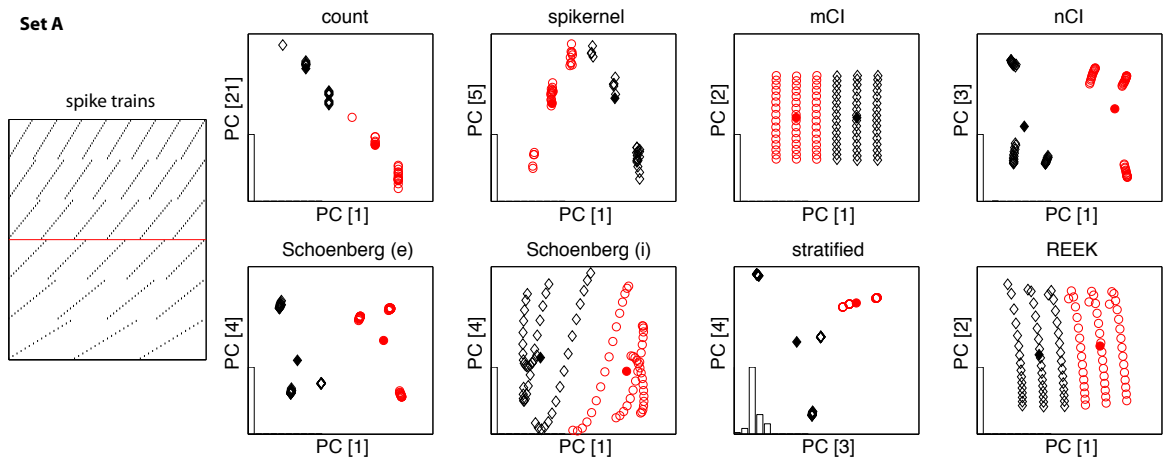


Figure 2: Visualization of maximally divergent PC for difference in the **count statistic**. The spike train distribution of Figure 1 is divided into two subsets with same phase distribution, but different count distribution – upper half of the raster has higher rate. Spike trains from each distribution conditioned on the count is plotted with different symbol, and the conditional mean is plotted as a solid symbol. Inset bar plots show the first ten normalized divergence (see (8)) per PC in the same order as eigenspectrum in Figure 1. Hypothesis testing rejected for all kernels with $p < 0.01$.

features, but the difference within are due to independent features (count statistic and location statistic [McFadden, 1965, Daley and Vere-Jones, 1988]).

All kernels successfully rejects the null for Set A as could be expected from sensitivity to count shown in Figure 1. In Figure 2, spike trains from each condition is projected to the PCs with highest divergence components. Interestingly, the PC1 captured the difference the most except for stratified kernel where the divergence was spread in several PCs. The projections show clear separation of the two clusters, however, individual projections show distinct structure of feature space.

For Set B (Figure 3), the count statistic was held constant, therefore K_{count} trivially fails to discriminate them. Except for count kernel, the separation is clearly made, and again the distinct structure of feature space contains the phase information. For stratified kernel, since the difference in phase are present in each of six strata, the divergence components are also spread. Note that the set of PCs with high divergence are oftentimes disjoint for set A and set B—$K_{\text{spikernel}}$, K_{nCI}, K^i_{Sch} and K^e_{Sch}, and K_{strat}—which is an indication that the count feature and location features are not mixed for this dataset.

Although in this example the two sets of spike trains were disjoint, and hence separable in features space, it is not crucial. We emphasize that it is not necessary for the kernel to *cluster* the sample points to derive a good divergence measure, but only the distance between the mean element matters i.e. even if the support of the projection of each class completely overlays on each other, theoretically, the difference between the probability laws will be captured in the distance between the corresponding mean elements *if the kernel is spd*.

