ARTICLE Communicated by Jonathan Victor

Improved Similarity Measures for Small Sets of Spike Trains

Richard Naud
richard.naud@epfl.ch
Felipe Gerhard
felipe.gerhard@epfl.ch
Skander Mensi
skander.mensi@epfl.ch
Wulfram Gerstner
wulfram.gerstner@epfl.ch
Brain Mind Institute and School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL, Switzerland

Multiple measures have been developed to quantify the similarity between two spike trains. These measures have been used for the quantification of the mismatch between neuron models and experiments as well as for the classification of neuronal responses in neuroprosthetic devices and electrophysiological experiments. Frequently only a few spike trains are available in each class. We derive analytical expressions for the small-sample bias present when comparing estimators of the time-dependent firing intensity. We then exploit analogies between the comparison of firing intensities and previously used spike train metrics and show that improved spike train measures can be successfully used for fitting neuron models to experimental data, for comparisons of spike trains, and classification of spike train data. In classification tasks, the improved similarity measures can increase the recovered information. We demonstrate that when similarity measures are used for fitting mathematical models, all previous methods systematically underestimate the noise. Finally, we show a striking implication of this deterministic bias by reevaluating the results of the single-neuron prediction challenge.

1 Introduction

In order to compare one spike train with another, many methods have been proposed, each relating to an underlying philosophy as to what feature of the spike train matters most. Comparing spike trains in terms of the time series of interspike intervals (Victor & Purpura, 1996, 1997; Quiroga, Kraskov, Kreuz, & Grassberger, 2002; Kreuz, Haas, Morelli, Abarbanel, & Politi, 2007) focuses on a local estimation of the firing rate. Comparing spike trains in terms of the presence of specific spike patterns (de Ruyter van Steveninck & Bialek, 1988; Lestienne, 1995; Victor & Purpura, 1996, 1997; Chi, Wu,

Haga, Hatsopoulos, & Margoliash, 2007) is important if spike patterns such as sequences of action potentials have a greater importance than single spikes (Abeles, 1991; Lisman, 1997; Izhikevich, Desai, Walcott, & Hoppensteadt, 2003; Eyherabide, Rokem, Herz, & Samengo, 2009). There are also many methods to compare spike trains in terms of spike timings. One can consider the minimum cost to transform one spike train into another by shifting, deleting or, adding spikes (Victor & Purpura, 1996, 1997), consider the similarities of spike trains in terms of how they would appear to a postsynaptic neuron (van Rossum, 2001; Houghton, 2009), or simply count the number of coincident spikes and consider the coincidence rate (Kistler, Gerstner, & Hemmen, 1997; Quiroga, Kraskov, et al., 2002; Hunter & Milton, 2003).

Spike train similarity measures were used for the classification of responses coming from different stimuli and thus to predict the stimulus that was presented to an animal (visual as in Optican & Richmond, 1987, and in Geisler, Albrecht, Salvi, & Saunders, 1991, or auditory as in Wang, Narayan, Grana, Shamir, & Sen, 2007). Decoding motor cortex activity with similarity measures can be used to predict the response of the animal such as the motion of its arm (Shpigelman, Singer, Paz, & Vaadia, 2005; Eichhorn et al., 2004), which demonstrates the potential role of spike train similarity measures for BMI and the construction of brain-controlled prosthesis. Quantifying the efficiency of the decoding can give an estimate of the amount of information recovered depending on the precision or the type of spike train metric (de Ruyter van Stevenick & Bialek, 1988; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996; Victor & Purpura, 1996, 1997; Reich, Victor, Knight, Ozaki, & Kaplan, 1997; Reich, Mechler, Purpura, & Victor, 2000; Reich, Mechler, & Victor, 2001; DiLorenzo & Victor, 2003). Spike train similarity measures have also been used to quantify how well the spike time predictions of a mathematical model match real data (Pillow, Paninski, Uzzell, Simoncelli, & Chichilnisky, 2005; Jolivet, Rauch, Lüscher, & Gerstner, 2006; Badel et al., 2008; Jolivet et al., 2008; Gerstner & Naud, 2009; Kobayashi, Tsubo, & Shinomoto, 2009). The study of neuronal noise and variability is yet another topic for which the problem of comparing spike trains arises (see Tiesinga, Fellous, & Sejnowski, 2008, for a review).

Neurons are said to have variable responses because different trials with the same stimulus give different spike trains. A small trial-to-trial variability is observed when the stimulus is directly injected into a neuron that receives no other input (Mainen & Sejnowski, 1995) and a much larger variability when one records a single neuron of the visual cortex during visual stimulation (Carandini, 2004). The source of variability in vivo is probably channel noise (Faisal, Selen, & Wolpert, 2008) in combination with cortical patterns impinging via a large number of synapses onto a single neuron (Calvin & Stevens, 1968; Arieli, Sterkin, Grinvald, & Aertsen, 1996; Tsodyks & Markram, 1997). Channel noise implies that spike generation is an intrinsically probabilistic process. One way to approach this problem

is to average the spike train similarity measures across several trials, as in Jolivet et al. (2006, 2008); the other is to look at the mismatch between the probability distributions of the different spiking processes.

Let us start with an example of the first approach, a pairwise comparison of spike trains using two standard measures (see Figure 1). We have a set of 50 spike trains that could come from experiments but have been generated here for simplicity by a rate-normalized inhomogeneous Poisson process defined by a firing rate that is a superposition of gaussians of different widths (see Figure 1a) . After each spike, further spiking was blocked during an absolute refractory period of 3 ms. We now generate a second set of 50 spike trains with the same point process (see Figure 1b) and generate a third set with the widths of the gaussians reduced by a factor of four (see Figure 1c). Since the processes in Figures 1a and 1b are identical, we might naıvely expect that the spike trains in these figures are more similar than the spike trains generated with a different process in Figure 1c. However, when we compare the spike trains using the well-known Victor-Purpura metric (Victor & Purpura, 1996) and the coincidence rate (Kistler et al., 1997) as a dissimilarity or similarity measure, respectively, we find that both the Victor-Purpura metric and the coincidence factor imply that the spike trains from dissimilar processes are more similar than spike trains from the same process. There is an easy explanation for this apparent contradiction: the average distance between spike trains depends not only on the difference between the distributions but also on the variability of each distribution. For example, suppose that 10 one-dimensional data points $\{x_1^{(1)}, \ldots, x_{10}^{(1)}\}$ are drawn from a gaussian distribution of width $\sigma_1 = 1$ and another 10 data points $\{x_1^{(2)}, \ldots, x_{10}^{(2)}\}$ from a gaussian distribution of arbitrary width σ_2. The average distance $\propto \sum_{i,j}^{10} \|x_i^{(1)} - x_j^{(2)}\|^2$ is smaller if $\sigma_2 < \sigma_1$ than for equal width $\sigma_1 = \sigma_2$ (see Figure 1d). Similarly, asking for the average coincidence count across spike trains to be maximum is not the same as asking for the probability distributions to match.

The question arises whether it could be possible to directly compare distributions of spike trains rather than pairwise distances. However, existing methods for estimating spike train probability distributions (parametric: Paiva, Park, & Príncipe, 2009b, or entropy based: Optican & Richmond, 1987; Strong, Köberle, de Ruyter van Steveninck, & Bialek, 1998, and Panzeri, Senatore, Montemurro, & Petersen, 2007) require the observation of a sufficiently high number of independent responses evoked by a given stimulus. To do this, one usually records the response of a given neuron to repeated presentations (trials) of the same stimulus. It is often impossible experimentally to accumulate enough trials to estimate these statistics adequately. When the neuron responses are recorded in patch-clamp, experimental drifts limit the total time of stable conditions to less than 1 hour, and the presence of long-range adaptation in single neurons imposes a lower limit on the trial-to-trial wait period of at least 10 seconds (La Camera

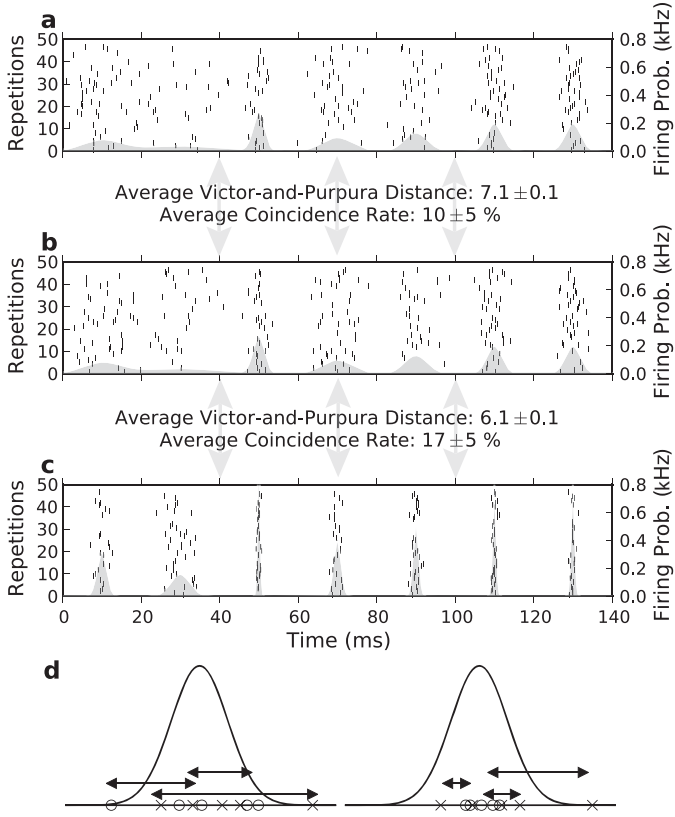

Figure 1: Similarity of spike trains does not predict similarity of distributions (a) Fifty spike trains (repetition number on vertical axis) containing an average of seven spikes (small vertical bars) per train were generated by a rate-normalized inhomogenous Poisson processes with a refractory period of 3 ms. The instantaneous rate of the point process is indicated by the shaded area and consists of seven gaussian bumps. (b) The same point process as in *a* was used to generate another set of 50 spike trains. (c) The jitter (i.e., the width of the gaussian bumps in gray) was reduced by a factor of 4 to generate a different set of spike trains. The average coincidence rate (Kistler et al., 1997) and the average Victor-Purpura metric (Victor & Purpura, 1996) between each spike train in *a* and *b* and between *b* and *c* are indicated in the gap between the subplots. Notice that the coincidence rate is greater between the dissimilar processes in *b* and *c* than between different instantiations of the same point process in *b* and *a*. Similarly, the larger averaged Victor-Purpura distance between *a* and *b* indicates that the process in *b* is more dissimilar from the process in *a* than from process in *c*. (d) The one-dimensional example is illustrated. The average point-wise distance between the crosses and the circles is larger if crosses and circles are drawn from the same distribution than if the circles are all clustered at the center.

et al., 2006; Lundstrom, Higgs, Spain, & Fairhall, 2008). Ignoring this lower limit will affect the independence of trials. This means that the similarity between the probabilistic spike processes often has to be calculated on a small number of samples where small-sample biases are known to occur.

Hence we arrive at the following dilemma: we do not have enough data to estimate probability distributions of spike trains, yet optimization based on naive pairwise comparisons yields undesirable results (see Figure 1). One could conceive several ways to address this dilemma. One possibility would be to work with pairwise measures but use a Kolmogorov-Smirnov test on the within-class and across-class distributions of pairwise distances. We briefly mention this method in section 6. A second option is to base comparisons not on the pairwise distances or the mismatch of the full probability distribution but only on the mean of the distributions, similar to some earlier approaches (Eggermont, Aertsen, & Johannesma, 1983; Gawne, McClurkin, Richmond, & Optican, 1991; Brown, Barbieri, Ventura, Kass, & Frank, 2002; Paninski, Pillow, & Simoncelli, 2005; Pillow et al., 2005, 2008). This is the option explored in the main body of this article. Since in the case of spike trains the mean of the distribution is the peristimulus time histograms (PSTHs), we study how to measure differences between the PSTHs. We derive bias-corrected measures of vectorial distance and correlation coefficient between PSTHs, and we estimate the variance of the distribution around the PSTH. A third option, which we study in depth, is to apply bias correction to previously used pair-wise distance measures applicable to small sets of spike trains.

The article is organized as follows. In order to establish the theoretical links between the different similarity measures, we review the binless vector space framework (Carnell & Richardson, 2005); in sections 2.1 and 2.2. A large set of seemingly distinct measures is encompassed by the vector space framework, from coincidence rate (Kistler et al., 1997) to correlation coefficient of peristimulus time histograms (Eggermont et al., 1983; Gawne et al., 1991; Brown et al., 2002; Paninski et al., 2005), and these relationships are reviewed in sections 2.6 and 2.7. In section 2.8 we discuss distance and angle measures between PSTHs. The analytical results on the deterministic bias in section 3.1 and small-sample bias in section 3.2 suggest new methods to compare sets of spike trains; these are set out in sections 3.3 and 3.4. In section 4, we use a set of test cases to show the validity of our analytical results and to show that these results generalize to other similarity measures that cannot be cast in the vector space framework (Eggermont et al., 1983; Victor & Purpura, 1996; van Rossum, 2001; Schreiber, Fellous, Whitmer, Tiesinga, & Sejnowski, 2003). We then show in section 4.7 that in classification tasks, the improved similarity measures can increase the recovered information. Finally we demonstrate in sections 4.5 and 4.6 that the improvements discussed in this article are crucial if the similarity measures are used for fitting or validating a stochastic neuron model.

2 The Space of Spike Trains

2.1 Spike Train Algebra. Similar to Carnell and Richardson (2005), we define the spike train vector as a weighted sum of Dirac-delta-pulses,

$$\mathbf{S}_i = \sum_{m=1}^{n_i} w_i^{(m)} \delta(t - \hat{t}_i^{(m)}), \tag{2.1}$$

with $0 < \hat{t} < T$ where $\hat{t}$ corresponds to a spike time, T to the total length of the spike train, and n_i to the number of spikes in spike train i. This mathematical object makes sense only when it has been integrated. For instance, the integral of such a spike train would be a mark accumulator process (Snyder, Miller, & Snyder, 1991). A spike train will have all weights $w_i^{(m)}$ equal to 1. The spike train vector is a mathematical generalization since it permits spikes to have fractional weights $w_i^{(m)}$. The mathematical framework also enables operations such as addition and multiplication by a scalar on the spike train vector. In general, we can write addition as

$$\mathbf{S}_i + \mathbf{S}_j = \sum_{m=1}^{n_i} w_i^{(m)} \delta(t - \hat{t}_i^{(m)}) + \sum_{p=1}^{n_j} w_j^{(p)} \delta(t - \hat{t}_j^{(p)}). \tag{2.2}$$

In the case where the spike $\hat{t}_i^{(m)}$ in spike train i is at the same time as a spike $\hat{t}_j^{(p)}$ in spike train j, then the sum of the spike train vectors has a spike with a weight $w_i^{(m)} + w_i^{(p)}$ at this time. We can also consider a spike train vector multiplied by a scalar,

$$a\mathbf{S}_i = \sum_{m=1}^{n_i} a w_i^{(m)} \delta(t - \hat{t}_i^{(m)}), \tag{2.3}$$

where $a \in \mathbb{R}$. Once addition and multiplication by a scalar are defined, it is easy to see that the closure axioms, the axioms for addition (commutativity, associativity, existence of zero element, existence of the negative), and the axioms for multiplication by a scalar (associativity, distributivity, and existence of identity) of a linear vector space are satisfied. Such a linear vector space made of functions of time is also called a function space. Figures 2A and 2B illustrate some of the properties of spike trains seen as vectors.

2.2 Scalar Products of Spike Trains. In order to define distances and angles between vectors, the function space must have an inner product. A linear space is said to have an inner (or scalar) product if for each vector pair $\mathbf{S}_i$ and $\mathbf{S}_j$, there exists a unique real number $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$ satisfying the following

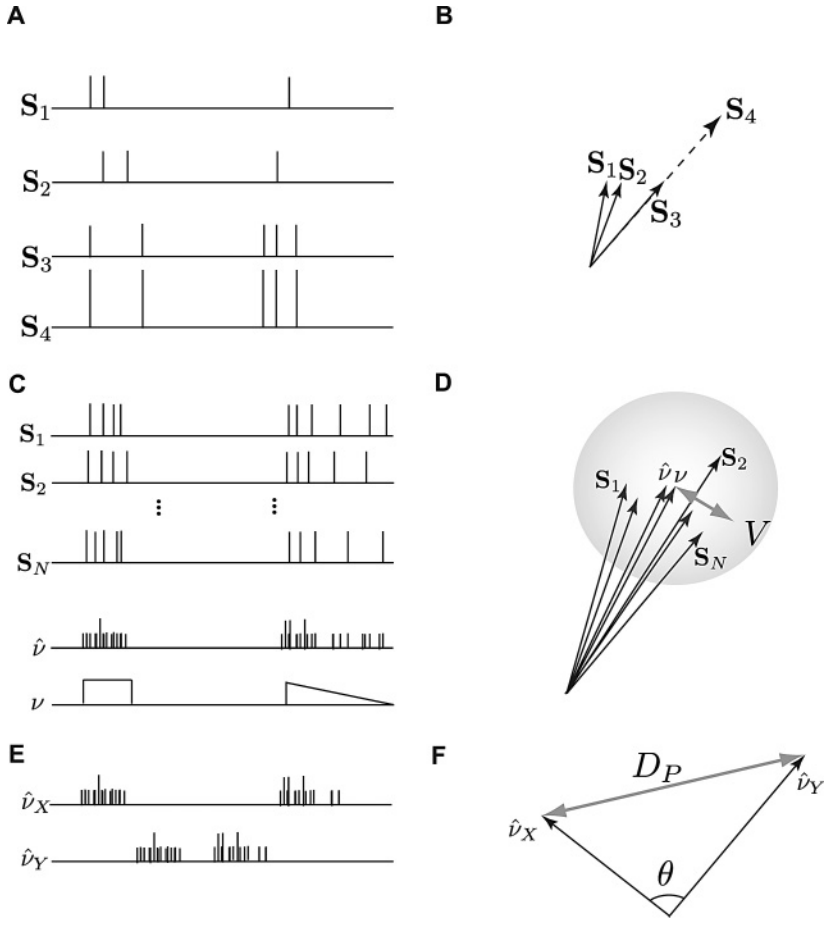

Figure 2: (A) Four spike trains with the spike train $\mathbf{S}_4$ having the same spike times as $\mathbf{S}_3$ but with twice the weight for each spike. (B) The abstract representation of the four spike trains in A as vectors. Vector $\mathbf{S}_3$ is slightly longer than $\mathbf{S}_1$ and $\mathbf{S}_2$ since the length is related to the spike count. Vector $\mathbf{S}_4$ is twice as long because each spike in the spike train (timing-wise equivalent to $\mathbf{S}_3$) has a weight of 2. (C) The population activity $\hat{\boldsymbol{\nu}}$, or PSTH, is the sum of all spikes divided by the total number N of spike trains. The population activity $\hat{\boldsymbol{\nu}}$ is an estimator of the instantaneous firing rate $\nu(t)$. (D) The abstract representation shows the vector ν as slightly different from its estimator $\hat{\boldsymbol{\nu}}$. The population activity vector lies in the center of the spike trains $\mathbf{S}_i$ seen as vectors. The mean squared distance from the population activity is the intrinsic variability V. (E) Example of two population activity vectors $\hat{\boldsymbol{\nu}}_X$ and $\hat{\boldsymbol{\nu}}_Y$ that show no overlap in spike timing. (F) The angle between two vectors of population activity is $\pi/2$ when no spikes are coincident at resolution Δ, such as in the examples of E.

axioms: commutativity, distributivity, associativity, and positivity. Typically there are multiple candidates of inner products satisfying the axioms. The inner products we consider here all have the general form

$$\langle \mathbf{S}_i, \mathbf{S}_j \rangle = \int_0^T \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_\Delta(s, s') S_i(t - s) S_j(t - s')\, ds\, ds' dt, \tag{2.4}$$

where K_Δ is a two-dimensional coincidence kernel with a scaling parameter Δ. K_Δ is required to be a nonnegative function with a global maximum at the origin. Moreover, $K_\Delta(s, s')$ should fall off rapidly so that $K_\Delta(s, s') \approx 0$ for all $|s|, |s'| > \Delta$. Typical examples of kernels include $K_\Delta(s, s') = h_1(s)h_2(s')$. For instance, $h_1(s) = h_2(s) = \frac{1}{\Delta} e^{-s/\Delta}$ for all $s > 0$ is the kernel used in van Rossum (2001). The scaling parameter should be small: $\Delta \ll T$. In the limit $\Delta \to 0$ under the constraint of a fixed integral (e.g., $\int h(s)ds = 6$ ms), the kernel approaches a Dirac-delta function. With $K_\Delta(s, s') = \delta(s)ds(s')$, we observe that $\langle \mathbf{S}_i, \mathbf{S}_i \rangle = \int_0^T S_i(t)dt = n_i$, where n_i is the number of spikes in $\mathbf{S}_i$. This inner product between spike trains is part of the reproducing kernel Hilbert space (RKHS) framework (Paiva et al., 2009b) and can be seen as a specific type of a spike pattern matching (SPM) kernel. We will see in section 2.6 how several standard similarity measures relate to this class of inner product.