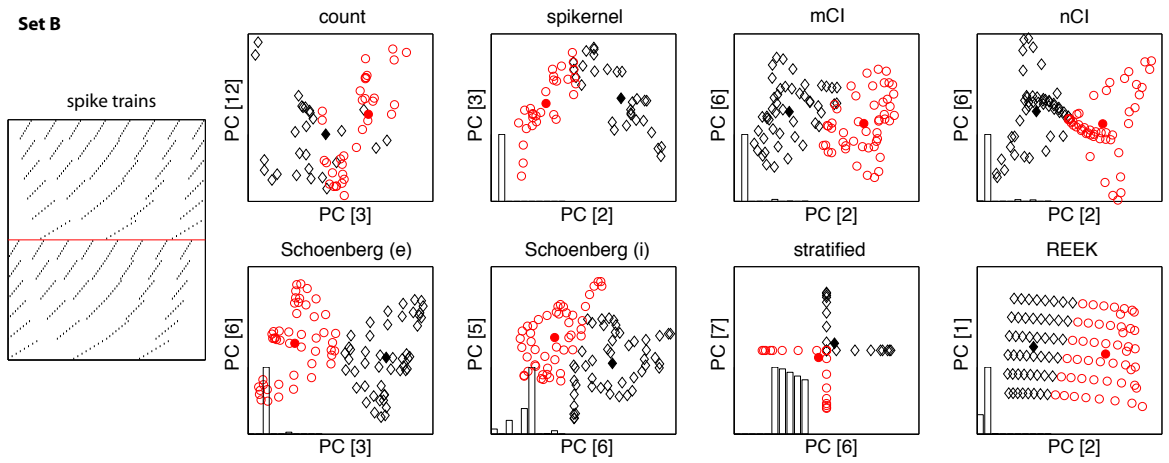


Figure 3: Visualization of maximally divergent PC for difference in the **location statistic**. The spike train distribution of Figure 1 (and Figure 2) is divided into two subsets with same count distribution, but different phase distribution. Hypothesis testing (as described in section 5.2) rejected the null for all kernels ($p < 0.05$) except for K_{count} ($p > 0.9$). Note that the scatter plot for count kernel is merely showing numerical error.

5.2 Hypothesis testing

In previous section, we used KPCA analysis to qualitatively assess feature sensitivity. Next, we quantify the power of each kernels with hypothesis testing i.e., given two sets of spike trains, we decide if the underlying point process that generated each of them are distinct. By comparing the performance for the detection of different statistics, we can quantify which features of interest the kernels are sensitive to. We also compare Wilcoxon rank sum test which a is non-parametric test widely used in neuroscience.

Surrogate test To generate a distribution of the statistic under the null hypothesis, we use the pooled observations as surrogate: Given realizations $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^m$, we randomly divide the pooled observations $\{x_1, \ldots, x_n, y_1, \ldots, y_m\}$ in segments of size n and m to generate the surrogate values. This procedure can be implemented efficiently by shuffling the Gram matrix generated by the pooled observations appropriately [Gretton et al., 2007, 2009]. Each bootstrapping had 9999 random shufflings of the kernel matrix. For the simulation studies, we report the rejection rate of the hypothesis from 200 independent trials.

Parameter selection Although strictly positive definite kernels provide an efficient way of inducing measures of divergence in the space of spike trains, an appropriate hypothesis testing requires choosing a suitable value for the free parameters, since the discriminability of two distributions heavily relies on the characteristics of the underlying kernel. Unfortunately, there are no established method for this task. A trivial solution to this problem could be to test the hypothesis over different values of the parameters and choose the one that provides best discriminability. But this approach faces the problem of *multiple testing*. To avoid such situation we consider a different approach proposed by Sriperumbudur et al. [2009] by re-defining the measure of divergence as $\mathbb{D}_{\mathcal{K}} = \max_{\kappa \in \mathcal{K}} \mathbb{D}_{\kappa}$ where $\mathcal{K}$ is a set of kernels, e.g. kernels over different parameter values. It has been shown that $\mathbb{D}_{\mathcal{K}}$ is a measure of divergence

if at least one κ in $\mathcal{K}$ is strictly positive definite [Sriperumbudur et al., 2009]. We exploit this fact, and use $\mathbb{D}_{\mathcal{K}}$ instead of $\mathbb{D}_{\kappa}$ to avoid multiple testing. We employ a grid of parameter values to construct $\mathcal{K}$ from a certain κ as follows.

As mentioned before, spikernel uses binned representation. To reduce computational load, we fix the bin size such that the number of bins would be 20 for the time interval of interest. Besides bin size, spikernel has four extra parameters [Shpigelman et al., 2005]. We fix $p_i = 1$, and evaluate the kernel for twelve different parameter settings by combining $\lambda \in \{0.5, 0.9\}, \mu \in \{0.7, 0.9\}, n \in \{3, 5, 10\}$ which are part of the range of parameters the authors chose [Shpigelman et al., 2005]. We use the implementation in C provided in the Spider package [3].

The free parameters for the other kernels are chosen to be 0.1, 0.5 and 0.9 quantiles of the numerator by random subsampling of spike trains, in addition to half of 0.1 quantile and twice the 0.9 quantile (total of 5 choices). This is a variation of common practice in kernel methods to match the kernel size to the scale of the data [Schölkopf, 1997]. When the parameters are nested ($K_{\text{nCI}}, K^{e}_{\text{Sch}}$), the overall scale parameter was chosen for each value of the parameter inside (e.g. size of the smoothing), thus resulting in 25 choices of parameter pairs in total.

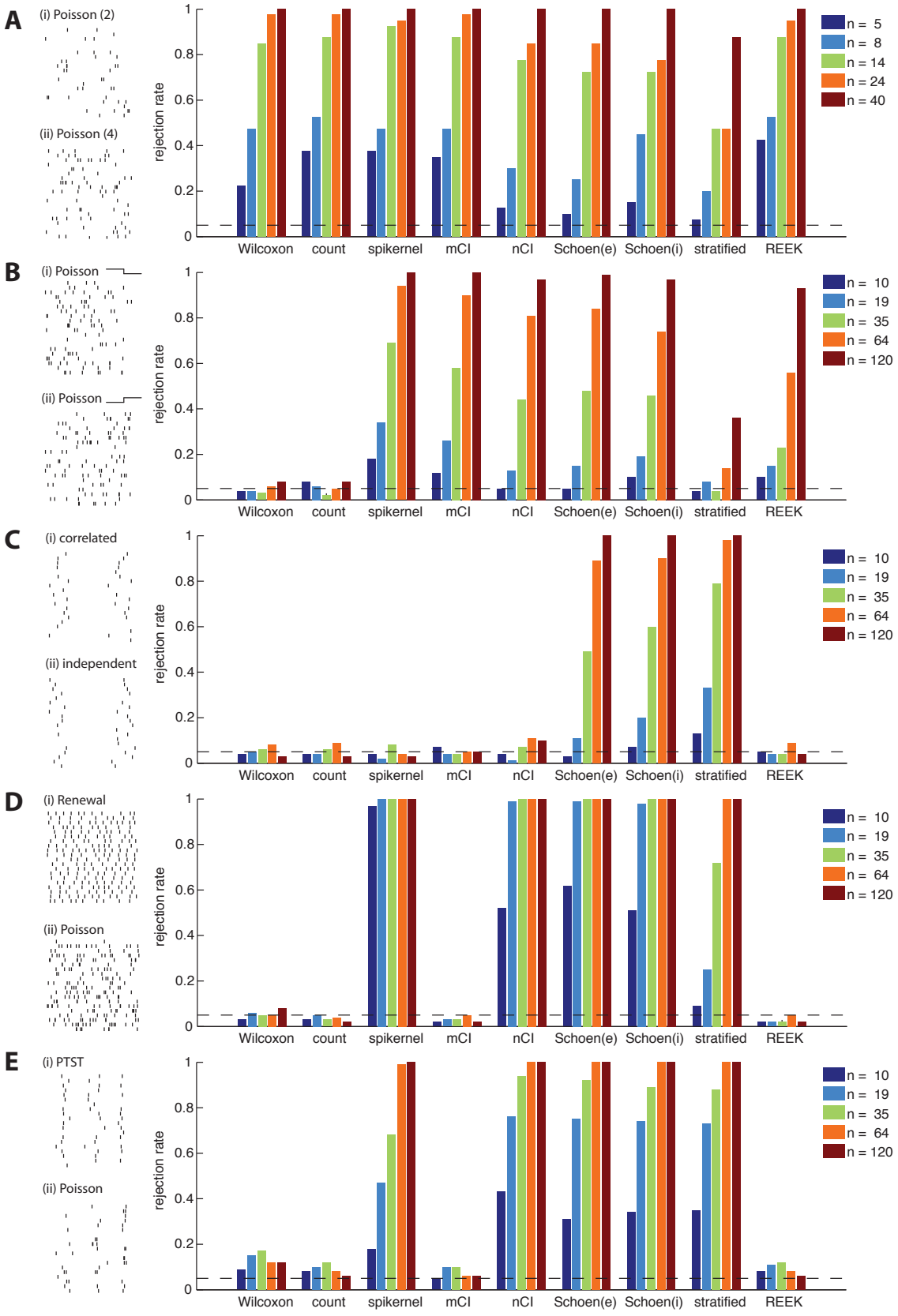
For the stratified kernel K_{strat}, we use a Gaussian kernel for each dimension. For each dimension we scale the kernel size such that it matches the empirical mean l_2 distance of a homogeneous Poisson process conditioned on the total number of spikes. The mean l_2 distance of a homogeneous Poisson process in each strata is a slowly growing function of dimension, where the rate of growth quickly reduces. The performance of such scaling is similar to using a single kernel size for all dimensions and does not change the conclusions.