For a comparison of spike trains seen as vectors, the notions of angular separation, distance, and norm of spike trains are particularly important. The squared norm of a spike train will be written as $\|\mathbf{S}_i\|^2 = \langle \mathbf{S}_i, \mathbf{S}_i \rangle$. Then the distance, D, between two spike train vectors is

$$D = \|\mathbf{S}_i - \mathbf{S}_j\|^2 = \langle \mathbf{S}_i - \mathbf{S}_j, \mathbf{S}_i - \mathbf{S}_j \rangle^2 = \|\mathbf{S}_i\|^2 + \|\mathbf{S}_j\|^2 - 2\langle \mathbf{S}_i, \mathbf{S}_j \rangle. \tag{2.5}$$

The cosine of the angle between $\mathbf{S}_i$ and $\mathbf{S}_j$ is

$$\cos \theta_{ij} = \frac{\langle \mathbf{S}_j, \mathbf{S}_i \rangle}{\|\mathbf{S}_i\| \, \|\mathbf{S}_j\|}. \tag{2.6}$$

Figures 2E and 2F illustrate the concepts of angle and distance for spike trains seen as vectors.

2.3 Ensemble of Stochastic Spike Trains. We consider spike trains arising from a stochastic process. Examples are theoretical processes (Poisson or other point processes) and data from neuronal recordings. From a mathematical point of view, we are interested in expected values of spike trains and functions thereof. From a practical point of view, we need to ask how well these mathematical properties can be estimated from a finite set of spike trains (e.g., derived from experimental recordings). We define the

firing intensity as the expectation over multiple measurements of a (potentially time-dependent) spiking process generating spikes with unit weight ($w = 1$):

$$\nu(t) = E\left[S(t)\right]. \tag{2.7}$$

Note that $\boldsymbol{\nu} \equiv \nu(t)$ belongs to the same function space as the spike trains $\mathbf{S}$ (see Figures 2C and 2D). Distance, norm, and angle between different firing intensities therefore have a straightforward definition. It is also possible to take expectations of inner products in the form of equation 2.4. Using equation 2.4, we can calculate the expected correlation of independent spike trains drawn from the same process (across several independent trials). Because of independence

$$E[\langle \mathbf{S}_i, \mathbf{S}_j \rangle_{i \neq j}] = \left\langle E[\mathbf{S}_i], E[\mathbf{S}_j] \right\rangle = \langle \boldsymbol{\nu}, \boldsymbol{\nu} \rangle = \|\boldsymbol{\nu}\|^2. \tag{2.8}$$

Therefore the expectation of the inner product of two independent spike trains belonging to the same probabilistic process is the squared norm of the firing intensity.

2.4 The PSTH as an Empirical Estimate. Let us now switch from a mathematical to a practical point of view. We consider a probabilistic time-dependent point process x from which we draw N_X independent and identically distributed (i.i.d.) spike trains, which we label $\mathbf{S}_i^{(x)}$ (where i runs from 1 to N_X). This process has a firing intensity $\nu_x(t)$ defined as a mathematical object in equation 2.7. In case of an inhomogeneous Poisson process, the rate $\nu_x(t)$ would completely define the point process, but we keep the treatment general and allow the point process to have arbitrary history dependence. Thus our treatment includes advanced spiking neuron models such as GLMs (Truccolo, Eden, Fellows, Donoghue, & Brown, 2005; Pillow et al., 2008) and experimental recordings.

The finite set of N_X spike trains is denoted by X. We suppose that the sample of N_X spike trains is the result of a measurement and that this sample contains all the information available about process x. It is natural to look at the averaged vector:

$$\hat{\boldsymbol{\nu}}_X = \frac{1}{N_X} \sum_{i=1}^{N_X} \mathbf{S}_i^{(x)}. \tag{2.9}$$

This resembles the equation for the center of mass in classical mechanics, but we will refer to $\hat{\boldsymbol{\nu}}$ as the population activity (Gerstner, 2000; Gerstner & Kistler, 2002) since multiple repetitions of the same stimulus can be seen as multiple identical, unconnected neurons receiving the same stimulus.

Other names include pooled train (Lindner, 2006) and superposition of point processes (Daley & Vere-Jones, 1988), although this last denomination often assumes sparsity of the individual spike trains, which we do not require here. The population activity becomes identical to a peristimulus time histogram (PSTH; Gerstein & Kiang, 1960) if we were to replace $\mathbf{S}$ by the vector made of the time series of binned spike counts. This procedure gives a discrete histogram. Alternatively, one could convolve the spike trains with a gaussian before or after averaging in order to get a smooth PSTH, also called spike density. In a stochastic framework, we can see $\hat{\boldsymbol{\nu}}_X$ as an estimator of the firing intensity since its expected value is $\boldsymbol{\nu}_x$.

In our approach for comparing ensembles of spike trains, the quantity $\hat{\boldsymbol{\nu}}_X$ has a central importance. We will derive similarity measures that quantify the match between $\hat{\boldsymbol{\nu}}$ of different ensembles. Moreover, we will discuss a quantity that relates to the variance around $\boldsymbol{\nu}_x$ seen as a central vector. Such ideas may sound deceptively similar to the central spike train discussed in the literature of spike train metrics. A central spike train, or prototype spike train, minimizes the sum of the squared distances to all the spike trains of the ensemble (Schoenberg & Tranbarger, 2008). The quantity $\hat{\boldsymbol{\nu}}_X$ is not a prototype spike train; it is actually not a spike train at all but rather the rescaled sum of multiple spike trains. Our approach is built on the theory of point processes where the firing intensity is of central importance, regardless of the process being Poisson. Our approach assumes that the spiking process is such that a time-dependent firing intensity is informative.

2.5 Intrinsic Variability and Reliability are Measures of the Variance. In order to link the notion of intrinsic reliability with the vector space picture, we first define $\hat{L}_X$ as the average of the norms of each spike train within the sample X:

$$\hat{L}_X = \frac{1}{N_X} \sum_{i=1}^{N_X} \|\mathbf{S}_i^{(x)}\|^2. \tag{2.10}$$

We note that $\hat{L}_X$ is related to the averaged spike count. L_X is exactly the averaged spike count if the inner product satisfies $\iint K_\Delta(s, s')\, ds ds' = 1$ and $K_\Delta(s, s') = 0$ whenever either s or s' is greater than the minimum interspike interval of any of the spike trains considered. The interpretation $L_X \sim$ *spike count* is helpful for the discussion in the remainder of this section.

Next, we consider the averaged squared deviation from the population activity or PSTH:

$$\hat{V}_X = \frac{1}{N_X - 1} \sum_{i=1}^{N_X} \left\| \mathbf{S}_i^{(x)} - \hat{\boldsymbol{\nu}}_X \right\|^2. \tag{2.11}$$

Croner, Purpura, and Kaplan (1993) used a similar measure to quantify the intrinsic variability in the responses to repeated representations of the same stimulus. We use the $N_X - 1$ normalization such that the sample variance is an unbiased estimator of the variance: $E[\hat{V}_X] = V_x$ for all N_X. Using equations 2.9 to 2.11, it is easy to show that the variance is proportional to the difference between the squared norm of the population activity and the averaged squared norms within the sample,

$$\hat{V}_X = \frac{N_X}{N_X - 1}(\hat{L}_X - \|\hat{\boldsymbol{\nu}}_X\|^2), \tag{2.12}$$

where $\hat{\nu}_X$ is the empirical population activity defined in equation 2.9. The sample variance $\hat{V}_X$ measures the amount of variability across the set of spike trains. To show this, we insert the definition of $\hat{\boldsymbol{\nu}}$ and separate from the double sum the elements containing the norm of the spike trains, which yields

$$\hat{V}_X = \hat{L}_X - \frac{2}{N_X(N_X - 1)} \sum_i \sum_{j<i} \langle \mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)} \rangle. \tag{2.13}$$

Under the assumptions indicated after equation 2.10, we can state that the sample variance $\hat{V}_X$ is the averaged spike count minus the average number of spikes that are coincident from one repetition to the next. One can thus think of $\hat{V}_X$ as the average number of spikes that are not coincident from one repetition to the next. We refer to $\hat{V}_X$ as the intrinsic variability.

If we take the expected value of the formula above, we can get an expression of the variance, using equations 2.10 and 2.13:

$$V_x = E\big[\langle \mathbf{S}_i^{(x)}, \mathbf{S}_i^{(x)} \rangle\big]_i - E\big[\langle \mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)} \rangle\big]_{i,j|i\neq j}. \tag{2.14}$$

We have used indices to the expected value in order to specify that the expectation is done in the first case on the ensemble of spike trains i, and in the second term, the expectation is on the same ensemble but without the terms corresponding to the inner product of a spike train to itself. Equation 2.14 will be important for the results of Section 3.2.

Since the variance quantifies the amount of variability within the set of spike trains, it may be convenient to have a number ranging between 0 and 1 (as in Schreiber et al., 2003) to quantify the reliability. We define $\hat{R}_X$, the estimator of the intrinsic reliability, as

$$\hat{R}_X = 1 - \frac{\hat{V}_X}{\hat{L}_X} = \frac{\frac{2}{(N_X-1)N_X} \sum_i \sum_{j<i} \langle \mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)} \rangle}{\frac{1}{N_X} \sum_i \|\mathbf{S}_i^{(x)}\|^2}. \tag{2.15}$$

We thus conclude that the intrinsic variability (see equation 2.13), the intrinsic reliability (see equation 2.15), and the variance of the spike trains around the population activity (or PSTH; see equation 2.11) all measure the same quantity but with different normalizations.

2.6 Similarity Measures Derived from an Inner Product. Schrauwen and Campenhout (2007) and Paiva et al. (2009b) have discussed the link between spike train vectors and standard measures such as those by Victor and Purpura (1996), van Rossum (2001), Hunter and Milton (2003), Schreiber et al. (2003). We briefly review these measures in order to link the formalism introduced in the previous section with known similarity measures. In the list of standard measures, we add the measure by Kistler et al. (1997) and the cross-correlation of spike densities (MacPherson & Aldridge, 1979; David & Gallant, 2005; Petersen et al., 2008). For brevity, we omit the measure of Houghton (2009) for which the link with vector spaces has already been discussed in Paiva, Park, and Príncipe (2010). We also omit the event synchronization of Quiroga, Kreuz, and Grassberger (2002), which is closely related to the coincidence factor of Kistler et al. (1997). Finally, we consider only spike time measures, since interspike interval, gamma-phase-based, template-based, or feature-based similarity measures pertain to different statistical quantities from instantaneous firing intensity (Victor & Purpura, 1996; Quiroga, Kreuz, et al., 2002; Tiesinga, 2004; Christen, Kohn, Ott, & Stoop, 2006; Chi et al., 2007; Kreuz et al., 2007; Kreuz, Chicharro, Andrzejak, Haas, & Abarbanel, 2009; Druckmann, Banitt, Gidon, & Schürmann, 2007; Naud, Marcille, Clopath, & Gerstner, 2008).

Table 1 shows the relationship between the space of spike trains and the coefficient of correlation between PSTHs (Eggermont et al., 1983; Gawne et al., 1991; Paninski et al., 2005), van Rossum's distance (van Rossum, 2001), and the coincidence factor (Kistler et al., 1997). The different measures correspond to different kernels K_Δ that define the inner product, equation 2.4. h_g, h_e, and h_r are gaussian, exponential, and rectangular coincidence kernels, respectively;

$$h_g(s;\Delta) = \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left(\frac{-s^2}{2\Delta^2}\right), \tag{2.16}$$

$$h_e(s;\Delta) = \Theta(s) e^{-s/\Delta}, \tag{2.17}$$

$$h_r(s;\Delta) = \Theta(s+\Delta)\Theta(\Delta - s). \tag{2.18}$$

where $\Theta(s)$ is the Heaviside function, with $\Theta(s) = 1$ for $s > 0$ and 0 otherwise.

We have included in Table 1 the generic angle and distance measures that we introduce in section 3. We also have included in Table 1 the measures loosely related to an inner product for completeness: the Victor and Purpura

Table 1: Summary of Spike Train Similarity Measures.

Name	Original Measure	$K_\Delta(s,s')$	References
Coefficient of correlation	$\rho_{XY} = \frac{\langle \hat{\nu}_X, \hat{\nu}_Y \rangle}{\|\hat{\nu}_X\| \, \|\hat{\nu}_Y\|}$	$h_g(s)h_g(s')$	Eggermont et al. (1983) Gawne et al. (1991) Paninski et al. (2005) David and Gallant (2005) Petersen et al. (2008)
Van Rossum's distance	$D_{\mathrm{VR}} = \|\mathbf{S}_i - \mathbf{S}_j\|^2$	$h_e(s)h_e(s')$	van Rossum (2001)
Coincidence factor	$\mathrm{CF} = \frac{\langle \mathbf{S}_i, \mathbf{S}_j \rangle - 2n_i n_j \Delta/T}{1/2\left(n_i+n_j\right)(1-2n_i\Delta/T)}$	$\delta(s)h_r(s')$	Kistler et al. (1997)
Victor-Purpura spike metric	$D_{\mathrm{spk}} = n_i + n_j - 2C(\mathbf{S}_i, \mathbf{S}_j)$	-	Victor and Purpura (1996)
Scaled Victor-Purpura	$\mathrm{VP} = \frac{2C(\mathbf{S}_i, \mathbf{S}_j)}{n_i+n_j}$	-	Kreiman, Krahe, Metzner, Koch, and Gabbiani (2000)
Coincidence factor without replacement	$\mathrm{CF2} = \frac{N_{coinc}(\mathbf{S}_i, \mathbf{S}_j) - 2n_i n_j \Delta/T}{1/2\left(n_i+n_j\right)(1-2n_i\Delta/T)}$	-	Gerstner and Naud (2009)
Hunter-Milton measure	$\mathrm{HM} = \frac{1}{n_i} \sum_k h_e(u_k(\mathbf{S}_i \to \mathbf{S}_j))$	-	Hunter and Milton (2003)
Angular separation	$M_a = \frac{\langle \hat{\nu}_X, \hat{\nu}_Y \rangle}{\|\hat{\nu}_X\| \, \|\hat{\nu}_Y\|}$	Unspecified	Section 2.8 Eggermont et al. (1983) Gawne et al. (1991) Paninski et al. (2005) David and Gallant (2005) Petersen et al. (2008)
Distance between population activities	$D_P = \|\hat{\nu}_X - \hat{\nu}_Y\|^2$	Unspecified	Section 2.8
Scaled distance	$M_D = \frac{2\langle \hat{\nu}_X, \hat{\nu}_Y \rangle}{\|\hat{\nu}_X\|^2 + \|\hat{\nu}_Y\|^2}$	Unspecified	Section 2.8

Notes: Here $h_g(s;\Delta) = \frac{1}{\sqrt{2\pi\Delta^2}} \exp(\frac{-s^2}{2\Delta^2})$ defines a gaussian filter while $h_e(s;\Delta) = \Theta(s)e^{-s/\Delta}$ is an exponential and $h_r(s;\Delta) = \Theta(s+\Delta)\Theta(\Delta - s)$ is a rectangular one. The kernel K_Δ defines the scalar product $\langle , \rangle$ as in equation 2.4.

metric (Victor & Purpura, 1996), the coincidence factor without replacement (Gerstner & Naud, 2009), and the Hunter and Milton measure (Hunter & Milton, 2003). The quantities $C(\mathbf{S}_i, \mathbf{S}_j)$, $N_{coinc}(\mathbf{S}_i, \mathbf{S}_j)$, and $\mathbf{u}(\mathbf{S}_i \to \mathbf{S}_j)$ are defined in the next section.