5.2.1 Homogeneous Poisson processes

The most common feature is the mean count, or homogeneous mean firing rate. We test the discriminability on homogeneous Poisson processes with rate 2 and 4 spikes within the interval where the randomness is maximizes given the mean rate [Rényi, 1964]. In Figure 4A, Wilcoxon test, and dissimilarities induced by K_{count}, $K_{\text{spikernel}}$, K_{mCI}, and K_{reef} performs well in this task, rejecting more than 90% with just 24 trials. Note that spd kernels, especially stratified kernel performs poorly compared to Wilcoxon for this task. However, as we have discussed before, features other than count distribution are not captured by Wilcoxon test nor divergence induced by the count kernel. Therefore, for the subsequent experiments, we intentionally keep the mean firing rates of the two point processes constant, in order to emphasize that the spd kernels are indeed sensitive to other statistical properties of the point processes.

[3] `http://people.kyb.tuebingen.mpg.de/spider`

Figure 4 *(following page)*: Performance of each kernel quantified by rejection rate of hypothesis testing. There are 5 artificially generated point processes A to E (see section 5.2 for details). Raster plots on the left shows 20 realizations from each point process. For each pair of point processes, the bar plots show the rejection rates as a function of number of samples for each kernel. n is the number of spike trains in each class.


A
(i) Poisson (2)
(ii) Poisson (4)
B
(i) Poisson
(ii) Poisson
C
(i) correlated
(ii) independent
D
(i) Renewal
(ii) Poisson
E
(i) PTST
(ii) Poisson
rejection rate
0
0.2
0.4
0.6
0.8
1
Wilcoxon
count
spikernel
mCI
nCI
Schoen(e)
Schoen(i)
stratified
REEK
n = 5
n = 8
n = 14
n = 24
n = 40
n = 10
n = 19
n = 35
n = 64
n = 120

5.2.2 Inhomogeneous Poisson process

As firing rate modulation over time is an abundantly observed phenomenon in neural code experiments, we test the relative performance of each measure. We use two inhomogeneous Poisson processes with same mean rate (5 spikes in the interval) but with different rate profile. The interval is divided into half, and a step change in the intensity function was introduced such that the first half and second half are constant rates. The rates were chosen to be 2 and 3 spikes per half interval for the two halves respectively for one process, and 3 and 2 for the other process.

Due to the high dispersion of count distribution that is identical for both processes, K_{strat} poorly performs but does not completely fail (Figure 4B). K_{mCI} is designed to be sensitive to the (smoothed) mean rate, hence performs well as expected. $K_{\text{spikernel}}$, K_{nCI}, K^e_{Sch}, K^i_{Sch} performs on par with K_{mCI}.

5.2.3 Two spikes with correlation

Neurons can integrate and hence be history dependent on firing, or can have a short membrane time constant and forget the past spiking. We explore a simple and instructive example to show if the divergences can detect such differences in temporal correlations. We investigate two point processes having at most two spikes each, where the processes share the same marginal intensity function but have different correlation structures [Seth et al., 2010a]. In the first process, the two event timings are correlated i.e. the interspike interval (ISI) has a narrow distribution, whereas in the second process the two event timings are independent i.e. each action potential has its own precise timing distribution. Both point processes have a lossy noise i.e. each action potential has a probability of missing with probability $p = 0.1$. The temporal jitter of each spike is 100 ms and the interval is 300 ms. Since the highest dimension of the problem is 2, this problem is easier for the stratified kernel (Figure 4C). Nevertheless, all spd kernel induced divergences quickly catch up as the number of sample increases while pd kernels are unable to discriminate.

5.2.4 Renewal vs Poisson process

In neural systems, spike timing features that deviate from Poisson process, such as refractory period, oscillation and bursting, are often modeled with renewal process [Nawrot et al., 2008]. In this experiment, we compare a stationary renewal process with gamma interval distribution and an equi-rate homogeneous Poisson process (Figure 4D). We observe that $K_{\text{spikernel}}$ performs the best, and K_{nCI}, perform on par with the spd kernels. However, recall that this performance of $K_{\text{spikernel}}$ comes at the price of more computational cost.

5.2.5 Precisely timed spike trains

When the same stimulation is presented to a neuronal system, the observed spike trains sometimes show a highly repeatable spatio-temporal pattern at the millisecond time scale. Recently these precisely timed spike trains (PTST) are abundantly reported both *in vivo* and *in vitro* preparations [Reinagel and Reid, 2002, DeWeese et al., 2003, Johansson and Birznieks,

2004]. Despite being highly reproducible, different forms of trial-to-trial variability have also been observed [Tiesinga et al., 2008].