2.7 Similarity Measures Loosely Related to an Inner Product

2.7.1 Coincidence Factor Without Replacement. The coincidence factor CF may count two coincidences if there are two spikes in $\mathbf{S}_j$ that fall within $\pm\Delta$ of a spike in $\mathbf{S}_i$. An alternative method to count the number of coincident spikes is to find the maximum number of spikes in $\mathbf{S}_i$ coincident with a spike in $\mathbf{S}_j$ under the constraint that a spike in $\mathbf{S}_i$ can be coincident with at maximum one spike in $\mathbf{S}_j$ (this method was used in the International Neuroinformatics Coordinating Facility's Quantitative Single Neuron Modeling Competition in 2009; Gerstner & Naud, 2009). In analogy to common probability models, we call this procedure a coincidence measure without replacement. This counting method can lead to a slightly different number of coincident spikes, which we call N_{coinc}. The coincidence factor without replacement (CF2) is given by replacing $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$ with $N_{coinc}(\mathbf{S}_i, \mathbf{S}_j)$ in the equation of CF (see Table 1). This similarity measure cannot be seen as an inner product on spike train vectors because it does not satisfy all the axioms for an inner product enumerated in section 2.2 (distributivity and associativity are not fulfilled). However, if the minimum interspike interval is larger than 2Δ, then the difference between $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$ and $N_{\text{coinc}}(\mathbf{S}_i, \mathbf{S}_j)$ vanishes since a single spike of $\mathbf{S}_i$ can never coincide with two spikes of $\mathbf{S}_j$.

2.7.2 Victor and Purpura Metric. Victor and Purpura (1996) developed three similarity metrics that measure the minimum cost to transform a spike train $S_i(t)$ into another spike train $S_j(t)$. One of the three metrics, $D_{\text{spk}}(S_i, S_j)$, has been used extensively to study the importance of spike timing in real neurons (see Victor, 2005, for a review). To calculate D_{spk}, one uses a set of three transformations, each having an associated cost: (1) shift a spike with a cost q per unit time, (2) remove a spike with a cost of 1, or (3) add a spike with a cost of 1. Note that if a spike is shifted by an interval larger than $\Delta = 2/q$ in order to achieve a match, it is always better to remove the spike and recreate it at the new location. Hence, the shifting cost increases linearly until it reaches a value of 2 and stays constant thereafter. Defined this way, $D_{\text{spk}}(\mathbf{S}_i, \mathbf{S}_j)$ is greater than or equal to 0 and cannot exceed the total number of spikes in $\mathbf{S}_i$ and $\mathbf{S}_j$ together (Victor & Purpura, 1996). Similar to the situation discussed for CF2, D_{spk} cannot be mapped to an inner product on spike train vectors, but the distinction arises only if the minimum interspike interval is shorter than 2Δ.

If the minimum interspike interval is larger than 2Δ, there can be at most one spike in i within $\pm\Delta$ of a spike in j. Then we do not have to worry about the matching between the spike trains. In this case, we can count the number of coincidences with an inner product $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$ so that the Victor-Purpura metric is

$$D_{\text{spk}}(\mathbf{S}_i, \mathbf{S}_j) = \|\mathbf{S}_i\|^2 + \|\mathbf{S}_j\|^2 - 2\langle \mathbf{S}_i, \mathbf{S}_j \rangle, \tag{2.19}$$

$$K_\Delta(s, s') = h_t(s; \Delta)\delta(s'), \tag{2.20}$$

$$h_t(s; \Delta) = (1 - |t|/\Delta)\Theta(1 - |t|/\Delta), \tag{2.21}$$

where h_t is a triangular function. This function quantifies the advantage of shifting a spike by s instead of deleting and adding it at a different time. Here, $\Delta = 2/q$ corresponds to the amount of shifting for which it is as costly to remove and add a spike. Equation 2.19 tells us that instead of cumulating the cost to make spike train i identical to spike train j, we may start from the maximum cost $n_i + n_j = \|\mathbf{S}_i\|^2 + \|\mathbf{S}_j\|^2$ and reduce this according to the inner product defined by equations 2.20 and 2.21.

In the general case, the minimal interspike interval may be smaller than 2Δ. Instead of the scalar product $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$, we work directly with the quantity $C(\mathbf{S}_i, \mathbf{S}_j)$ that we derive from the Victor-Purpura metric by the definition

$$C(\mathbf{S}_i, \mathbf{S}_j) = \frac{1}{2}\left[n_i + n_j - D_{\text{spk}}(\mathbf{S}_i, \mathbf{S}_j)\right], \tag{2.22}$$

where D_{spk} is evaluated as in Victor and Purpura (1996). Finally, in order to compare D_{spk} with the same criteria and on the same scale as the previous similarity measures, we create a measure equivalent to D_{spk} but ranging from 0 (worst match) to 1 (best match) by normalizing with the total number of spikes (Kreiman et al., 2000):

$$\text{VP}(\mathbf{S}_i, \mathbf{S}_j) = \frac{n_i + n_j - D_{\text{spk}}(\mathbf{S}_i, \mathbf{S}_j)}{n_i + n_j} = \frac{C(\mathbf{S}_i, \mathbf{S}_j)}{\frac{1}{2}(n_i + n_j)}. \tag{2.23}$$

2.7.3 Hunter and Milton. Let $\mathbf{u}(\mathbf{S}_i \to \mathbf{S}_j)$ be a vector with elements, $u_k(\mathbf{S}_i \to \mathbf{S}_j) > 0$, made of the times between spike k in the spike train i and the nearest spike in spike train j (where $1 < k < n_i$). The average of each n_i component of $\mathbf{u}$ could be taken as a measure of how spike train i is different from spike train j. Instead of simply averaging the components of $\mathbf{u}(\mathbf{S}_i \to \mathbf{S}_j)$, one could also consider averages of nonlinear transformations such as h_e. This generalization leads to what is known as the Hunter and Milton measure (Hunter & Milton, 2003):

$$\mathrm{HM}(\mathbf{S}_i, \mathbf{S}_j) = \frac{1}{n_i} \sum_{k=0}^{n_i} h_e(u_k(\mathbf{S}_i \to \mathbf{S}_j); \Delta). \tag{2.24}$$

Chi and Margoliash (2001) have used the same measure but without the nonlinear function h_e. Here again associativity and distributivity are not fulfilled, and formally, we cannot see HM as an inner product on the function space as defined in section 2. HM can be loosely related to an inner product the same way as VP was loosely related to an inner product.[1] Moreover, HM can be defined as an inner product in the RKHS (Paiva et al., 2009b).

2.8 Comparing Sets of Spike Trains: The Distance Between Two PSTHs. When comparing sets of spike trains, we would like to know whether set X is distinct from set Y—more precisely, whether the spike trains in the set X were generated by a probabilistic process distinct from the one that generates the spike trains in set Y. For each set, we calculate the population activity vector, equation 2.9. From the vector space perspective, we define two criteria to distinguish set X from set Y based on their population activity only: the angular separation between the two population activity vectors of each set and the distance between the two population activity vectors. The angular separation between population activities is equivalent to the coefficient of correlation between PSTHs. Surprisingly, we will see that these criteria can be related to those matching the spike timings (angular separation) only or the spike times and the total number of spikes (distance).

2.8.1 Angular Separation Measures Temporal Coding Characteristics. Let the set X consist of N_X i.i.d. spike trains: $\{\mathbf{S}_1^{(x)}, \mathbf{S}_2^{(x)}, \ldots, \mathbf{S}_{N_X}^{(x)}\}$ with population activity $\hat{\boldsymbol{\nu}}_X$, and the set Y of N_Y i.i.d. spike trains labeled $\mathbf{S}_j^{(y)}$ with population activity $\hat{\boldsymbol{\nu}}_Y$. The first similarity measure is the angular separation between the population activities:

$$M_a(X, Y) = \cos\hat{\theta}(X, Y) = \frac{\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\|\hat{\boldsymbol{\nu}}_X\| \, \|\hat{\boldsymbol{\nu}}_Y\|}. \tag{2.25}$$

This quantity was extensively used as a similarity measure between populations of spike trains and is called the correlation coefficient, or Pearson coefficient, between two PSTHs (Eggermont et al., 1983; Gawne et al., 1991;

[1] To illustrate this point, replace h_e by h_t in equation 2.24 and consider spike trains as having no interspike intervals smaller than 2Δ. The measure reduces to HM $= \frac{1}{n_i} \iiint h_t(s; \Delta)\delta(s')S_i(t-s)S_j(t-s')\,dsds'dt$, which is the same inner product as for the VP distance (see equation 2.20).

Petersen et al., 2008; David & Gallant, 2005). The cosine goes from 0 (perpendicular population activities) to 1 (parallel population activities and hence perfect match). The spike train vectors cannot be antiparallel since the vectors—and, consequently, the population activities—are always nonnegative. The M_a measure (M stands for *match* and the index a for *angular*) is dependent on the time course of the firing intensity only, but not its amplitude. In particular, it is independent of the total firing rate. To see this, consider $\nu_x(t) = A_1 f_x(t)$, and $\nu_y(t) = A_2 f_y(t)$, with scalar parameter $A_1 > 0$ and $A_2 > 0$, for which we have $\frac{\langle \boldsymbol{\nu}_x, \boldsymbol{\nu}_y \rangle}{\|\boldsymbol{\nu}_x\| \, \|\boldsymbol{\nu}_y\|}$ independent of the respective amplitudes A_1 and A_2.

2.8.2 Distance Measures Both Temporal Coding and Spike Count Characteristics. The second similarity measure is the distance between the population activities:

$$D_P(X, Y) = \|\hat{\boldsymbol{\nu}}_X - \hat{\boldsymbol{\nu}}_Y\|^2 = \|\hat{\boldsymbol{\nu}}_X\|^2 + \|\hat{\boldsymbol{\nu}}_Y\|^2 - 2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle. \tag{2.26}$$

Here both the shape and the amplitude of the firing intensity have to match for the distance to be 0. For comparison, let us scale and offset this measure so that (similar to the angular match M_a in equation 2.25) 0 indicates the worst match and 1 the best match. The distance-based match M_D is defined as

$$M_D(X, Y) = 1 - \frac{D_P(X, Y)}{D_{\max}} = \frac{2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\|\hat{\boldsymbol{\nu}}_X\|^2 + \|\hat{\boldsymbol{\nu}}_Y\|^2}, \tag{2.27}$$

where $D_{\max} = \|\hat{\boldsymbol{\nu}}_X\|^2 + \|\hat{\boldsymbol{\nu}}_Y\|^2$ is the maximum distance obtained for population activities with no overlap.

3 Analytical Results

We start by considering in section 3.1 pairwise comparisons of spike trains and show that there is a systematic bias toward noise-free models. We then turn to a comparison of the means—the PSTHs. We show in section 3.2 that for small sample sizes, the distance between two PSTHs is biased but that this bias can be removed.

3.1 Train by Train Comparison Has a Deterministic Bias. We start with the example from Figure 1—a direct comparison of spike trains. To do so, we average the pairwise distance or similarity by comparing a spike train from set X to one of set Y (as was done with the coincidence factor in Jolivet et al., 2008). We will demonstrate that such a pairwise comparison implies a big bias toward the deterministic process. To illustrate this aspect, let us consider the distance D_{VR} (see Table 1) between spike train vectors. In view

of the one-dimensional example (see Figure 1D), it is not surprising that this method wrongly measures the distance between the distributions of spike trains, since we are trying to estimate the distance between distributions from element-by-element distances. First, we note that the expected value of the element-by-element average of the distance is equal to the expected value of the distance between two arbitrary spike trains—one chosen from the ensemble X and the other from the ensemble Y:

$$E\left[\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\|\mathbf{S}_i^{(x)}-\mathbf{S}_j^{(y)}\|^2\right]=E[\|\mathbf{S}^{(x)}-\mathbf{S}^{(y)}\|^2]. \tag{3.1}$$

Using the results of the previous section (in particular, equations 2.7, 2.8, and 2.14), we can derive an expression of the expected value as a function of the underlying firing intensities. This gives

$$E\left[\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\|\mathbf{S}_i^{(x)}-\mathbf{S}_j^{(y)}\|^2\right]=\|\boldsymbol{\nu}_x-\boldsymbol{\nu}_y\|^2+V_x+V_y. \tag{3.2}$$

In other words, the expected value of the element-by-element average of distances between spike trains is not the distance between the two firing intensities. The bias is proportional to the sum of the variability intrinsic to x and y. Moreover, we note that the bias does not decrease as we increase the number of spike trains in the sample since it is independent of N_X and N_Y.

Now we come to an important deduction from equation 3.2. When the similarity measure is used for fitting a neuron model to experimental data, the variance terms $V_x + V_y$ in equation 3.2 lead to what can be called a deterministic bias. Suppose that the distance in equation 3.2 is to be minimized by tuning the model, say y, to best fit the data, say x. Equation 3.2 tells us that the best model may not have its firing characteristics equal to those of the neuron ($\|\boldsymbol{\nu}_x-\boldsymbol{\nu}_y\|^2=0$ and $V_x=V_y$) because it can be more advantageous to set the variance V_y of the model to 0 at the expense of a minor mismatch $\|\boldsymbol{\nu}_x-\boldsymbol{\nu}_y\|^2>0$. Therefore, the best match will be achieved with a process of less variability than the data, that is, a more deterministic model. Again, we stress that this deterministic bias does not arise from the spike train, only from the fact that we are trying to assess the distance between probability distributions by looking at the averaged distance between its elements. A deterministic bias can arise when comparing a model with data or two data sets, and such a deterministic bias is not N-dependent. We explore this in detail with simulations and fits to real data in sections 4 and 5.

3.2 Sample Bias in the Norm of the Population Activity. The results from section 3.1, in particular, equation 3.2, suggest that we should avoid averaged pairwise comparisons. It might be much better to compare directly the distances $\|\hat{\boldsymbol{\nu}}_X - \hat{\boldsymbol{\nu}}_Y\|^2$ of the mean rate of the set of spike trains. Let us again recall the one-dimensional example of section 1. A set of $N = 10$ one-dimensional data points $\{x_1, \ldots, x_{10}\}$ will have a sample mean $\overline{x} = \frac{1}{N}\sum_{i=1}^{10} x_i$. It is well known that even if the sample mean is an unbiased estimator of the mean, the sample mean squared would be a biased estimator of the squared mean:

$$E[\overline{x}^2] - E[x]^2 = E[(\overline{x} - E[x])^2] = \frac{\sigma_x^2}{N}, \tag{3.3}$$

where σ_x^2 is the variance of the sample: $\sigma_x^2 = E\left[(x - E[x])^2\right]$. Only when the sample is large ($N \to \infty$) does the bias between $E[\overline{x}^2]$ and $E[\overline{x}]^2$ vanish. We refer to this effect in the following as a small-sample bias or, for short, a sample bias.

The one-dimensional example shows that the small-sample bias will appear in comparisons of spike trains whenever the similarity measure contains the norm of the (empirical) population activity. In the following, we expand on this simple instance of the small-sample bias and apply it to the space of spike trains.

We saw in equation 2.8 that the expected value of the inner product $\langle \mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)} \rangle$ between two independent spike trains generated by the same point process is $\|\boldsymbol{\nu}_x\|^2$. The norm of the population activity $\|\hat{\boldsymbol{\nu}}_X\|^2$ is made principally of inner products between independent spike trains (off-diagonal terms) but also contains inner products of a spike train with itself (diagonal terms). To see this, we square the definition of the population activity in equation 2.9:

$$\|\hat{\boldsymbol{\nu}}_X\|^2 = \left\langle \frac{1}{N_X}\sum_i \mathbf{S}_i, \frac{1}{N_X}\sum_j \mathbf{S}_j \right\rangle = \frac{2}{N_X^2}\sum_i \sum_{j>i} \langle \mathbf{S}_i, \mathbf{S}_j \rangle + \frac{1}{N_X^2}\sum_i \langle \mathbf{S}_i, \mathbf{S}_i \rangle. \tag{3.4}$$

One expects that these diagonal terms (the last sum in equation 3.4) will make a vanishingly small contribution when $N \to \infty$, but when N is small, they are not negligible. Therefore, the norm of the population activity is biased if it is used as an estimator of the true $\|\boldsymbol{\nu}_x\|^2$, and this bias propagates to all the similarity measures discussed so far (M_a, D_P, M_D). Taking the expectation of equation 3.4, we can now see from a comparison with equation 2.14 that using $\|\hat{\boldsymbol{\nu}}_X\|^2$ as the estimator for $\|\boldsymbol{\nu}_x\|^2$ is biased by

$$E\left[\|\hat{\boldsymbol{v}}_X\|^2\right] - \|\boldsymbol{v}_x\|^2 = \frac{V_x}{N_X}, \tag{3.5}$$

in complete analogy to equation 3.3. But V_x can be estimated to order $(1/N_X)^2$ from $\hat{V}_X$ as we can see from equations 2.13 and 2.14. Using $\hat{V}_X$ and N_X, we can therefore build an unbiased estimator simply by subtracting the bias from the biased estimator. Let us define the off-diagonal terms:

$$\hat{C}^*_{XX} = \frac{2}{(N_X - 1)N_X} \sum_i \sum_{j<i} \langle \mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)} \rangle. \tag{3.6}$$

It is clear from equation 2.8 that $E[\hat{C}^*_{XX}] = \|\boldsymbol{v}_x\|^2$. Hence, $\hat{C}^*_{XX}$ is an unbiased estimator of the norm of the population activity. The C here stands for "coincidences" as $\hat{C}^*_{XX}$ counts the coincidences across repetitions, that is, across several spike trains from the ensemble X. Across multiple tests with sample size N, this new estimator has a variance $\mathrm{Var}[\hat{C}^*_{XX}] = \frac{2}{N(N-1)}\mathrm{Var}[\langle \mathbf{S}_i, \mathbf{S}_j \rangle_{i \neq j}]$ and is a member of a general class of unbiased estimators with minimum variance (Voinov & Nikulin, 1996).[2] Equation 3.6 is reminiscent of the equation of intrinsic reliability R_X; indeed, we can write

$$\hat{C}^*_{XX} = \hat{R}_X \hat{L}_X. \tag{3.7}$$

Informally, this means that number of coincidences across repetitions $\hat{C}^*_{XX}$ equals the intrinsic reliability times the number of spikes. In terms of the intrinsic variability estimator $\hat{V}_X$, the number of coincidences across repetitions is the total number of spikes minus the intrinsic variability: $\hat{C}^*_{XX} = \hat{L}_X - \hat{V}_X$.

3.3 Sample Bias in Distance and Angle Measures. The sample bias causes an overestimation of the distance measure D_P (defined in equation 2.26) proportional to the intrinsic variability V_x in each process. Again the analogy to the set of one-dimensional data points discussed in section 1 is useful. The squared distance between the sample means $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and $\overline{y} = \frac{1}{N}\sum_{i=1}^N y_i$ is given by $(\overline{x} - \overline{y})^2 = \overline{x}^2 + \overline{y}^2 - 2\overline{x}\,\overline{y}$, which contains the biased quantities $\overline{x}^2$ and $\overline{y}^2$ (see section 3.2). In the case of spike trains, the bias-corrected estimator C^*_{XX} can be used to form an unbiased estimator of the distance between two population activities:

[2] We can see why the coefficient $\frac{2}{(N_X-1)N_X}$ replaces the coefficient $\frac{2}{N_X^2}$ in equation 3.4. The factor $\frac{2}{N_X^2}$ overestimates the number of terms in the corresponding sum by an amount that will become vanishingly small at large N_X: $\frac{2}{N_X^2} = \frac{2}{N_X(N_X-1)}(1 - \frac{1}{N_X})$.

$$D_P^*(X,Y) = \|\hat{\boldsymbol{\nu}}_X - \hat{\boldsymbol{\nu}}_Y\|^2 - \frac{\hat{V}_X}{N_X} - \frac{\hat{V}_Y}{N_Y} = \hat{C}_{XX}^* + \hat{C}_{YY}^* - 2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle, \tag{3.8}$$

such that $\mathrm{E}[D_P^*(X,Y)] = \|\boldsymbol{\nu}_x - \boldsymbol{\nu}_y\|^2$. The reader may want to compare the right-hand side of equation 2.26 with the right-hand side of equation 3.8. Note that the expectation $E[D_P^*(X,Y)]$ can never fall below 0 if X and Y are statistically independent point processes. Nevertheless, for small-sample size, $D_P^*(X,Y) < 0$ is possible so that D_P^* is not a metric.[3]

Because the sample bias causes an overestimate of $\|\hat{\boldsymbol{\nu}}_X\|^2$, this leads to an underestimation of the angular separation M_a and M_D. Replacing $\|\hat{\boldsymbol{\nu}}_X\|^2$ by $\hat{C}_{XX}^*$ should remove the sample bias present in M_a and M_D. We therefore consider instead of equation 2.25 a bias-corrected angular separation:

$$M_a^*(X,Y) = \frac{\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\sqrt{\hat{C}_{XX}^* \hat{C}_{YY}^*}}. \tag{3.9}$$

Similarly the distance measure of equation 2.27 is replaced by

$$M_D^*(X,Y) = \frac{2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\hat{C}_{XX}^* + \hat{C}_{YY}^*}. \tag{3.10}$$

Note that when comparing a set of model data X with experimental data Y, the instantaneous firing intensity of the model is often known exactly. In this case, the normalization should be done with $C_{XX}^* = \|\boldsymbol{\nu}_{model}\|^2$.

However, even if $\hat{C}_{XX}^*$ is an unbiased estimator of $\|\boldsymbol{\nu}_x\|^2$, it does not imply that $(\hat{C}_{XX}^*)^{-1}$ is an unbiased estimator of $\|\boldsymbol{\nu}_x\|^{-2}$ (Voinov, 1985). We expect that this additional bias will be sizable only when $\hat{C}_{XX}^*$ is close to 0 on the scale of its standard deviation. Before using the bias-corrected measures M_a^* and M_D^* described above, one should check if the magnitude of $\hat{C}^*$ is very close to 0. We have observed that this can happen when there are very few spikes or when the point process has low intrinsic reliability, for instance, a homogeneous gamma renewal process with random initiation and small firing rate. In this case, D_P^* should be preferred to M_a^* or M_D^*. If an angular separation is required, then one can use the method employed by David and Gallant (2005) to estimate the first-order bias of $\|\hat{\boldsymbol{\nu}}_X\|^{-2}$.

[3] $D_P^*(X,Y) < 0$ will arise when the number of coincidences across repetitions $\hat{C}_{XX}^*$ is smaller than the coincidence count between X and Y, $\hat{C}_{XX}^* + \hat{C}_{YY}^* < 2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle$. For instance, two sets of spike trains, each made of only two repetitions of a homogeneous Poisson process, could happen to have no coincidences across distinct repetitions ($\hat{C}_{XX}^* = \hat{C}_{YY}^* = 0$) but still coincidences between sets X and Y ($\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle \neq 0$).

The need to correct for the sample bias has been highlighted in the appendix of David and Gallant (2005)[4] and of Petersen et al. (2008).[5] In these studies, they arrive at empirical estimates of the bias that accurately approximate the bias-corrected angular separation of equation 3.9. In contrast to these earlier approaches, our analytical debiasing method presented here has minimal variance.

3.4 Applications to Measures Loosely Related to an Inner Product. In section 3.1, we saw that pairwise comparisons contain deterministic bias. Moreover, in section 3.2, we saw that working with squared quantities can give rise to a bias for small-sample sizes. With these insights in mind, we now return to classical spike train metrics. The derivations we presented in sections 3.1 and 3.2 rely heavily on the inner product and the quantity $\hat{\boldsymbol{\nu}}$, that is, the population activity defined as the empirical mean of the set of spike trains. We could go one step further and try to conciliate our approach with spike train metrics. This is motivated by our point-process approach and strays from the formalisms of spike train metrics (see the discussion in section 6). The similarity measures obtained above encompass all the spike train metrics that are related to an inner product, but we use the spike train metrics to subserve a comparison of the population activity. In spike train metrics loosely related to an inner product, there is no such thing as a population activity. In this section, we exploit the analogy with the inner product to compute a distance that should relate to the distance between the distributions rather than the distance between the individual spike trains.

To do this, we have to identify a quantity that is related to $\|\hat{\boldsymbol{\nu}_X}\|^2$ and replace this quantity by a term analogous to $\hat{C}^*_{XX}$ in equation 3.6. In this section we explain how this can be done for the Victor and Purpura metric (D_{spk}), the coincidence rate without replacement (CF2), and the Hunter and Milton measure (HM). The resulting measures are summarized in Table 2.

3.4.1 Victor and Purpura D_{spk}. Let us consider the element-by-element average of the Victor-Purpura metric D_{spk} defined in Victor and Purpura (1997). We now use the definition of $C(\mathbf{S}_i, \mathbf{S}_j)$ in equation 2.22 so as to

[4]David and Gallant (2005) remove the bias by first noting that $\|\hat{\boldsymbol{\nu}}\|^{-2} = \|\boldsymbol{\nu}\|^{-2} + \epsilon/N$, where ϵ scales the magnitude of the noise. By successively calculating $\|\hat{\boldsymbol{\nu}}\|^{-2}$ for subsets of the total sample of different sizes and plotting as a function of $1/N$, $\|\boldsymbol{\nu}\|^{-2}$ can be estimated by linear extrapolation to find the offset at $1/N = 0$. This method yields a good first-order bias correction that also removes bias due to the inverse.

[5]Using our notation, Petersen et al. (2008) estimate $\|\boldsymbol{\nu}\|^2$ by taking the inner product of separate halves of the total available samples. This is a clever and simple way to remove the sample bias, though the additional bias due to the inverse may still be important. We note that because of the separation of the total sample in two subsamples, this estimator has a variance that is roughly twice greater than $\text{Var}[\hat{C}^*]$, its variance would correspond to $\frac{4}{N^2}\text{Var}[\langle \mathbf{S}_i, \mathbf{S}_j \rangle_{i \neq j}]$.

Table 2: Summary of Similarity Measures for Comparing Groups of Spike Trains.

Name	Group Measure	C^*_{XX}	Section
Angular separation	$M^*_a = \frac{\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\sqrt{\hat{C}_{XX}\hat{C}_{YY}}}$	$\frac{2}{N(N-1)} \sum_{i,j>i} \langle \mathbf{S}_i, \mathbf{S}_j \rangle$	3.3
Distance	$D_P = \hat{C}^*_{XX} + \hat{C}^*_{YY} - 2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle$	$\frac{2}{N(N-1)} \sum_{i,j>i} \langle \mathbf{S}_i, \mathbf{S}_j \rangle$	3.3
Scaled distance	$M^*_D = \frac{2\langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle}{\hat{C}^*_{XX} + \hat{C}^*_{YY}}$	$\frac{2}{N(N-1)} \sum_{i,j>i} \langle \mathbf{S}_i, \mathbf{S}_j \rangle$	3.3
Victor-Purpura spike metric	$D^*_{\text{spk}} = \hat{C}^*_{XX} + \hat{C}^*_{YY} - 2\hat{C}_{XY}$	$\frac{2}{N(N-1)} \sum_{i,j>i} C(\mathbf{S}_i, \mathbf{S}_j)$	3.4
Scaled Victor-Purpura	$\text{VP}= \frac{2\hat{C}_{XY}}{\hat{C}^*_{XX} + \hat{C}^*_{YY}}$	$\frac{2}{N(N-1)} \sum_{i,j>i} C(\mathbf{S}_i, \mathbf{S}_j)$	3.4
Coincidence rate without replacement	$\text{CF2}^* \frac{\hat{C}_{XY}}{\frac{1}{2}(\hat{C}^*_{XX} + \hat{C}^*_{YY})}$	$\sum_{i,j>i} \frac{N_{\text{coinc}}(\mathbf{S}_i, \mathbf{S}_j) - 2n_i n_j \Delta / T}{N(N-1)/2}$	3.4
Hunter-Milton measure	$\text{HM}^* = \frac{\hat{C}_{XY}}{\frac{1}{2}(\hat{C}^*_{XX} + \hat{C}^*_{YY})}$	$\sum_{i,j>i} \frac{\text{HM}(\mathbf{S}_i \to \mathbf{S}_j) + \text{HM}(\mathbf{S}_j \to \mathbf{S}_i)}{N(N-1)}$	3.4

evaluate the average distance between N_X spike trains in set X and N_Y spike trains in set Y:

$$\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} D_{\text{spk}}(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)}) = \overline{n}_X + \overline{n}_Y - \frac{2}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} C(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)}). \tag{3.11}$$

Since there are N_X spike trains in ensemble X and $n_i^{(x)}$ is the number of spikes in trial i, $\overline{n}_X = \sum_{i=1}^{N_X} n_i^{(x)}/N_X$ is just the average number of spikes in a train from ensemble X. Similarly, $\overline{n}_Y$ is the average number of spikes per train across the spike trains in set Y. $C(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)})$ is a number discounting any coincidence from the maximum cost $n_i + n_j$ of transforming the spike train i into spike train j (see section 2.7). In equation 2.19, we replaced C by the inner product defined with the triangular function (which holds only when the interspike interval is greater than 2Δ). The analogy is instructive; if C corresponds to the inner product, then the average number of spikes in equation 3.11 should play a role equivalent to $\|\hat{\mathbf{v}}\|^2$ in equation 2.26.

Why and how should we replace the average number of spike per repetition in equation 3.11? We have argued that the similarity measure we are looking for should account for the similarity between statistical properties of the point processes. D_{spk} as defined above indicates a distance of 0 only when the spike trains (or the set of spike trains) are identical, spike per spike, through all the repetitions. Formulated in this way, D_{spk} has no reason to be 0 when we compare spike trains coming from the same stochastic point process. For instance, two sets of spike trains coming from the same inhomogeneous Poisson process will not have all their spikes coinciding. We can create a measure that will have expected distance 0 when comparing spike trains from the same point process by replacing the average number of spikes $\overline{n}_X$ or $\overline{n}_Y$ by the average coincidence count across distinct pairs of repetition within each set, $\hat{C}^*_{XX}$ for set X and $\hat{C}^*_{YY}$ for set Y:

$$\hat{C}^*_{XX} = \frac{2}{N_X(N_X-1)}\sum_{i=1}^{N_X}\sum_{j>i}^{N_X} C(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)}). \tag{3.12}$$

This leads to the new distance across sets of spike trains:

$$D^*_{\text{spk}}(X,Y) = \hat{C}^*_{XX} + \hat{C}^*_{YY} - 2\hat{C}_{XY}, \tag{3.13}$$

where $\hat{C}_{XY} \equiv \frac{1}{N_X}\frac{1}{N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} C(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)})$ is the average coincidence count between individual spike trains of set X and individual spike trains of set Y

(compare equation 3.13 with equation 3.8 to see that $\hat{C}_{XY} \sim \langle \hat{\boldsymbol{\nu}}_X, \hat{\boldsymbol{\nu}}_Y \rangle$). Then D^*_{spk} is 0 when the average coincidence count across X and Y is the same as the average coincidence count across distinct repetitions within set X and within set Y. This distance is maximum when the processes x and y are so dissimilar that there are never any coincidences between spike trains from X and spike trains from Y. Similar to D^*_P treated in section 3.2, D^*_{spk} cannot be seen as a metric since it is greater than or equal to 0 only in expectation.

In equation 2.23 we performed a normalization of $D_{\text{spk}}(\mathbf{S}_i, \mathbf{S}_j) \in [0, n_i + n_j]$ into a measure $\text{VP}(\mathbf{S}_i, \mathbf{S}_j) \in [0, 1]$. For a scaled distance between sets of spike trains analogous to the VP distance, we introduce the expression

$$\text{VP}^*(\text{X}, \text{Y}) = \frac{\hat{\text{C}}_{\text{XY}}}{\frac{1}{2}(\hat{\text{C}}^*_{\text{XX}} + \hat{\text{C}}^*_{\text{YY}})}. \tag{3.14}$$

This is very similar to M^*_D, but in equation 3.14, the quantity that we called average coincidence count ($\hat{C}^*_{XX}, \hat{C}^*_{YY}, \hat{C}_{XY}$) is based on element-by-element comparison rather than on the inner product on the population activity. Through numerical simulations, we will show in section 3.4 that VP* should be used instead of VP to properly distinguish between small sets of spike trains derived from different point processes.

3.4.2 Coincidence Factor Without Replacement. In section 2.7, we distinguished between two methods for counting the number of coincidences. In the first one, the number of coincidences can be counted using the inner product for CF in Table 1. In the second, we count the number of coincidence by allowing at most one spike from train i to be coincident with a spike in train j. We called this second quantity $N_{\text{coinc}}(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)})$. The analogy with the inner product is clear with $\langle \mathbf{S}_i, \mathbf{S}_j \rangle \sim N_{\text{coinc}}(\mathbf{S}_i, \mathbf{S}_j)$. Here, we keep the discounting term for the chance level (as in the equation for CF in Table 1) and identify

$$\hat{C}_{XY} = \frac{1}{N_X} \frac{1}{N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \left(N_{\text{coinc}}(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(y)}) - 2 n_i^{(x)} n_j^{(y)} \Delta / T \right) \tag{3.15}$$

as the averaged number of coincidences that is loosely related to an inner product. $\hat{C}_{XY}$ counts the number of coincidences without replacement above the expected number of coincidences that one would obtain if both spike trains were produced by a Poisson process ($2 n_i n_j \Delta / T$). Accordingly, we can define the average number of coincidences across distinct repetitions:

$$\hat{C}^*_{XX} = \frac{1}{N_X(N_X-1)} \sum_{i=1}^{N_X} \sum_{j\neq i}^{N_X} \left(N_{\text{coinc}}(\mathbf{S}_i^{(x)}, \mathbf{S}_j^{(x)}) - 2n_i n_j \Delta/T\right). \tag{3.16}$$

Here again, the expectation of $\hat{C}_{XY}$ will be always smaller than the expectation of $\hat{C}^*_{XX}$ or $\hat{C}^*_{YY}$. We then formulate a new version of CF2 (distance based) as

$$\text{CF2}^* = \frac{\hat{C}_{XY}}{\frac{1}{2}(\hat{C}^*_{XX} + \hat{C}^*_{YY})}. \tag{3.17}$$

We will show in the examples of section 3.4 that equation 3.17 removes the deterministic bias from CF2.

3.4.3 Hunter and Milton. It is also possible to create a bias-corrected measure equivalent to HM by considering the average coincidence count:

$$\hat{C}_{XY} = \frac{1}{N_X} \frac{1}{N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \frac{\text{HM}(\mathbf{S}_i^{(x)} \to \mathbf{S}_j^{(y)}) + \text{HM}(\mathbf{S}_j^{(y)} \to \mathbf{S}_i^{(x)})}{2}, \tag{3.18}$$

from which naturally follows the estimate of its upper bound:

$$\hat{C}^*_{XX} = \frac{1}{N_X(N_X-1)} \sum_{i=1}^{N_X} \sum_{j=i+1}^{N_X} \text{HM}(\mathbf{S}_i^{(x)} \to \mathbf{S}_j^{(x)}) + \text{HM}(\mathbf{S}_j^{(x)} \to \mathbf{S}_i^{(x)}), \tag{3.19}$$

which yields a measure similar to the distance-based M_D:

$$\text{HM}^* = \frac{\hat{C}_{XY}}{\frac{1}{2}(\hat{C}^*_{XX} + \hat{C}^*_{YY})}. \tag{3.20}$$

Through case studies, we will show in section 4 that VP*, CF2*, and HM* correctly distinguish differences between point processes.

4 Examples

In this section, we show that the sample bias can cause spike train measures to be unsuitable to distinguish between point processes despite the fact that the firing rate or temporal characteristics of the two point processes are different by a factor of 2 or 3.

4.1 A Framework for Comparing the Different Measures. Suppose we have a set of spike trains X that were generated with process x—for instance, N_X spike trains could come from the measurement of N_X neuronal responses to repeated presentation of a fixed (but potentially time-dependent) stimulus. Furthermore, we have a second set of spike trains Y generated with the process y. This second process could have been generated by a model. In order to measure whether the model process y is similar to the neuronal process x, we would like a similarity function between 0 and 1, $M(X, Y)$, to measure the match between X and Y. The letter M stands for *match* (or *similarity*).

In statistical terms we want to evaluate if we can reject the hypothesis that x is the same as y using the restricted data contained in X and Y. To do this, we consider the null hypothesis that Y is no more different from X than X is to another instantiation of x that we call X'. Let X' be another set of N_Y spike trains taken from process x. Our confidence in rejecting the hypothesis that x is the same as y grows with the probability of observing $M(X, X') > M(X, Y)$. With the test statistics

$$\mathcal{D} = M(X, X') - M(X, Y), \tag{4.1}$$

this probability is

$$\mathrm{p} = \int_0^\infty p(\mathcal{D})\, d\mathcal{D}, \tag{4.2}$$

where $p(\mathcal{D})$ is the probability density of $\mathcal{D}$ and the integral runs over positive values $\mathcal{D} > 0$.

In addition to p, another useful quantity is the mean of $\mathcal{D}$, defined as $\overline{\mathcal{D}} = E[\mathcal{D}] = \int_{-1}^{1} \mathcal{D} p(\mathcal{D}) d\mathcal{D}$. One would expect that whenever $E[M(X, Y)] < E[M(X, X')]$, $\overline{\mathcal{D}} > 0$, indicating that the match between process x and process y is worse than the match between process x and itself. Suppose that we construct an artificial data set for which we know a priori that x is different from y. Thus, we know that $\overline{\mathcal{D}}$ should be positive. If we nevertheless find a value $\overline{\mathcal{D}} < 0$, the only remaining possibility is that the measure of match M on which we have based our statistical variable $\overline{\mathcal{D}}$ is ill suited for assessing the similarity between collections of spike trains. We will show in this section that some measures have $\overline{\mathcal{D}} < 0$ even if process x is distinct from y.