We model a precisely timed spike train in an interval by L pairs of probability density and probability pairs $\{(f_i(t), p_i)\}_{i=1}^{L}$. Each $f_i(t)$ corresponds to the temporal jitter, and p_i corresponds to the probability of generating (otherwise missing) the spike. Each realization of the PTST model produces at most L spikes i.e. L determines the maximum dimension for the PTST. The equi-intensity Poisson process has the rate function $\lambda(t) = \sum_i p_i f_i(t)$. We test if the kernels can differentiate between the PTST and the equi-intensity Poisson process for $L = 3$. We choose $f_i(t)$ to be equal variance Gaussian distributions on a grid sampled from a uniform random variable, and set $p_i = 0.9$.

We observe that the spd kernels again succeed in correctly rejecting the null hypothesis (Figure 4E). However, several pd kernels, $K_{\text{spikernel}}$ and K_{nCI} also performs on par, while the other pd kernels fail similarily to the renewal process experiment. The success of these pd kernels can be explained in terms their rather extensive construction than the simpler kernels such as K_{mCI} that are linear in nature.

6 Real data

6.1 Retinal spiking response

To test the practical effectiveness of knerel induced point process divergences, we applied them to publicly available sensory response data[4]. We used the rat retinal ganglion cell recordings from Nirenberg's lab [Jacobs et al., 2006]. There were 22 stimulus conditions repeated 30 times while recording 3.1 seconds long spike trains from 15 neurons. The task is to discriminate different stimulus conditions given the spike trains from a single neuron.

We compute the rejection rate of the result of hypothesis testings. If an experimenter were to select neurons that were sensitive to differences in the stimulus conditions, the rejection rate can be interpreted as the probability that he or she would categorize the neuron as being sensitive correctly (given a fixed false positive rate of 10%). The rejection rate of the null hypothesis is shown in Figure 5. We organized the stimulus conditions in two ways: we randomly paired the 22 stimulus conditions to form 11 hypothesis tests, or for each neuron, we paired the stimulus conditions to have close median firing rate response by sorting the conditions and pairing neighbors for the test. The latter pairing is to emphasize detectability for features other than mean count.

For spikernel, we spanned 24 parameters (same as simulation section, but allowing bin size of 50 ms as well as 150 ms). Despite being most optimized, $K_{\text{spikernel}}$ was the slowest kernel, and yet did not perform as well as K_{mCI}. The difference between spikernel and mCI, nCI, Schoenberg kernels was more pronounced for the neighboring rate pair condition. Overall K^e_{Sch} and K_{nCI} were the best, getting close to rejecting all responses due to different stimuli. Stratified kernel performed the worst due to large number of spikes (on average 20 to 50 spikes per trial).

[4] Available from `http://neurodatabase.org` [Gardner, 2004].

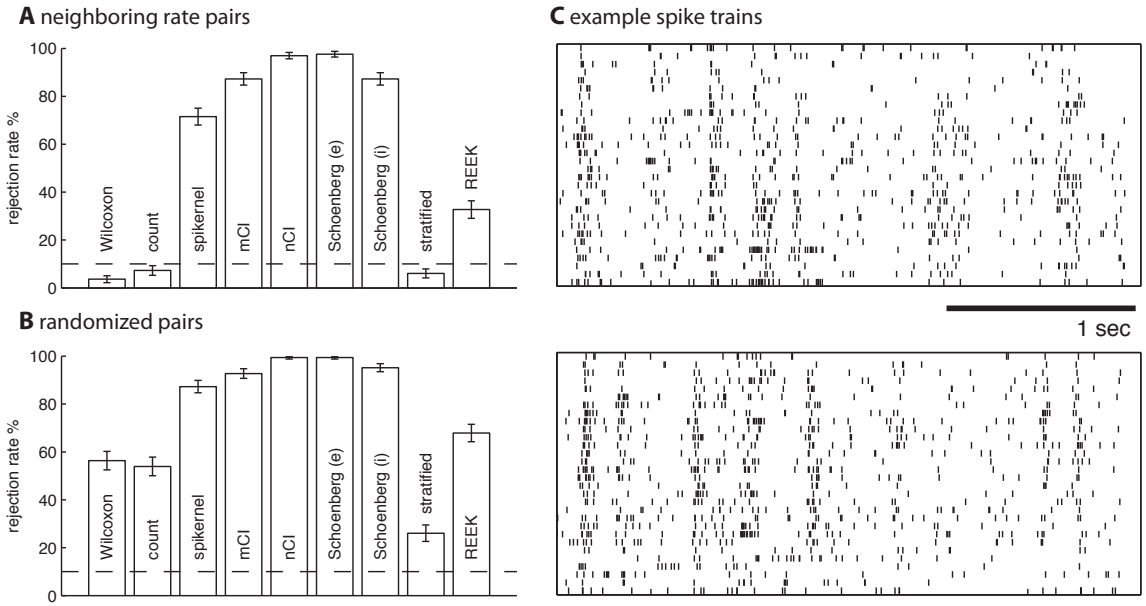


Figure 5: Performance on spike trains recorded from retinal ganglion neurons. (A) Rejection rate of each divergence measure for neighboring rate conditions. Dashed line is the test size ($p = 0.1$). Error bar shows standard error over 165 tests. (B) Rejection rate for randomized stimulus conditions. (C) 30 spike trains from two neighboring stimulus conditions for cell 113. Mean number of spikes per trial was 30. Only Schoenberg (e) and nCI rejected ($p < 0.05$). mCI was the next closest to rejecting ($p = 0.11$).

6.2 Directionality detection task in monkey MT

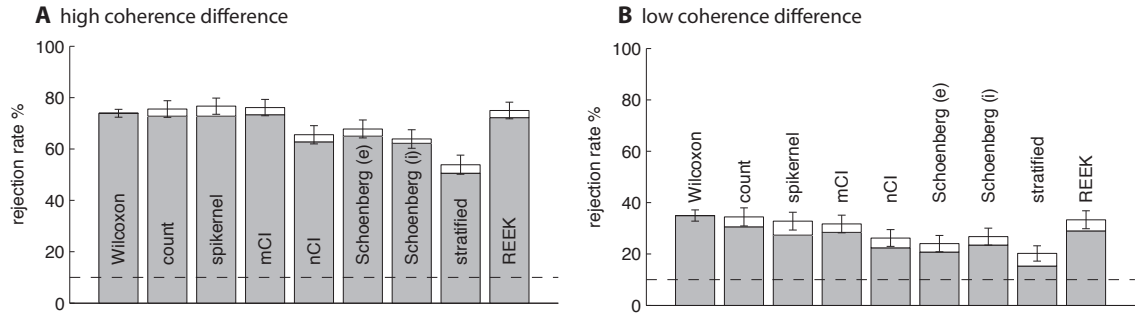


Figure 6: Distinguishing the preferred direction from the null direction stimulus from single neuron responses in MT. (A) high coherence trials (B) low coherence trials. Gray area represents fraction of rejections that agrees with Wilcoxon rank sum test.

Area MT (middle temporal) of the visual system is known to have neurons tuned to directional visual flow. The classical theory of MT is that the mean firing rate of these neurons encode the strength of directional motion [Albright, 1984]. Detecting such a tuning without any underlying assumptions, and with small number of repeats is often crucial for the neuroscientists. We apply the proposed kernel based divergences to single MT neuron recordings

from a detection task performed by monkeys [Britten et al., 1996].

We extracted dataset from a freely available online repository[5] to obtain responses from 180 direction sensitive MT neurons [Britten et al., 1996]. We select 30 trials for high coherence level 12.8% and 25.6% (or low coherence 3.2% and 6.4% for a more difficult task) each for the preferred direction and null direction. Constant strength motion stimulus is presented for 2 seconds, and we select a window at the end of 2 seconds that corresponds to average 3 spikes over the selected set of trials (for both directions).

Consistent with the rate coding under constant stimuli, the hypothesis testing for both pairs of conditions suggest mean spike count to be the main feature for discrimination (Fig. 6). The high coherence task is easier to the monkey, as well as for the hypothesis test.

Interestingly the kernels that showed better performance for simulations and results from retina performed worse in this dataset, compared to Wilcoxon rank sum test, and count, spikernel, mCI, and REEF kernel. This dataset is collected from single electrode electrophysiology where the experimenter explorered the neurons in MT and selected direction selective cells [Britten et al., 1996]. We argue that this selection bias could have resulted in strong discriminability feature in the mean count, and therefore simpler tests performed better, while spd kernels are distracted by other features of the data.

7 Conclusion

In this paper, we have studied strictly positive definite kernels on the space of spike trains. A strictly positive definite kernel is more powerful than a positive definite kernel in the sense that it offers a unique functional representation of a probability law. Therefore, the former induces a divergence —a tool that can be applied in testing equality of distribution—while the latter only a dissimilarity. However, the performance of a kernel is not only governed by its mathematical properties, but also by its computational simplicity and choice of free parameters. We have performed an extensive set of simulations to empirically compare the performance and complexity of both the existing positive definite kernels and the proposed strictly definite kernels, qualitatively, through kernel PCA, and quantitatively, through hypothesis testing, on both synthetic as well as real data. These results reveal useful insight regarding the characteristics of these kernels, and also explore their practicality.