In the following, we study the hypothesis that a model y (for which we can choose the parameters) is statistically different from neuronal data x. In section 5, the neuronal data come from recordings, but here our "neuronal data" are artificially generated. For a given data set X, the self-match $M(X, X')$ is fixed and close to 1, while the match $M(X, Y)$ depends on the choice of parameters of the model process y. Hence, maximizing the match between model and data is the same as minimizing the statistical

discriminability $\overline{\mathcal{D}}$ (see equation 4.1). Our aim is to shed light on the flaws of existing measures by showing that the mean discriminability $\overline{\mathcal{D}}$ can go negative—which should never occur for sensible measures of match.

4.2 Numerical Methods. In the following sections, we compute $\overline{\mathcal{D}}$ in different scenarios to investigate whether the metric has a bias indicated by $\overline{\mathcal{D}} < 0$. We calculate $\overline{\mathcal{D}}$ numerically for the following measures of match: angle-based M_a, bias-corrected M_a^*, distance-based M_D, bias-corrected distance M_D^*, rescaled Victor-Purpura VP, coincidence factor without replacement CF2, and Hunter and Milton HM, as well as our novel debiased measures VP*, HM*, and CF2*. We generate N spike trains of the process y (data set Y) and $2N$ spike trains of the process x (data sets X and X'). To calculate the expected value, we repeated the simulation of the $3N$ spike trains between 1000 and 10,000 times to obtain smooth discriminability profiles.

Artificial data sets were generated with different processes. To simulate inhomogeneous Poisson processes and the spike response model (characterized by the firing intensity $\nu(t)$ or the conditional intensity $\lambda(t)$), we discretized time with bins of $\Delta t = 0.1$ ms and compared the probability of spiking in that bin ($P_F = 1 - e^{\nu(t)\Delta t}$ or $P_F = 1 - e^{\lambda(t)\Delta t}$) to a random number of uniform distribution. All programs for the simulations were written in Matlab. For the VP measure, we used the Matlab code by Daniel Reich and Jonathan Victor.[6]

Different measures often correspond to a different kernel shape. To calculate M_a and M_D and their bias-corrected equivalents, we used $K_\Delta(s, s') = h_t(s)h_t(s')$ with the triangular kernel $h_t = (1 - |t|/2\Delta)\Theta(1 - |t|/2\Delta)$ as defined with the VP metric. In order to compare the other measures on the same basis, we chose the kernel's parameter (see equations 2.16, 2.17, 2.18, and 2.21) such that the integral of the kernel was equal to 4 ms: $\int_{-\infty}^{\infty} h_e(s;\Delta)ds = \int_{-\infty}^{\infty} h_t(s;\Delta)ds = \int_{-\infty}^{\infty} h_r(s;\Delta)\,ds = \int_{-\infty}^{\infty} h_g(s;\Delta)\,ds = 4$ ms. For example, h_r intrinsic to CF2 has an integral of 2Δ; therefore, we took $\Delta = 2$ ms for computing CF2. The integral of the kernel was 4 ms for all simulations and the figures shown in this article, except in section 4.7 where Δ was varied over a large range.

4.3 Case Study 1: Match Based on PSTH or Population Activity. In order to have full control over all details, we work with data sets X, X', and Y, which are generated artificially. We study three different point processes.

4.3.1 Homogeneous Gamma Renewal Process. We investigate how well the multiple measures discriminate between two homogeneous gamma renewal processes with different firing intensities. This can be seen as probing the sensitivity of a measure to distinguish two different messages in a pure

[6] http://www-users.med.cornell.edu/jdvicto/metricdf.html, accessed June 2010.

rate code. For example, a neuron might respond to stimulus A with spike trains that can be described as a gamma renewal process of rate ν_x. For a stimulus B, it is again a gamma renewal process, but of rate ν_y. Can we distinguish between stimulus A and B?

We sample from the gamma interval distribution of order 2 with spike rate ν. The probability density of observing an interspike interval of s is therefore

$$p(s) = 4\nu e^{-2\nu s}(\nu s). \tag{4.3}$$

Using a total time of $T = 150$ seconds and a first spike randomly put within the first second, we generate $2N$ spike trains with $\nu = \nu_x = 10$ Hz to form X and X' and N spike trains with $\nu = \nu_y$ to form Y. ν_y can vary between 1 and 25 Hz. Sample spike trains from X with $\nu_x = 10$ Hz and for Y with $\nu_y = 1$ Hz are shown in Figure 3a. Changing the order of the gamma renewal process to 1 (Poisson) does not induce sizable changes in the curves of Figures 4a and 4b and Figures 5a to 5c (results not shown).

Figure 4 shows the discriminability profiles for $N = 20$, 60 for the angle-based match (M_a, equation 2.25) and the distance-based match (M_D, equation 2.27) and the bias corrected measures (M_a^*, equation 3.9, and M_D^*, equation 3.10), with $N = 20$. Biased similarity measures of angular separation show a region where $\overline{\mathcal{D}} < 0$. This bias does not disappear if we go from a small sample of $N = 20$ to a rather large sample of $N = 60$. As discussed in section 5.1, a value of $\overline{\mathcal{D}} < 0$ indicates that the measure of match is flawed.

When we correct for sample bias, we observe an elimination of the negative $\overline{\mathcal{D}}$ areas. Three conclusions can be drawn from Figures 4a and 4b. First, a naive measure based on angles or distances is biased, as indicated by negative values of $\overline{\mathcal{D}}$. Second, even a bias-corrected measure based on angles is not able to discriminate between two ensembles of different firing rates. Third, the bias-corrected measure based on distance is highly sensitive to different firing rates.

Let us mention in passing a minor point. We argued in section 3 that the angular separation should be completely insensitive to the modifications in the amplitude of the firing intensity. However, in Figure 4a, the bias-corrected angle-based measure shows positive $\overline{\mathcal{D}}$ at $\nu_y = 1$ Hz. This can be explained by recalling that it is not entirely unbiased when the norm of the population activity is close to 0 (see section 3.2). Indeed, for $\nu_y = 1$ Hz, we calculate $\hat{C}^*_{YY} = 0.11$ with a distribution that is clearly not gaussian. Under these conditions, the inverse of an unbiased estimator is not itself unbiased and this additional bias can explain why M_a^* appears sensitive to the firing rate difference between $\nu_x = 10$ Hz and $\nu_y = 1$ Hz.

4.3.2 Spike Time Jitter. Next we investigate how well the multiple measures discriminate features of a spike timing code. We can think, for instance, of the first-spike latency code discussed in Gawne, Kjaer, and Richmond

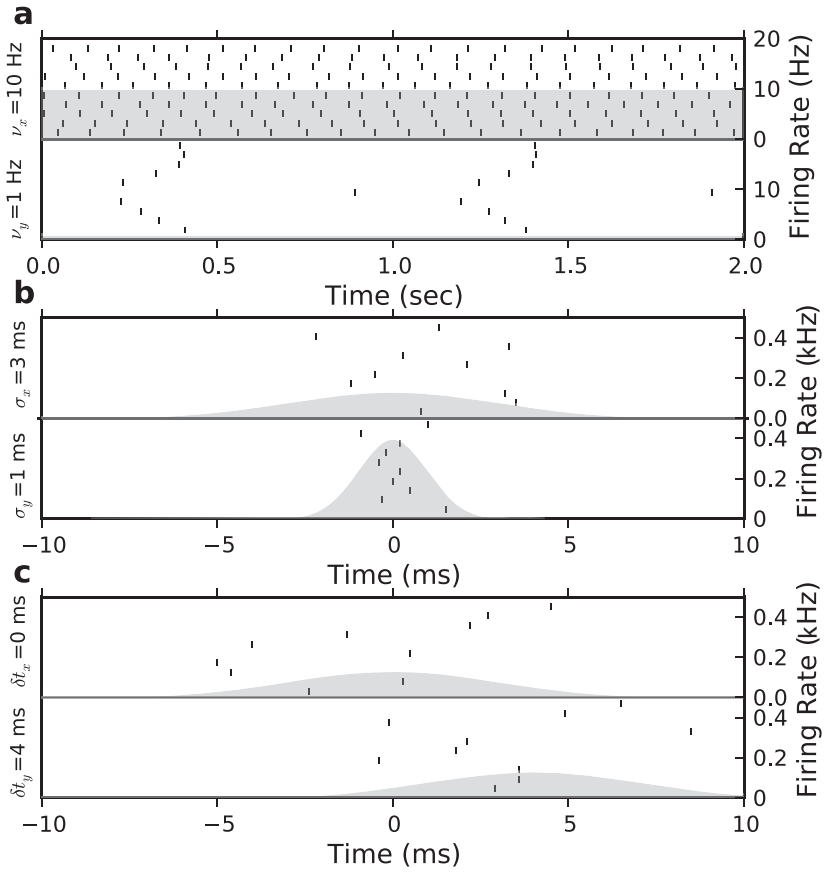

Figure 3: Raster plots and firing rates illustrating the different case studies. Two sets of $N = 10$ spike trains (examples of X and Y) are shown: (a) comparing two gamma renewal processes, one with $\nu_x = 10$ Hz (top) and the other with $\nu_y = 1$ Hz (bottom); (b) spike time jitter with $\sigma_x = 3$ ms (top) and $\sigma_y = 1$ ms (bottom); (c) spike time latency with $\delta t_x = 0$ ms (top) and $\delta t_y = 4$ ms (bottom). The shaded area shows the instantaneous firing rate of the process with its axis on the right.

(1996) and in Gollisch and Meister (2008). We consider the arrival of a single spike with a given average latency and at a given precision. We look at the sensitivity of the similarity measures to changes in precision. In this scenario, we generate $2N$ spike trains made of a single spike shifted by a gaussian random number having a standard deviation of $\sigma_x = 3$ ms. These spike trains form the sets X and X'. We ask whether we can discriminate the spike trains from N other spike trains made of a single spike, each shifted by a gaussian random number having a different standard deviation of σ_y,

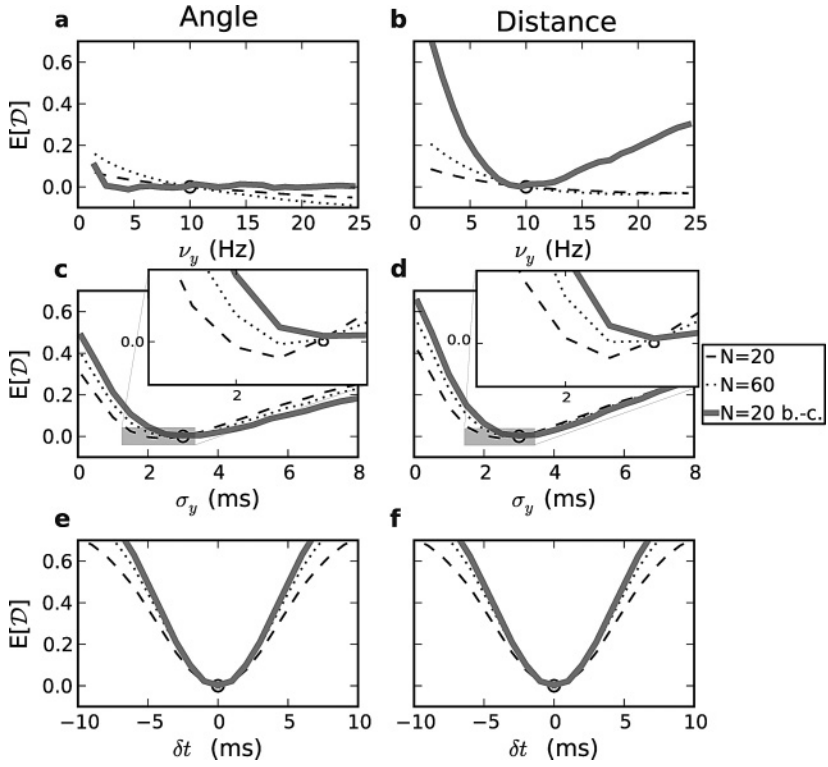

Figure 4: Discriminability profiles for three similarity measures of spike trains seen as vectors: M_a, M_a^* (a, c, e) and M_D, M_D^* (b, d, f). The discriminability is plotted for different values of either the firing intensity of a gamma renewal process (first row; a, b), the spike time jitter (second row; c, d), or the average latency (third row; e, f). The discriminability profiles are shown with N spike trains with dashes ($N = 20$), dots ($N = 60$), and a thick line for the bias-corrected measures ($N = 20$). The reference values of the firing intensity, the jitter, or the latency in ensemble X and X', are indicated with a circle at $\overline{\mathcal{D}} = 0$: $\nu_x = 10$ Hz in a and b, $\sigma_x = 3$ ms in c and d, $\delta t = 0$ ms in e and f. The zoomed-in plot of the shaded area is shown as an inset for c and d.

which can vary between 0 and 8 ms. Sample spike trains from X with $\sigma_x = 3$ ms and for Y with $\sigma_y = 1$ ms are shown in Figure 3b.

We find that for $N = 20$, naive implementations of the angle-based match M_a and the distance-based match M_D show a region where $\overline{\mathcal{D}} < 0$. The size of the region decreases with the increasing size of the sets, as we can see in the insets of Figures 4c and 4d. The region where $\overline{\mathcal{D}} < 0$ occurs for parameters $\sigma_y < \sigma_x$ and indicates the deterministic bias: a spiking process

is seen as more similar to a process with a smaller jitter than to a process with the correct jitter. In contrast, the bias-corrected measures M_a^* and M_D^* can discriminate spikes arriving with any two distinct jitters.

Since each spike train consists of only a single spike, we expect from the results in section 3.2 that M_a^* and M_D^* can have an additional bias due to the inverse of the unbiased estimator. Nonetheless, we do not observe a region with negative discriminability for the bias-corrected measures, indicating that in this case, the additional bias due to the inverse is negligible. We conclude that for timing-based codes, the raw measures of match M_a (based on angle) and M_D (based on distance) exhibit a deterministic bias, that is, they favor a model interpretation with less jitter than the real process. Both bias-corrected matches M_D^* and M_a^* work well.

4.3.3 Spike Time Latency. To probe the sensitivity to average latency, we generate $2N$ spike trains made of a single spike shifted by δt_x plus a gaussian random number with standard deviation of $\sigma = 3$ ms. We ask whether we can distinguish these spike trains from N spike trains made of a single spike coming with a different latency δt_y plus a gaussian random number with a standard deviation of $\sigma = 3$ ms. This will produce a firing intensity given by

$$\nu(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\delta t_y)^2}{2\sigma^2}\right). \tag{4.4}$$

Sample spike trains from X with $\delta t_x = 0$ ms and for Y with $\delta t_y = 4$ ms are shown in Figure 3c.

Figures 4e and 4f show the discriminability profiles for $N = 20$, and 60 for M_a and M_D and the bias-corrected measures M_a^* and M_D^* with $N = 20$. We observe that increasing the size of the sample increased the discriminability. We note that the bias-corrected measures with $N = 20$ show better discriminability than the raw measures for a much larger sample size of $N = 60$.

The overall conclusion of section 4.3 is that the bias-corrected measures always work better than the biased ones and the distance-based match M_D^* works on all three scenarios, whereas the angle-based match M_a^* works only for temporal codes and is insensitive to firing rate codes.

4.4 Case Study 2: Measures Loosely Related to an Inner Product. Using the same paradigms (gamma renewal process, spike time jitter, and spike time latency), we check the quality of those well-known similarity measures that relate only loosely to the vector space framework. In particular, we consider the HM measure, the coincidence factor CF2, the Victor-Purpura metric VP (see section 2.7), along with HM*, VP*, and CF2*. We calculate numerically the value of $\overline{\mathcal{D}}$ for each measure and plot the results

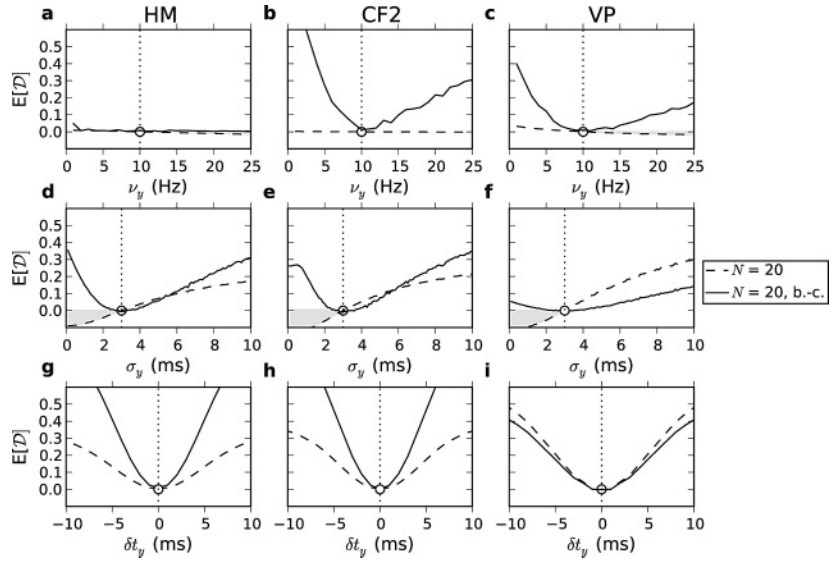

Figure 5: Discriminability profiles for three standard similarity measures: the Hunter and Milton measure (HM; a, d, g), the coincidence factor without replacement (CF2; b, e, h), and the scaled Victor and Purpura metric (VP; c, f, i). The discriminability is plotted (dashed line) for different values of the firing intensity of a gamma renewal process (first row; a, b, c), the spike time jitter (second row; d, e, f), or the average latency (third row; g, h, i). The discriminability profiles for the analogous measures HM*, CF2*, and VP* are shown for all cases with a solid line. The target value of the firing intensity, the jitter, or the latency is indicated with a dotted line with a circle at $\overline{\mathcal{D}} = 0$. The area where a similarity measure shows a negative discriminability is indicated in gray.

in Figure 5 using $N = 20$ (changing N here will affect only the standard deviation, as explained in section 3.1). For a gamma renewal process (see Figures 5a–5c), there is an area where $\overline{\mathcal{D}} < 0$ for all three measures, but for both VP and HM, it is more pronounced than for CF2.

The fact that $\overline{\mathcal{D}} < 0$ for large firing rates indicates that a gamma renewal process that in reality has a low firing rate of 10 Hz will be seen as deceptively similar to a gamma renewal process at a higher firing rate. This can be understood by a greater number of random coincidences at higher firing rates that are not discounted in the similarity measures. On the other hand, CF2 appears not to be sensitive to different firing rates. This is not too surprising because CF2 corrects for random coincidences. A comparison with Figure 4a therefore suggests that we can interpret CF2 as an angular separation between spike train vectors, since both angular match and CF2 are insensitive to firing rates.

Modified versions of the three measures (HM*, CF2*, and VP*) have $\overline{\mathcal{D}} \geq 0$ for all firing rates ν_y. We see in Figure 5a that HM* is insensitive to changes in the overall firing rate, which would make this measure conceptually related to an angular separation. CF2* and VP* have a strong sensitivity to the overall firing rate and so are to be interpreted as distance-based measures.

When investigating the discriminability for spike time jitter, we find that HM, CF2, and VP exhibit a region of negative discriminability (see Figures 5d–5f). This means that a process generating spikes with a greater precision will be systematically overrated when compared to a process with larger jitter. This is what we called deterministic bias in section 3.1. The modification suggested in section 3.4 removes this deterministic bias in the new measures HM*, CF2*, and VP*, which show $\overline{\mathcal{D}} \geq 0$ for all jitter σ_y.

All measures retained a high and positive discriminability for average latency (see Figures 5 g–5i). This indicates that the three measures HM, VP, and CF2 are equally useful for the detection of spike latency codes. The discriminability was greater when using HM* instead of HM and when using CF2* instead of CF2, but not when using VP* instead of VP.

We conclude that HM, CF2, and VP all fail to discriminate the level of precision and sometimes the overall firing rate. The measures HM*, CF2*, and VP* have $\overline{\mathcal{D}} \geq 0$ everywhere. We note that VP* and CF2* were sensitive to changes of firing statistics from all scenarios considered, with CF2* showing a slightly greater sensitivity.

4.5 Case Study 3: Phase Code. We can explore the deterministic bias further by considering an inhomogeneous Poisson process with a tunable proportion of spikes that arrive at specific times. Let us consider an inhomogeneous Poisson process with a time-varying intensity made up of 50 equally spaced gaussian bumps superimposed on a constant baseline. The amplitude of the baseline is chosen to complement the amplitude of the gaussian such that the expected number of spikes is always 50 for a single trial. The amplitudes are regulated by a parameter α:

$$\nu(t) = 50\frac{\alpha}{T} + \frac{1-\alpha}{\sqrt{2\pi\sigma^2}} \sum_{k=1}^{50} \exp\left(-\frac{(t - t_0 - k\delta t)^2}{2\sigma^2}\right), \tag{4.5}$$

with the total time $T = 5$ s, the interbump interval $\delta t = 100$ ms, and the latency of the first bump $t_0 = 50$ ms. The parameter α tunes the proportion of spikes that will arise from the gaussians versus the number of spikes that will arrive at random times (see Figure 6). When $\alpha = 1$, the spike train is Poisson, and, conversely, when $\alpha = 0$, spiking is entirely periodic with a jitter σ of 3 ms. We call α the random fraction. This type of spike-generating

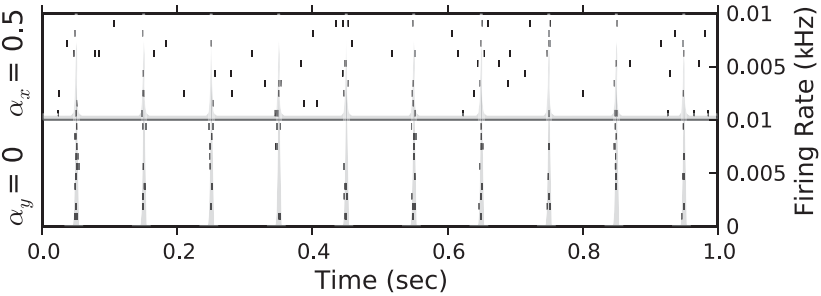

Figure 6: Raster plots and firing rates illustrating the phase coding. Two sets of $N = 10$ spike trains (examples of X and Y) are shown for comparing mixtures of random spike times and precisely locked spike times—one with $\alpha_x = 0.5$ (top) and the other with $\alpha_y = 0$ (bottom). The shaded area shows the instantaneous firing rate of the process with its axis on the right.

process can be related to a phase code riding on top of a random background (Lisman, 2005; Kayser, Montemurro, Logothetis, & Panzeri, 2009).

The inhomogeneous Poisson process is simulated using equation 4.5. We generate $2N$ spike trains with 50% of phase-locked spikes ($\alpha_x = 0.5$). We then compare those spike trains with N spike trains generated with α_y and calculate $\overline{\mathcal{D}}$ numerically.

Figure 7 shows the discriminability profile for M_a^*, M_D^*, HM, CF2, and VP with $N = 10$. The thick solid line gives the mean value of $\mathcal{D}$ averaged over many samples with $N = 10$ each. We can observe again that only the measures M_a^* and M_D^* do not have a region where $\overline{\mathcal{D}} < 0$ when $\alpha_y < \alpha_x$. All the commonly used measures such as VP, HM, and CF2 have a strong bias. If the real data have a random fraction α_x and the model α_y, model parameters optimized with VP, HM, or CF2 will give a best value of $\alpha_y = 0$. That is, a smaller random fraction increases the value of the similarity measures HM, CF2, and VP despite the fact that the process for generating the ensemble X has a stochastic fraction $\alpha_x = 0.5$. This observation confirms a strong deterministic bias that is to be expected in view of our theoretical results in section 3.

The standard deviations in Figure 7 are much larger for the measures M_a^* and M_D^* than for HM, CF2, and VP. This comes from the estimation of the variance required for correcting the bias and is less precise at small sample sizes. The question arises whether the large standard deviation is a disadvantage of M_a^* or M_D^* as compared to their biased versions M_a and M_D.

We compared the biased and bias-corrected distance and looked at the variance of $\mathcal{D}$. Figure 8a shows that the standard deviation is greater for bias-corrected M_D^* than for biased M_D by superposing the curves on the

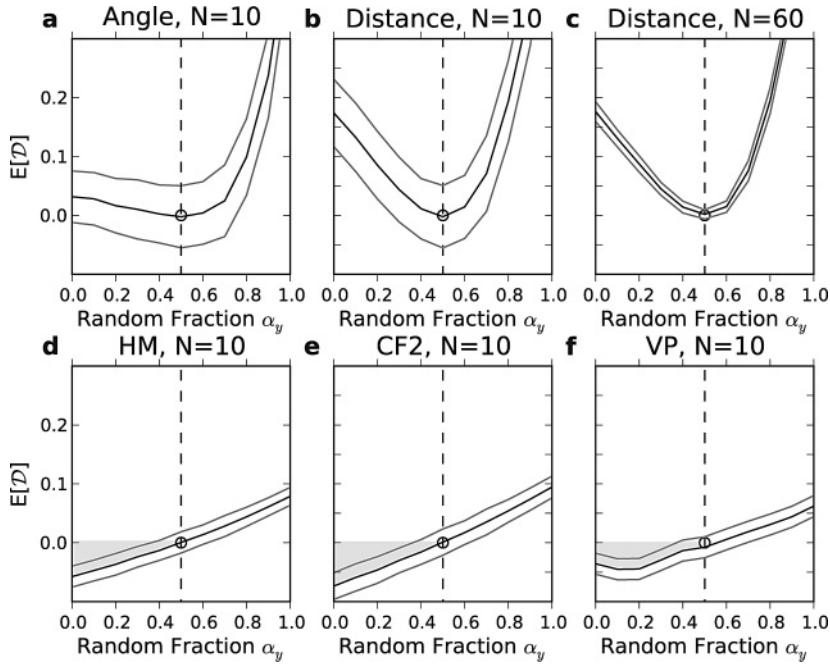

Figure 7: Phase coding: Discriminability profiles for: M_a^* (a: $N = 10$), M_D^* (b: $N = 10$; c: $N = 60$), the Hunter and Milton measure (d: $N = 10$), the coincidence factor without replacement (e: $N = 10$), and the scaled Victor and Purpura metric, VP, (f: $N = 10$). The discriminability is plotted (solid black line) for different values of the fraction of spikes that arrive at a fixed time.The target value of the fixed fraction is indicated with a dashed line with a circle at $\overline{\mathcal{D}} = 0$. The area where a similarity measure shows a negative discriminability is indicated in gray.

same plot. Keeping the random fraction fixed at $\alpha_y = 0.2$ and changing the size of the sample used to construct the population activity, we see that the standard deviation decreases rapidly as we increase N (see Figure 8c). The point where the lower line of the standard deviation crosses 0 is $N = 6$ for the bias-corrected measure and $N = 30$ for the biased measure. Moreover, it takes roughly $N = 400$ samples for the biased measure to achieve the same discriminability as the bias corrected.

The probability of observing $\mathcal{D} < 0$ is the confidence level of observing $\mathcal{D}$ within the confidence interval of negative discriminability. In Figure 8b, we show that the probability of observing $\mathcal{D} < 0$ as a function of the sample size N is always smaller for the bias-corrected M_D^*.

In summary, the bias-corrected match M_D^* works extremely well for phase-coded input. It is significantly better than the angle-based match M_a^*. The classical measures such as HM, CF2, and VP fail completely.

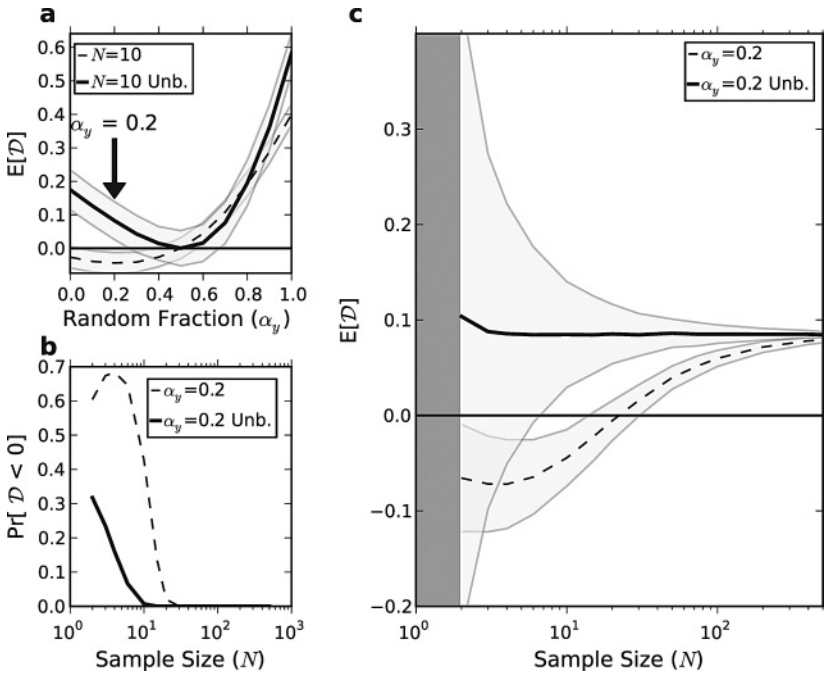

Figure 8: The similarity measure M_D^* increases discriminability for all sample sizes greater than 1. (a) The discriminability for the scaled distance between population activities with $N = 10$ (M_D, dashed line) and the same measure corrected for small sample bias (M_D^*, solid thick line). (b, c) The locked fraction of 0.2 is chosen. (b) The probability of observing a negative discriminability is smaller for the bias correction (M_D^*, solid thick line) than without the bias correction (M_D, dashed line) at all sample sizes above 1. (c) Discriminability is plotted against the sample size with (M_D^*, solid thick line) and without (M_D, dashed line) bias correction. The area within 1 standard deviation of the expected value is shown in light gray.

4.6 Case Study 4: Spike Response Model. We evaluate the discriminability of the similarity measures on more realistic data by using the spike-response model with escape noise to generate the spike trains (SRM; Plesser & Gerstner, 2000; Gerstner, 2008). The SRM considered here is a generalized linear model (GLM) for which the probability of spiking depends on the spiking history (Gerstner, van Hemmen, & Cowan, 1996; Truccolo et al., 2005; Pillow et al., 2005). We use a simple exponential for both the input filter (having a resistance R and a time constant τ_m: $\kappa(t) = \frac{R}{\tau_m} e^{-t/\tau_m}$) and the spike history kernel (having an amplitude A and a time constant τ_A: $\eta(t) = -Ae^{-t/\tau_A}$). The conditional firing intensity for this model is

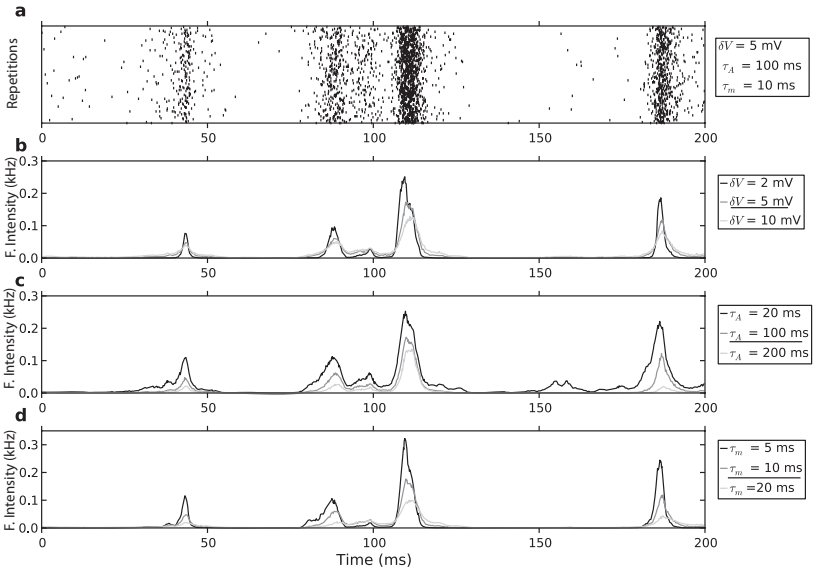

Figure 9: Illustrating the effect of the model parameters on the PSTH. (a) The spike times of $N = 100$ repeated presentation of the same stimulus to a spike response model with $\delta V = 5\,\text{mV}$, $\tau_A = 100\,\text{ms}$, and $\tau_m = 10\,\text{ms}$. The PSTHs (thick lines) in panels b, c, and d are produced by averaging the number of spikes in each 0.1 ms time bin over $N = 10{,}000$ repetitions, using the same paradigm as in panel a. Using the same time-dependent input, three PSTHs are produced for different values of (b) the scaling factor (δV), (c) the adaptation time constant (τ_A), and (d) the time constant of the input filter (τ_m).

$$\lambda(t|I, S_t) = f\left(\frac{[\kappa * I](t) + [\eta * S_t](t)}{\delta V} - \theta_0\right), \tag{4.6}$$

where δV scales the stochasticity of the model (also called the temperature), θ_0 relates to the firing threshold and regulates the offset, f is the link function (here an exponential $f(x) = e^x$), S_t is the spike train of all times before t, and I is the stimulus current. We model the stimulus as an Ornstein-Uhlenbeck process (mean: 0 pA, std. 20 pA, correlation time: 3 ms) representing current arriving from a large number of synapses. For all the simulations presented, we fixed $A = 2\,\text{mV}$, $R = 0.1\,\text{M}\Omega$, and $\theta_0 = 2$. We used spike trains of 1 second to calculate numerically the discriminability $\overline{D}$ of M_D^* and VP.

Figure 9 shows that changing the parameters δV, τ_m, τ_A has a nontrivial effect on the instantaneous firing intensity. Let us now assume that the real neuron has parameters $\delta V = 5$ mV, $\tau_A = 100$ ms, and $\tau_m = 10$ ms. We use this neuron model to generate the data set X with $N = 10$ spike trains. We

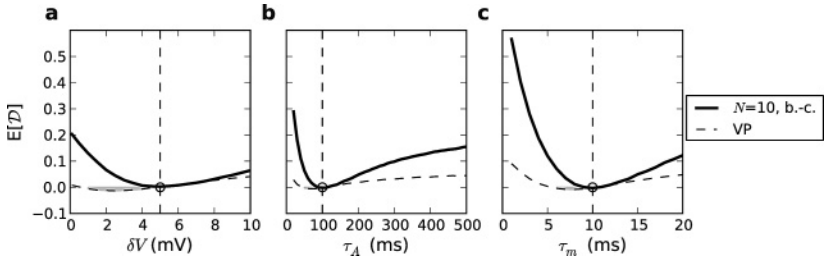

Figure 10: The discriminability of the Victor and Purpura measure (dashed line) and M_D^* ($N = 10$, solid thick line) is shown for varying model parameters: (a) the scaling factor (δV), (b) the adaptation time constant (τ_A), and (c) the time constant of the input filter (τ_m). The target values are indicated with a dashed line with a circle at $\overline{\mathcal{D}} = 0$. The area where a similarity measure displays a negative discriminability is indicated in gray.

now try to extract the correct parameters by simulating the same model but with variable parameters.

Naively, one would expect that we would find the correct set of parameters by minimizing a discriminability $\mathcal{D}$ between our simulations Y and reference sets X, X'. Figure 10 shows that this is not the case if we base the distance measure on the VP metric. The bias in the VP measure (and other measures in the literature) gives rise to a pseudo-optimal set of parameters with $\delta V = 2$, $\tau_m = 8$ ms, and $\tau_A = 80$ ms. On the other hand, if we use the unbiased distance measure M_D^*, we retrieve the correct parameters for δV, τ_A, and τ_m. We conclude that the bias in firing rate and spike jitter that we saw earlier in the VP measure in Figure 5 gives rise to a strong mismatch of parameters.

4.7 Case Study 5: Classification. Spike train similarity measures have been used extensively for the classification of spike trains. The previous examples have shown that two-category classification requires methods such as M_d^*, VP*, HM*, or CF2* to distinguish processes that differ in their level of stochasticity. Here, we compare classification performance using the classical D_{spk} defined in Victor and Purpura (1997) and the new D_{spk}^* defined in Table 2.