The simulation results reveal that some pd kernels often perform equally well compared to spd kernel. This observation is counter-intuitive but not surprising, since while spd kernels should asymptotically perform better in theory, this does not guarantee good performance for the small sample regime. In fact, it is natural that a pd kernel designed to capture the exact discriminating statistical feature to outperform a generic spd kernel. For example, if a neuronal system exhibits a rate coding scheme then it is sufficient and perhaps the best to use a kernel that is only sensitive to the rate of the neural signal. Therefore, when studying neural codes, the choice of kernel should depend on how information is encoded and what the trial-to-trial variability structure is. However, the assuming a neural code beforehand can lead to bias in the analysis, therefore when in doubt, it is advantageous to have a generalist spd kernel.

[5] Neural Signal Archive `http://www.neuralsignal.org/`

We have also observed that for the real datasets, the count kernel, which fails for all but one artificial datasets by design, often performs on par or better compared to the sophisticated kernels in many datasets including one shown in section 6.2. We believe that this is because the neurons were selected under the assumption that the mean firing rate, which is the statistic that the count kernel or Wilcoxon test is sensitive to, is modulated. It is possible that for neurons that modulate responses in different ways are overlooked due to the unavailability of state-of-the-art analysis tools. Therefore, the use of strictly positive definite kernels would potentially allow new discoveries through observing these seemingly irrelevant neuronal responses. We would like to conclude the paper by stating two partially novel aspects of this framework that we have not explicitly addressed in the paper.

1. An spd (or a pd) kernel induces a (pseudo-) metric, since the kernel can be treated as an inner product in a reproducing kernel Hilbert space [Schölkopf and Smola, 2002]. Therefore, in essence, this approach also provides families of distance measures. This attribute relates this paper to the work on spike train metrics (e.g. [Victor and Purpura, 1997, van Rossum, 2001]). However, both Victor and Purpura [1997] and van Rossum [2001] restrict themselves to designing distance measures between two spike train realizations of point processes whereas our work is more general since it introduces distance between two point processes (or collection of spike trains).

2. The use of strictly positive definite kernels is not merely restricted to the design of divergence, but the characteristic nature of such kernels allow them to embed a probability law—a point process in our case—as an element in the reproducing kernel Hilbert space [Smola et al., 2007]. This in turn allows probabilistic operations such as Bayesian inference to be performed in terms of functional operations without the need of estimating or approximating the probability law [Song et al., 2010]. Therefore, the content of the paper naturally links these advance machine learning tools to neuroscience.

We leave exploring these directions as future work. Code for spike train kernels and divergence tests are available at the author's website.

Acknowledgements

We are grateful to the referees for their very helpful comments, which improved the paper significantly. This work was supported by NSF grant ECCS-0856441, and DARPA grant N66001-10-C-2008.

References

T. D. Albright. Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, 52(6):1106–1130, December 1984.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer-Verlag, 1984.

K. H. Britten, W. T. Newsome, M. N. Shadlen, S. Celebrini, and J. A. Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience*, 13(01):87–100, 1996.

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 406–414. 2010.

M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369–378, February 2010.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, 1988.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA, USA, 2001.

M. R. DeWeese, M. Wehr, and A. M. Zador. Binary spiking in auditory cortex. *Journal of Neuroscience*, 23(21):7940–7949, August 2003.

C. Diks and V. Panchenko. Nonparametric tests for serial independence based on quadratic forms. *Statistica Sinica*, 17:81–98, 2007.

J. Eichhorn, A. Tolias, A. Zien, M. Kuss, C. E. Rasmussen, J. Weston, N. Logothetis, and B. Schölkopf. Prediction on spike data using kernel algorithms. In *Advances in Neural Information Processing Systems 16*, pages 1367–1374. MIT Press, 2004.

N. K. Fisher and A. Banerjee. A novel kernel for learning a neuron model from spike train data. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 595–603. 2010.

D. Gardner. Neurodatabase.org: networking the microelectrode. *Nature Neuroscience*, 7(5): 486–487, 2004.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, 2009.

C. Houghton. Studying spike trains using a van Rossum metric with a synapse-like filter. *Journal of Computational Neuroscience*, 26(1):149–155, February 2009.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148:574–591, October 1959.

A. Jacobs, N. Grzywacz, and S. Nirenberg. Decoding the parallel pathways of the retina. In *Society for Neurosceince*, page Program No. 47.10/G1, 2006.

R. S. Johansson and I. Birznieks. First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nature Neuroscience*, 7(2):170–177, February 2004.