We follow the leave-one-out classification method of Victor and Purpura (1997). We use the phase coding of section 4.5 to conceive five different classes of spike trains according to five different values of the random fraction ($\alpha = [0, .25, .5, .75, 1]$). We draw 25 spike trains from each of the five different classes, which makes a total of 125 spike trains. We will estimate the classification performance by randomly choosing one of the 125 spike trains and estimating with a spike train similarity measure to which of the five classes it should belong. To do this, we measure the average D_{spk}

of the singled-out spike train $S_i^{(\nu)}$ with $S_j^{(\mu)}$ the 25 or 24 spike trains of each class μ:

$$d(S_i^{(\nu)}, \mu) = \left[\frac{1}{N} \sum_{j=1}^{N} D_{\text{spk}} \left(S_i^{(\nu)}, S_j^{(\mu)} \right)^z \right]^{1/z}, \tag{4.7}$$

where the exponent z is introduced to span various topological situations (Victor & Purpura, 1997). The singled-out spike train is assigned to the class with the minimum average distance. We then repeat the operation N_{tot} times and build the matrix $Z_{\nu\mu}$ made of the number of times a spike train from class ν was attributed to class μ. The classification performance gives an estimate of the transmitted information H in bits (Victor & Purpura, 1997):

$$H = \frac{1}{N_{tot}} \sum_{\nu,\mu} Z_{\nu\mu} \left[\log_2 Z_{\nu\mu} - \log_2 \sum_{\nu} Z_{\nu,\mu} - \log_2 \sum_{m} Z_{\nu,m} + \log_2 N_{tot} \right]. \tag{4.8}$$

We repeat the classification five times for multiple values of $q = 2/\Delta$, and then repeat the whole procedure again with D^*_{spk}.

Using D^*_{spk} instead of d almost doubles the amount of transmitted information H recovered when q is sufficiently large (see Figure 11). The increase in transmitted information is robust to changes in the exponent z (see Figure 11, right). At small q, the classification is almost random. This can be explained by the fact that for the small cost q of moving a spike corresponds to a large timescale on which the metric is sensitive. Therefore, increasing q makes the metric gradually more sensitive to the presence of precisely timed spikes coming with a gaussian jitter of ± 2 ms. The advantage of using D^*_{spk} instead of D_{spk} starts to be significant at $q = 8$ cost per second and increases further until a maximum is reached at $q = 256$ cost per second. The optimum of D^*_{spk} at $q = 256$ cost per second corresponds to a maximal temporal shift of $\Delta = 4$ ms.

We conclude that the improved similarity measure can increase the information recovered with spike train classification. Since the classification performance depends heavily on the statistics of the spike train classes, further investigation is required to assess the extent to which this increase is significant and general.

5 Fitting Cortical Neurons

In 2009, the International Neuroinformatics Coordinating Facility (INCF) organized the third competition in modeling single neuron activity. The

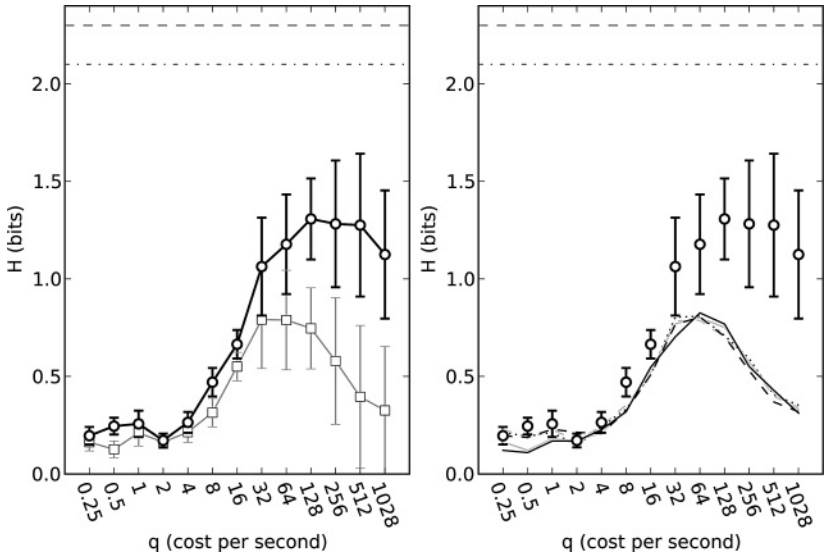

Figure 11: Classification of spike trains coming from different point processes is improved by using D^*_{spk} instead of the measure of equation 4.7, d. (Left) The transmitted information H is shown as a function of the cost of moving one spike q for classifications performed using D^*_{spk} (thick black line, circle markers) or the original Victor-Purpura distance d with $z = 2$ according to equation 4.7 (gray line, square markers). Leave-one-out classification was used on five different classes of spike trains. Perfect classification in a scenario with five equally possible classes would have resulted in $H = \log_2(5) = 2.3$ bits shown with the dashed line. The maximum likelihood classifier under the assumption that the generative model is known would be able to achieve $H = 2.1$ bits (dot-dashed). Error bars indicate 1 standard deviation. (Right) The transmitted information H is shown as a function of the cost of moving one spike q for classifications performed using D^*_{spk} (circle markers with error bars) or the original Victor-Purpura distance d with various z exponents ($z = -2$ black dotted, $z = 4$ solid gray line, $z = -4$ dashed gray line, $z = 8$ solid black line, $z = -8$ dashed black line).

aim was to provide common benchmark data and evaluation criteria to compare the performance of different neural models (Gerstner & Naud, 2009). We use a subset of this publicly available data to demonstrate the impact of biased performance measures on the model selection process (see section 3.1).

5.1 Experimental Details. An L5 pyramidal cell in a brain slice of a 14-day-old rat was stimulated with a time-dependent, in vivo–like current. Somatic recordings were made to obtain the voltage trace of the membrane

potential. In current-clamp mode, a current was injected through the same patch pipette. The current waveform was generated using a superposition of Poissonian excitatory and inhibitory spike trains, convolved with two exponential kernels. In total, there were 13 repetitions of 60 seconds each. Full experimental details can be found elsewhere.[7]

The training set consists of the first 38 seconds for which both the injected current and the voltage trace are provided. For the test set, the injected current for the last 22 seconds is provided, and the spike times have to be predicted.

To compare the recorded spike trains with the predicted ones, the coincidence factor without replacement (CF2) with a window width of $\Delta = 4$ ms was used (see section 2.7). The question we address here is whether the choice of CF2 to quantify model performance could have had a strong influence on the final ranking of the submissions.

5.2 Neural Model and Fitting. We modeled the spiking activity of the cell with an SRM having an exponential link function ($f(x) = e^x$ in equation 4.6). As in section 4.6, the instantaneous firing rate $\lambda(t|H_t)$ is conditioned on a filtered version of the injected current $[\kappa * I](t)$, the filtered past activity of the neuron $[\eta * S_t](t)$, and a constant offset θ_0 to model baseline activity (see equation 4.6).

The spike history filter η is described as a sum of $m_h = 9$ basis splines (B-splines of order 3). Knot points are spaced on a roughly logarithmic scale up to 320 ms. Let the jth spline be of shape $A_j(\Delta t)$; then the contributing term to equation 4.6 is

$$\frac{[\eta * S_t](t)}{\delta V} = \sum_j^{m_h} \theta_{h,j}[A_j * S](t), \tag{5.1}$$

where $\theta_{h,j}$ are parameters of the spike history filter. Similarly, the current is filtered with a kernel that is made up of $m_c = 14$ spline basis functions $B_j(\Delta t)$ with knot points roughly logarithmically spaced up to 250 ms. The contribution from the current to the linear sum of equation 4.6 is thus given by

$$\frac{[\kappa * I](t)}{\delta V} = \sum_j^{m_c} \theta_{c,j}[B_j * I](t), \tag{5.2}$$

with parameters $\theta_{c,j}$ for the current filter.

[7] INCF Web site: http://www.incf.org/Events/competitions/spike-time-prediction/2009.

Spike times are extracted from the voltage trace by searching for zero crossings from below. The resulting series of spikes is transformed into a binary sequence with a bin size of 1 ms. The subthreshold fluctuations of the voltage trace are ignored for the fitting procedure because the aim of the competition organized by the INCF was a prediction of spike times rather than the subthreshold membrane potential. All parameters are determined using standard IRLS (iteratively reweighted least squares; McCullagh & Nelder, 1998) method to obtain a maximum likelihood estimate of the 24 parameters in $\vec{\theta}$.

The estimated spike history filter η captures mainly refractory effects. The kernel used for the current shows an approximately exponential decay with a positive influence of the injected current on spiking probability. Applying the time-rescaling theorem (Brown et al., 2002) suggests a good fit on the training data. Samples of spike trains are generated using the current trace of the test set to generate the model's prediction.

5.3 Deterministic Bias Penalizes the Maximum Likelihood Solution. Since the model is stochastic, we generated 13 different realizations of a spike train, conditioned on the same current—the 22 seconds of current given in the test set of the INCF competition. These 13 realizations are then compared with the experimentally observed spike trains. This procedure gives a first value of the CF2 score. The mean CF2 score, averaged over $N = 100$ repetitions, is $CF2^{(\mathrm{ML})} = 0.36$. Here, the index (ML) indicates this solution is found with a stochastic model using an optimization based on maximum likelihood as an objective. We now vary the stochasticity of the model to see whether the CF2 exhibits a bias toward deterministic models. The level of stochasticity is controlled by the parameter δV in equation 4.6. The maximum likelihood solution corresponds to $\delta V = 1$ mV. For $\delta V \to \infty$, the model tends to assign a constant spiking probability for each bin—a homogeneous Poisson process—while for $\delta V \to 0$, the model becomes more deterministic, the transfer function approaches a step function, and the neuron fires as soon as the right-hand side of equation 4.6 becomes positive. We vary the parameter δV in a wide range and adjust the constant offset θ_0 numerically to keep the firing rate approximately equal to the firing rate observed at $\delta V = 1$ mV.

For each value of δV, we generated 13 realizations from this modified model and calculated the CF2 performance. We repeat this procedure several times so as to find the mean CF2 values as a function of δV (see Figure 12, solid line). As expected, the performances approach 0 (chance level) for large values of δV since the model turns into a homogeneous Poisson process. However, for values lower than $\delta V = 1$ mV, a clear increase in performance is visible: the maximum likelihood solution does not yield the highest CF2 score; instead, the purely deterministic model ($\delta V \to 0$) exhibits the peak performance of around 0.45 for this model setup.

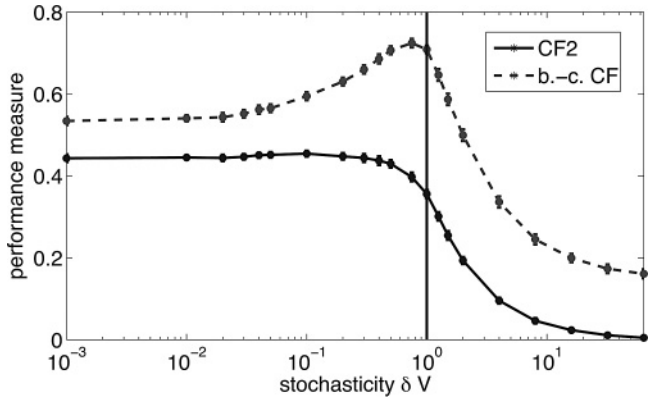

Figure 12: Fitting cortical neurons. The model parameters obtained by maximum likelihood estimation ($\delta V = 1$ mV) were scaled by a temperature-like factor and the performances of these modified models were evaluated using the CF2 (solid line). The vertical line indicates the performance at the maximum likelihood estimate. For each δV, $N = 100$ predictions were generated with 13 trials each. Points denote mean performance; error bars denote standard deviation. The dashed line shows the bias-corrected distance measure based on the coincidence factor (M_D^* with the kernel K_Δ defined for CF in Table 1).

We reevaluated the same spike time predictions using a debiased version of the coincidence factor, M_D^*, using the kernel K_Δ defined for CF in Table 1 (see Figure 12, dashed line). The peak performance in this metric now almost coincides with the maximum likelihood solution, and hence the bias toward a deterministic threshold model has disappeared. These results also suggest that the maximum likelihood estimate is optimal in the sense that the model performance cannot be trivially enhanced by moving away from the maximum likelihood estimate (choosing $\delta V \neq 1$ mV). We note that changing δV is equivalent to rescaling all model parameters by a constant factor.

Dividing performance measure CF2 by the intrinsic reliability of the neuron (which is a constant for a given neuron; here, it was 0.782) as in the original competition will scale CF2 uniformly, independent of δV. This scaling gives insight into the performance of the model in comparison to the intrinsic noise in the data but will not remove the deterministic bias.

5.4 New Ranking in INCF Competition. The INCF competition accepted both stochastic models and deterministic models. Based on our theoretical results in section 3 and the simulations of section 4, we expect that the CF2 measure penalized stochastic models and that reevaluating the submitted predictions with bias correction is important to make a fair ranking of how well the models match the data.

Table 3: Ranking of the INCF Competition.

Participant	$CF2/R$ (%)	Participant	CF2* (%)
1 (type II)	76.2	# 7 (type II)	74.3
2 (type I)	75.4	# 6 (type II)	72.3
3 (type I)	73.4	# 11 (type III)	68.7
4 (type I)	71.4	# 8 (type II)	68.4
5 (type I)	71.0	# 1 (type II)	66.4
6 (type II)	68.9	# 2 (type I)	66.1
7 (type II)	68.3	# 3 (type I)	64.5
8 (type II)	67.0	# 4 (type I)	62.7
9 (type I)	63.3	# 5 (type I)	62.7
10 (type III)	58.4	# 10 (type III)	59.6
11 (type III)	58.3	# 9 (type I)	57.0
12 (type III)	56.0	# 12 (type III)	49.6

Notes: Columns on the left: Evaluation of CF2 normalized by the intrinsic reliability of the data as published on the INCF Web page. Right: reordering with the bias-corrected measure CF2*. We find that all deterministic models move several ranks down when the corrected CF2* measure is used, that is, after removal of the deterministic bias. Submissions are labeled according to types of models. Type I: deterministic models with a single predicted spike train for all repetitions. Type II: deterministic models fitted independently on each repetition such that there are different predicted spike trains for each repetition. Type III: stochastic models.

Table 3 shows the ranking of all participations in challenge A in the INCF competition 2009 for both CF2 and the new CF2*. The two measures are therefore equivalent and differ mainly in how we normalize the number of coincident spikes. We hide the name and affiliation of participants but classify the models in three types: type I: deterministic models with a single predicted spike train for all repetitions; type II: deterministic models fitted independently on each repetition such that there is a different predicted spike train each repetition; and type III: stochastic models.

The first striking observation in Table 3 is that all deterministic models move several ranks down the list once deterministic bias is removed from the similarity measures. The second observation that catches the eye is that a stochastic model (#11, type III) that was ranked as one of the three worst models when using CF2 is now in the top three when we use the new measure for model ranking. With CF2*, the top three models are either stochastic (type III) or deterministic models fitted independently on each repetition (type II). Fitting independently on each repetition introduces variability in the predicted spike trains, which goes toward the variability intrinsic to real neurons. Some of the stochastic models performed badly with CF2 and still perform badly with CF2*. This indicates that the bias correction does not simply give an advantage to stochastic models but requires the correct type of stochasticity.

We observe that most of the submissions were made using deterministic models and that this could be explained by the fact that these models

had an unfair advantage against the stochastic models because CF2 was used to quantify the respective performances. It could be said that CF2 does not adequately take into account the stochastic nature of the neurons. Following the theoretical results in section 3 and the case studies in section 4, we propose that bias-corrected measures based on distance between population activities such as M_D^* or the analogous D_P^* and CF2* are more appropriate to quantify whether neuron models match real neurons.

6 Conclusion

6.1 Spikes as Vectors in a Vector Space. In the formalism proposed by Carnell and Richardson (2005), the inner product between spike trains yields natural definitions of distance or angular separation. The principle encompasses existing similarity measures such as those proposed by van Rossum (2001) or Schreiber et al. (2003). We noted that the coincidence factor of Kistler et al. (1997), as well as the VP measure (for large q values), can also be related to an inner product and thus to vector space measures. In a direct extension of the formalism, we introduced the population activity, that is, the vector made by averaging the spike train vectors within a sample, and showed that the scalar product between two population activities reduces to the commonly used correlation factor for PSTHs (Eggermont et al., 1983; Gawne et al., 1991; Paninski et al., 2005; David & Gallant, 2005; Petersen et al., 2008).

When spike trains, or pools of spike trains, are seen as vectors in a vector space, it is natural to consider quantities such as distance or angular separation to compare these spike trains. We have shown that angular separation can be interpreted as a similarity measure sensitive to the timing of events only, while the distance is a similarity measure sensitive to the timing of events and the overall firing rate (see section 3). We confirmed this difference with surrogate spike trains in section 4.3.

6.2 Kolmogorov-Smirnov Test. The improved similarity measures we have presented all boil down to a distance between PSTHs. As mentioned in section 1, this is just one way of taking into account the intrinsic variability. Another option would be to apply a Kolmogorov-Smirnov test or use Kullback-Leibler divergence to assess if the distances across spike trains within one sample have the same probability distribution as the distances across spike trains from two different samples. While the Kullback-Leibler divergence is notoriously hard to estimate for small samples sizes (Akaike, 1973), the Kolmogorov-Smirnov (K-S) test could provide a measure of distance between two sets of spike trains. The basic idea is to compare the distribution of pairwise distances $D(S_i^{(x)}, S_j^{(x)})$ within a set to that of the pairwise distances $D(S_i^{(x)}, S_j^{(y)})$ across the set X and the set Y, where D can be calculated with any of the commonly used spike train metrics. We will

pursue the direct analysis of this alternative method elsewhere. Preliminary results indicate that a K-S test on sets of spike trains can efficiently discriminate point processes with different levels of stochasticity and that there would be no sample bias (results not shown). We note that only a subset of the within-sample and the across-sample distances can be considered as i.i.d. (as required for the K-S test).

6.3 Beyond the PSTH. Consider two classes of perfectly periodic spike trains. The first class contains doublets of spikes with a period of $2T$, and the second class is made of single spikes with a period of T, but for both processes, the phase changes randomly between one trial and the next so that both processes have the same constant firing intensity. Because the PSTH is identical for both classes, the two classes cannot be distinguished with the methods presented in this article, although they would appear immediately distinguishable with parametric methods (Paiva et al., 2009b) or with the spike pattern metric of Victor and Purpura (1996) under the K-S test described in the previous paragraph.

6.4 Small-Sample Bias. We studied multiple types of spike train similarity measures in terms of their potential to discriminate among different probabilistic point processes. We avoided the variability introduced by the size of the coincidence kernel (Shimazaki & Shinomoto, 2007; Paiva, Park, & Príncipe, 2009a) by considering coincidence kernels normalized by a condition on the integral. We have shown that many of these measures perform badly on simple tasks and that this can be remedied by taking into account a small-sample bias in the comparison of firing intensities. This bias is relevant when the sample consists of fewer than approximately 400 spike trains (see Figure 8) and it can lead to dramatic failures when the available sample contains fewer than 30 spike trains.

From our case studies, we conclude that even if the spike time metrics are valid measures for comparing spike trains, one cannot correctly discriminate between the underlying—stochastic—point processes by simply looking at their element-by-element average. For each similarity measure, we proposed an improved version, and we showed that the novel variant is indeed sensitive to the statistical features of the point processes generating the spike trains. The relevance of the small-sample bias has been previously shown to be important in the comparison of PSTHs (David & Gallant, 2005; Petersen et al., 2008).

6.5 Implication for the Classification of Neuronal Responses. Our improved similarity measures for classification should partially improve the estimates of mutual information, even though there will always remain a sample bias because higher-order statistics are difficult to estimate from a limited number of elements (Treves & Panzeri, 1995; Panzeri et al., 2007). The extent to which our improved measures will increase classification

performance depends strongly on how the classes differ in terms of the statistical properties of the point processes. For instance, we expect that if the different classes differ in their spike timings, there should be a smaller advantage than when the differ in their spike time jitter or overall firing rate (see Figure 4). We have shown an example of the improved classification using the variable locked fraction scenario in section 4.7. In this example, the improved similarity measure increased the amount of information recovered. However, further investigations would be required to assess how significant this increase with real spike trains is.

6.6 Implication for the Measures of Intrinsic Reliability. Measures of intrinsic reliability such as R_X in equation 2.15 should be free of sample bias because they do not depend on $\|\hat{\mathbf{v}}\|$. Still, R_X can be biased when L_x is small and has a high variance, since the inverse of an unbiased estimator is not itself unbiased (Voinov, 1985). The methods presented here do not correct for this specific type of bias. On the other hand, our analytical results introduced a geometrical interpretation of the intrinsic reliability R_X: it corresponds to the variance of spike train vectors around the population activity (i.e., their center-of-mass, or PSTH). We also noted that intrinsic variability (Croner et al., 1993) corresponds to intrinsic reliability (Schreiber et al., 2003) under different normalizations.

6.7 Implications for Fitting: The Deterministic Bias. Why should none of the standard similarity measures be used for fitting neuronal processes? We have shown that a simple element-by-element average induces a preference for more deterministic models when a spike train similarity measure is used as a cost function for fitting a mathematical model to data (see sections 3.1 and 5.3). Moreover, we have indicated that the parameters obtained when optimizing a biased similarity measure are wrong. When the available sample size is small, we have shown that it is possible to use bias-corrected measures; otherwise, the estimated parameters such as the level of stochasticity in the model will be inadequately estimated. In general, element-by-element comparison yields an underestimate of the similarity between a stochastic model and a real neuron. This deterministic bias turns out to be responsible for shifting the maximum of the similarity measures such as coincidence factors away from the maximum likelihood solution, which explains a previously reported discrepancy (Shinomoto, 2010). Perhaps the most striking example of the importance of the deterministic bias was the reordering of the results of the single-neuron prediction challenge (see section 5.4).

Appendix A: Exemplar Code

Here we provide an example code called MDcode to compute the normalized, bias-corrected distance M_D^*. The Matlab program can be run by

```
function MDstar =  MDcode(Set1, Set2, Window, Type)
Ncoinc = (length(Window)-1)/2;
Nrep1 = size(Set1,1);  Nrep2 = size(Set2,1);
Set1f = zeros(size(Set1));  Set2f = zeros(size(Set2));
for i = 1:Nrep1
    Set1f(i,:) = circshift( fftfilt(Window, Set1(i,:)'), -Ncoinc)';
end
for i = 1:Nrep2
    Set2f(i,:) = circshift( fftfilt(Window, Set2(i,:)'), -Ncoinc)';
end
m1 = 0; m2 = 0;
if Type==1% van Rossum like
    nu1 = mean(Set1f,1); nu2 = mean(Set2f,1);
    for i = 1:Nrep1;
        m1 = m1 + sum(Set1f(i,:).*sum(Set1f(i+1:end,:),1));
    end
    for i = 1:Nrep2
        m2 = m2 + sum(Set2f(i,:).*sum(Set2f(i+1:end,:),1));
    end
else % Type 2 coincidence factor-like
    nu1 = mean(Set1f,1); nu2 = mean(Set2,1);
    for i = 1:Nrep1;
        m1 = m1 + sum(Set1(i,:).*sum(Set1f(i+1:end,:),1));
    end
    for i = 1:Nrep2
        m2 = m2 + sum(Set2(i,:).*sum(Set2f(i+1:end,:),1));
    end
end
m1 = m1/Nrep1/(Nrep1-1)); m2 = m2/Nrep2/(Nrep2-1);
MDstar = sum(nu1.*nu2)/(m1 + m2);
```

saving a ASCII file containing the above lines of code under the name MDcode.m.

The program takes four inputs: Set1, Set2, Window, Type. Set1 and Set2 are matrices containing the set of spike trains to be compared. They are built by stacking (adding rows) binned spike trains made of 0s and 1s. The number of rows should correspond to the number of repetitions, and the number of columns should correspond to the number of (small) time bins. Window corresponds to the coincidence windows used to compute the inner products. It must have odd length and should be symmetric

around its center if the inner product type is 2. Type = 1 corresponds to the Van Rossum metric (see Table 1) with an arbitrary coincidence window. If Type = 2, it corresponds to the Coincidence factor with an arbitrary coincidence window.

This Matlab program can easily be modified to compute the bias-corrected distance D^* or the bias-corrected angular separation M_a^* by modifying the last line of code according to equations 3.8 and 3.9. The use of *fftfilt* requires a special toolbox, but it can easily be replaced by *conv*, which is provided with the basic distribution.

References

Abeles, M. (1991). *Corticonics*. Cambridge: Cambridge University Press.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In F. Csáki & B. N. Petrov (Eds.), *Proc. of the 2nd Int. Symp. on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science*, *273*, 1868–1871.

Badel, L., Lefort, S., Brette, R., Petersen, C., Gerstner, W., & Richardson, M.J.E. (2008). Dynamic IV curves are reliable predictors of naturalistic pyramidal-neuron voltage traces. *Journal of Neurophysiology*, *99*(2), 656–666.

Brown, E., Barbieri, R., Ventura, V., Kass, R., & Frank, L. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, *14*, 325–346.

Calvin, W., & Stevens, C. (1968). Synaptic noise and other sources of randomness in motoneuron interspike intervals. *Journal of Neurophysiology*, *31*, 574–587.

Carandini, M. (2004). Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS Biology*, *2*(9), 1483–1493.

Carnell, A., & Richardson, D. (2005). Linear algebra for time series of spikes. In M. Verleysen (Ed.), *Proceedings of the 13th European Symposium on Artificial Neural Networks* (pp. 363–368). Evere, Belgium: d-side.

Chi, Z., & Margoliash, D. (2001). Temporal precision and temporal drift in brain and behavior of zebra finch song. *Neuron*, *32*(5), 899–910.

Chi, Z., Wu, W., Haga, Z., Hatsopoulos, N. G., & Margoliash, D. (2007). Template-based spike pattern identification with linear convolution and dynamic time warping. *Journal of Neurophysiology*, *97*(2), 1221–1235.

Christen, M., Kohn, A., Ott, T., & Stoop, R. (2006). Measuring spike pattern reliability with the Lempel-Ziv-distance. *J. Neurosci. Methods*, *156*(1–2), 342–350.

Croner, L. J., Purpura, K., & Kaplan, E. (1993). Response variability in retinal ganglion cells of primates. *Proc. Natl. Acad. Sci. USA*, *90*(17), 8128–8130.

Daley, D., & Vere-Jones, D. (1988). *An introduction to the theory of point processes*. New York: Springer.

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network*, *16*(2–3), 239–260.

de Ruyter van Steveninck, R. R., & Bialek, W. (1988). Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proc. R. Soc. B*, *234*, 379–414.

DiLorenzo, P. M., & Victor, J. D. (2003). Taste response variability and temporal coding in the nucleus of the solitary tract of the rat. *Journal of Neurophysiology*, *90*(3), 1418–1431.

Druckmann, S., Banitt, Y., Gidon, A., & Schürmann, F. (2007). A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Frontiers in Neuroscience*, *1*(1), 7–18.

Eggermont, J. J., Aertsen, A. M., & Johannesma, P. I. (1983). Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. *Hear. Res.*, *10*(2), 167–190.

Eichhorn, J., Tolias, A., Zien, A., Kuss, M., Rasmussen, C., Weston, J., et al. (2004). Prediction on spike data using kernel algorithms. In S. Thrün, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems*, *16* (pp. 1367–1374). Cambridge, MA: MIT Press.

Eyherabide, H., Rokem, A., Herz, A., & Samengo, I. (2009). Bursts generate a non-reducible spike-pattern code. *Front. Neurosci.*, *3*(1), 8–14.

Faisal, A., Selen, L., & Wolpert, D. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.

Gawne, T. J., Kjaer, T. W., & Richmond, B. J. (1996). Latency: Another potential code for feature binding in striate cortex. *Journal of Neurophysiology*, *76*(2), 1356–1360.

Gawne, T. J., McClurkin, J. W., Richmond, B. J., & Optican, L. M. (1991). Lateral geniculate neurons in behaving primates. III. Response predictions of a channel model with multiple spatial-to-temporal filters. *Journal of Neurophysiology*, *66*(3), 809–823.

Geisler, W. S., Albrecht, D. G., Salvi, R. J., & Saunders, S. S. (1991). Discrimination performance of single neurons: Rate and temporal-pattern information. *Journal of Neurophysiology*, *66*(1), 334–362.

Gerstein, G., & Kiang, N. (1960). An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophysical Journal*, *1*(1), 15–28.

Gerstner, W. (2000). Population dynamics of spiking neurons: Fast transients, asynchronous states, and locking? *Neural computation*, *12*, 43–89.

Gerstner, W. (2008). Spike-response model. *Scholarpedia*, *3*(12), 1343.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models*. Cambridge: Cambridge University Press.

Gerstner, W., & Naud, R. (2009). How good are neuron models. *Science*, *326*, 379–380.

Gerstner, W., van Hemmen, J., & Cowan, J. (1996). What matters in neuronal locking? *Neural Computation*, *8*, 1653–1676.

Gollisch, T., & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*, 1108–1111.

Houghton, C. (2009). Studying spike trains using a van rossum metric with a synapse-like filter. *J. Comput. Neurosci.*, *26*(1), 149–155.

Hunter, J. D., & Milton, J. G. (2003). Amplitude and frequency dependence of spike timing: implications for dynamic regulation. *Journal of Neurophysiology*, *90*(1), 387–394.

Izhikevich, E., Desai, N., Walcott, E., & Hoppensteadt, F. (2003). Bursts as a unit of neural information: Selective communication via resonance. *Trends in Neurosciences*, *26*(3), 161–167.

Jolivet, R., Kobayashi, R., Rauch, A., Naud, R., Shinomoto, S., & Gerstner, W. (2008). A benchmark test for a quantitative assessment of simple neuron models. *Journal of Neuroscience Methods*, *169*, 417–424.

Jolivet, R., Rauch, A., Lüscher, H., & Gerstner, W. (2006). Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of Computational Neuroscience*, *21*, 35–49.

Kayser, C., Montemurro, M., Logothetis, N., & Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, *61*(4), 597–608.

Kistler, W., Gerstner, W., & Hemmen, J. (1997). Reduction of the Hodgkin-Huxley equations to a single-variable threshold model. *Neural Computation*, *9*, 1015–1045.

Kobayashi, R., Tsubo, Y., & Shinomoto, S. (2009). Made-to-order spiking neuron model equipped with a multi-timescale adaptive threshold. *Frontiers in Computational Neuroscience*, *3*, 9.

Kreiman, G., Krahe, R., Metzner, W., Koch, C., & Gabbiani, F. (2000). Robustness and variability of neuronal coding by amplitude-sensitive afferents in the weakly electric fish eigenmannia. *Journal of Neurophysiology*, *84*(1), 189–204.

Kreuz, T., Chicharro, D., Andrzejak, R. G., Haas, J. S., & Abarbanel, H.D.I. (2009). Measuring multiple spike train synchrony. *J. Neurosci. Methods*, *183*(2), 287–299.

Kreuz, T., Haas, J., Morelli, A., Abarbanel, H., & Politi, A. (2007). Measuring spike train synchrony. *Journal of Neuroscience Methods*, *165*(1), 151–161.

La Camera, G, Rauch, A., Thurbon, D., Lüscher, H., Senn, W., & Fusi, S. (2006). Multiple time scales of temporal response in pyramidal and fast spiking cortical neurons. *Journal of Neurophysiology*, *96*, 3448–3464.

Lestienne, R. (1995). Determination of the precision of spike timing in the visual cortex of anaesthetised cats. *Biological Cybernetics*, *74*(1), 55–61.

Lindner, B. (2006). Superposition of many independent spike trains is generally not a Poisson process. *Phys. Rev. E*, *73*(2), 022,901.

Lisman, J. (1997). Bursts as a unit of neural information: Making unreliable synapses reliable. *Trends in Neurosciences*, *20*(1), 38–43.

Lisman, J. (2005). The theta/gamma discrete phase code occurring during the hippocampal phase precession may be a more general brain coding scheme. *Hippocampus*, *15*(7), 913–922.

Lundstrom, B., Higgs, M., Spain, W., & Fairhall, A. (2008). Fractional differentiation by neocortical pyramidal neurons. *Nature Neuroscience*, *11*, 1335–1342.

MacPherson, J. M., & Aldridge, J. W. (1979). A quantitative method of computer analysis of spike train data collected from behaving animals. *Brain Research*, *175*(1), 183–187.

Mainen, Z. F., & Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, *268*(5216), 1503–1506.

McCullagh, P., & Nelder, J. A. (1998). *Generalized linear models* (2nd ed.). London: Chapman & Hall/CRC.

Naud, R., Marcille, N., Clopath, C., & Gerstner, W. (2008). Firing patterns in the adaptive exponential integrate-and-fire model. *Biological Cybernetics*, *99*, 335–347.

Optican, L. M., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. 3. Information theoretic analysis. *J. Neurophysiol.*, *57*, 162–178.

Paiva, A.R.C, Park, I., & Príncipe, J. (2009a). A comparison of binless spike train measures. *Neural Computing and Applications*, *19*(3), 1–15.

Paiva, A.R.C, Park, I., & Príncipe, J. (2009b). A reproducing kernel Hilbert space framework for spike train signal processing. *Neural Computation*, *21*(2), 424–449.

Paiva, A.R.C, Park, I., & Príncipe, J. (2010). Inner products for representation and learning in the spike train domain. In K. G. Oweiss (Ed.), *Statistical signal processing for neuroscience and neurotechnology*. Orlando, FL: Academic Press.

Paninski, L., Pillow, J., & Simoncelli, E. (2005). Comparing integrate-and-fire models estimated using intracellular and extracellular data. *Neurocomputing*, *65–66*, 379–385.

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, *98*(3), 1064–1072.

Petersen, R. S., Brambilla, M., Bale, M. R., Alenda, A., Panzeri, S., Montemurro, M. A., et al . (2008). Diverse and temporally precise kinetic feature selectivity in the VPM thalamic nucleus. *Neuron*, *60*(5), 890–903.

Pillow, J., Paninski, L., Uzzell, V., Simoncelli, E., & Chichilnisky, E. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *Journal of Neuroscience*, *25*(47), 11,003.

Pillow, J., Shlens, J., Paninski, L., Sher, A., Litke, A., Chichilnisky, E., et al. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999.

Plesser, H. E., & Gerstner, W. (2000). Escape rate models for noisy integrate-and-fire neurons. *Neurocomputing*, *32–33*, 219–224.

Quiroga, R. Q., Kraskov, A., Kreuz, T., & Grassberger, P. (2002). Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys. Rev. E*, *65*(4), 041903.

Quiroga, R. Q., Kreuz, T., & Grassberger, P. (2002). Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys. Rev. E*, *66*(4), 041904.

Reich, D., Mechler, F., Purpura, K., & Victor, J. D. (2000). Interspike intervals, receptive fields, and information encoding in primary visual cortex. *Journal of Neuroscience*, *20*(5), 1964–1974.

Reich, D. S., Mechler, F., & Victor, J. D. (2001). Temporal coding of contrast in primary visual cortex: When, what, and why. *Journal of Neurophysiology*, *85*(3), 1039–1050.

Reich, D. S., Victor, J. D., Knight, B. W., Ozaki, T., & Kaplan, E. (1997). Response variability and timing precision of neuronal spike trains in vivo. *Journal of Neurophysiology*, *77*(5), 2836–2841.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1996). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Schoenberg, F. P., & Tranbarger, K. E. (2008). Description of earthquake aftershock sequences using prototype point patterns. *Environmetrics*, *19*, 271–286.

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

Schreiber, S., Fellous, J., Whitmer, D., Tiesinga, P., & Sejnowski, T. J. (2003). A new correlation-based measure of spike timing reliability. *Neurocomputing*, *52*(54), 925–931.

Shimazaki, H., & Shinomoto, S. (2007). A method for selecting the bin size of a time histogram. *Neural Computation*, *19*(6), 1503–1527.

Shinomoto, S. (2010). Fitting a stochastic spiking model to neuronal current injection data. *Neural Networks*, *23*(6), 764–769.

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

Snyder, D. L., Miller, M. I., & Snyder, D. L. (1991). *Random point processes in time and space* (2nd ed.). New York: Springer-Verlag.

Strong, S. P., Köberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, *80*, 197–200.

Tiesinga, P. H. E. (2004). Chaos-induced modulation of reliability boosts output firing rate in downstream cortical areas. *Physical Review E*, *69*(3 Pt. 1), 031912.

Tiesinga, P., Fellous, J., & Sejnowski, T. J. (2008). Regulation of spike timing in visual cortical circuits. *Nature Reviews Neuroscience*, *9*(2), 97–109.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, *7*, 399–407.

Truccolo, W., Eden, U., Fellows, M., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, *93*, 1074–1089.

Tsodyks, M., & Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl. Academy of Sci. USA*, *94*, 719–723.

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

Victor, J. D. (2005). Spike train metrics. *Current Opinion in Neurobiology*, *15*, 585–592.

Victor, J. D., & Purpura, K. (1996). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *Journal of Neurophysiology*, *76*(2), 1310–1326.

Victor, J. D., & Purpura, K. (1997). Metric-space analysis of spike trains: Theory, algorithms, and application. *Network: Comput. Neural Syst.*, *8*, 127–164.

Voinov, V. (1985). Unbiased estimation of powers of the inverse of mean and related problems. *Sankhyā: The Indian Journal of Statistics, Series B*, *47*(3), 354–364.

Voinov, V., & Nikulin, M. S. (1996). Unbiased estimators and their applications. *Mathematics and its applications*. Dordrecht: Kluwer.

Wang, L., Narayan, R., Grana, G., Shamir, M., & Sen, K. (2007). Cortical discrimination of complex natural stimuli: Can single neurons match behavior? *Journal of Neuroscience*, *27*(3), 582–589.

Received January 13, 2011; accepted June 2, 2011.