K. Kang and S. Amari. Discrimination with spike times and ISI distributions. *Neural Computation*, 20(6):1411–1426, June 2008.

L. Li, I. M. Park, S. Seth, J. Choi, J. T. Francis, J. C. Sanchez, and J. C. Príncipe. An adaptive decoder from spike trains to micro-stimulationusing kernel least-mean-square (KLMS) algorithm. In *IEEE Machine learning for Signal Processing (MLSP)*, 2011.

J. A. McFadden. The entropy of a point process. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):988–994, Dec 1965.

R. Naud, F. Gerhard, S. Mensi, and W. Gerstner. Improved similarity measures for small sets of spike trains. *Neural Computation*, 23(12):3016–3069, September 2011.

M. P. P. Nawrot, C. Boucsein, V. R. Molina, A. Riehle, A. Aertsen, and S. Rotter. Measurement of variability dynamics in cortical spike trains. *Journal of Neuroscience Methods*, 169(2):374–390, April 2008.

A. R. C. Paiva, I. Park, and J. C. Príncipe. Innovating signal processing for spike train data. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Lyon, France*, 2007.

A. R. C. Paiva, I. Park, and J. C. Príncipe. A reproducing kernel Hilbert space framework for spike trains. *Neural Computation*, 21(2):424–449, Feburary 2009.

A. R. C. Paiva, I. Park, and J. C. Príncipe. A comparison of binless spike train measures. *Neural Computing & Applications*, 19:405–419, 2010.

I. Park and J. C. Príncipe. Quantification of inter-trial non-stationarity in spike trains from periodically stimulated neural cultures. In *IEEE International conference on acoustics, speech, and signal processing (ICASSP)*, pages 5442–5445, 2010. Special session on Multivariate Analysis of Brain Signals: Methods and Applications.

E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005.

D. H. Perkel, G. L. Gerstein, and G. P. Moore. Neuronal spike trains and stochastic point processes. I. The single spike train. *Biophysical Journal*, 7(4):391–418, 1967.

A. Pinkus. Strictly hermitian positive definite functions. *Journal d'Analyse Mathématique*, 94(1):293–318, December 2004.

M. Rao, S. Seth, J. Xu, Y. Chen, H. Tagare, and J. C. Príncipe. A test of independence based on a generalized correlation function. *Signal Processing*, 91(1):15–27, 2011.

P. Reinagel and R. C. Reid. Precise firing events are conserved across neurons. *Journal of Neuroscience*, 22(16):6837–6841, Aug 2002.

A. Rényi. On an extremal property of the Poisson process. *Annals of the Institute of Statistical Mathematics*, 16(1):129–133, December 1964.

R. Schaback and H. Wendland. *Multivariate approximation and applications*, chapter Characterization and construction of radial basis functions, pages 1–24. Cambridge University Press, 2001.

B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

B. Schölkopf. *Support vector learning*. PhD thesis, Berlin, Techn. Univ., München, Germany, 1997. Zugleich: Berlin, Techn. Univ., Diss., 1997.

B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

B. Schrauwen and J. Campenhout. Linking non-binned spike train kernels to several existing spike train metrics. *Neurocomputing*, 70(7-9):1247–1253, 2007.

S. Seth, I. Park, A. Brockmeier, M. Semework, J. Choi, J. Francis, and J. C. Príncipe. A novel family of non-parametric cumulative based divergences for point processes. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2119–2127. 2010a.

S. Seth, M. Rao, I. M. Park, and J. C. Príncipe. A unified framework for quadratic measures of independence. *IEEE Transactions on Signal Processing*, 59:3624–3635, 2010b.

L. Shpigelman, Y. Singer, R. Paz, and E. Vaadia. Spikernels: Predicting arm movements by embedding population spike rate patterns in inner-product spaces. *Neural Computation*, 17(3):671–690, March 2005.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT '07: Proceedings of the 18th international conference on Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer-Verlag.

L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of Hidden Markov models. In *International Conference on Machine Learning (ICML)*, 2010.

B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pages 1750–1758. 2009.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 9:773–780, 2010.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proc. of Annual Conference on Learning Theory*, 2008.

P. Tiesinga, J.-M. Fellous, and T. J. Sejnowski. Regulation of spike timing in visual cortical circuits. *Nature Reviews Neuroscience*, 9:97–107, Feb 2008.

M. C. W. van Rossum. A novel spike distance. *Neural Computation*, 13:751–763, 2001.

J. D. Victor and K. P. Purpura. Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164, 1997.

J. D. Victor. Binless strategies for estimation of information from neural data. *Physical Review E*, 66(5):051903, Nov 2002.