Published in final edited form as:

Neural Comput. 2015 January ; 27(1): 104–150. doi:10.1162/NECO_a_00681.

Spatio-temporal conditional inference and hypothesis tests for neural ensemble spiking precision

Matthew T. Harrison¹, Asohan Amarasingham², and Wilson Truccolo³

¹Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912

²Department of Mathematics, The City College of New York, and Departments of Biology and Psychology, The Graduate Center, The City University of New York, New York, New York 10031

³Department of Neuroscience and Brown Institute for Brain Science, Brown University, Providence, Rhode Island 02912

Abstract

The collective dynamics of neural ensembles create complex spike patterns with many spatial and temporal scales. Understanding the statistical structure of these patterns can help resolve fundamental questions about neural computation and neural dynamics. Spatio-temporal conditional inference (STCI) is introduced here as a semiparametric statistical framework for investigating the nature of precise spiking patterns from collections of neurons that is robust to arbitrarily complex and nonstationary coarse spiking dynamics. The main idea is to focus statistical modeling and inference, not on the full distribution of the data, but rather on families of conditional distributions of precise spiking given different types of coarse spiking. The framework is then used to develop families of hypothesis tests for probing the spatio-temporal precision of spiking patterns. Relationships among different conditional distributions are used to improve multiple hypothesis testing adjustments and to design novel Monte Carlo spike resampling algorithms. Of special note are algorithms that can locally jitter spike times while still preserving the instantaneous peri-stimulus time histogram (PSTH) or the instantaneous total spike count from a group of recorded neurons. The framework can also be used to test whether first-order maximum entropy models with possibly random and time-varying parameters can account for observed patterns of spiking. STCI provides a detailed example of the generic principle of conditional inference, which may be applicable in other areas of neurostatistical analysis.

Keywords

jitter; maximum entropy; network; resampling; spike timing; spike train; synchrony

1 Introduction

The spatio-temporal dynamics of multiple simultaneously recorded spike trains can reveal clues about the processes that underlie neural information processing (Abeles, 1982; Engel and Singer, 2001; Schneidman et al., 2006; Fujisawa et al., 2008; Pillow et al., 2008; Haider and McCormick, 2009; Truccolo et al., 2010), as well as properties of ongoing neural dynamics critical for understanding neurological disorders such as Parkinson's disease and epilepsy (Netoff and Schiff, 2002; Uhlhaas and Singer, 2006; Truccolo et al., 2011). These

clues are often statistical in nature, and there is growing demand for statistical models and methods that can appropriately handle the complexities and nonstationarities commonly observed in neural spiking data (Brody, 1998; Grün et al., 2002; Ventura et al., 2005b; Amarasingham et al., 2006; Churchland et al., 2007; Staude and Rotter, 2009; Louis et al., 2010; Cohen and Kohn, 2011; Amarasingham et al., 2012; Kelly and Kass, 2012). Amarasingham et al. (2012) argue that conditional inference is a useful statistical tool for focusing on precise spike timing in the presence of complex, nonstationary, coarse-temporal dynamics. They describe how many commonly-used spike train resampling algorithms, particularly those related to jitter (random, local perturbations of spike times), can be viewed as a type of hypothesis testing based on conditional inference. Here we introduce several extensions of jitter that can be used to further probe the nature of precise spiking. We also situate these extensions within an abstract framework, **spatio-temporal conditional inference (STCI)**, that can be used much more broadly to develop statistical models and methods for investigating precise spiking dynamics.

The spike train resampling algorithm for interval jitter (see Amarasingham et al., 2012, and the references therein) creates surrogate datasets by uniformly perturbing spikes within small intervals (or coarse time bins), separately and independently for each neuron. Interval jitter creates surrogate datasets that are similar on coarse time scales, but maximally random on fine time scales. The distinction between “fine” and “coarse” is roughly determined by the length of the jitter intervals. The techniques introduced here grew out of a desire to use interval jitter, but at the same time preserve certain additional information, such as the instantaneous peri-stimulus time histogram (PSTH) in datasets with multiple repeated trials or the instantaneous population (firing) rate (Okun et al., 2012) in datasets with multiple simultaneously recorded neurons, both of which would normally be blurred by interval jitter. While developing these resampling algorithms, which are presented below, we realized that there is a large family of similar algorithms that can be described within a common framework and that might have interesting neurostatistical applications. This paper describes that common framework and a few of the interesting resampling algorithms derived from it. Figure 1 shows surrogate data from some of these algorithms. Panel E shows interval jitter subject to also preserving the population rate (left) or PSTH (right).

Formally, STCI begins by factoring the distribution over spike trains into a term that captures coarse-precision spiking and another that captures fine-precision spiking given coarse-precision spiking:

$$\text{Prob}(\text{spiking}) = \text{Prob}(\text{coarse spiking}) \times \text{Prob}(\text{fine spiking} | \text{coarse spiking}). \quad (1)$$

The definition of “coarse spiking” is an important modeling choice and need not correspond to usual notions of coarseness. Once a definition of “coarse spiking” has been chosen, statistical modeling and inference focuses exclusively on the final factor in Equation (1), namely, the conditional distribution of fine spiking given coarse spiking. This is in contrast to the usual approach of basing statistical conclusions on the entire distribution of spiking. When appropriate, this refinement of focus provides striking statistical robustness by allowing any possible distribution for coarse spiking and by not having to infer the details of

this distribution from data. As such, STCI can be viewed as a type of semiparametric modeling and inference. The coarse-spiking factor is completely nonparametric and the fine-spiking factor is possibly parametric. Figure 2 provides a schematic illustration of why STCI can be sensible for scientific and statistical questions about fine-precision spiking.

STCI can be used for any type of statistical inference: parameter estimation, confidence intervals, hypothesis tests, etc. In this paper we only develop the details for a specific type of hypothesis test, namely, the null hypothesis that $\text{Prob}(\text{fine spiking} \mid \text{coarse spiking})$ is uniform, or maximum entropy. Loosely speaking, this null hypothesis corresponds to the idea that there is no additional precise spiking structure beyond what is implied by the observed coarse spiking structure. Algorithms for sampling from the null hypothesis correspond to spike train resampling algorithms, such as those illustrated in Figure 1. The features of the original data that are preserved in the surrogate data are exactly the “coarse spiking” features used in Equation (1).

2 Space-time bins and coarsened data

In this section we formally describe the family of allowable definitions of “coarse spiking” for STCI, along with some important special cases. This is certainly not the only useful family for conditional inference. For instance, both interval jitter and pattern jitter (Harrison and Geman, 2009) use Equation (1), but only interval jitter is an example of STCI. Nevertheless, STCI should provide a template for developing other interesting uses of Equation (1).

The main idea is that “coarse spiking” is essentially a blurring, or averaging, of the spike train, but that this blurring need not be a pure temporal blurring (as in interval jitter or Figure 2), but can involve “spatial” (for lack of a better term) blurring over subgroups of spike trains. Furthermore, the blurring need not be contiguous in either space or time, but can, for instance, involve blurring within distinct neuronal subtypes that are not spatially proximate or blurring within distinct epochs of temporally separated trials. An important, but subtle point is that we allow “coarse spiking” to refer simultaneously to multiple distinct types of blurring; cf., Figures 1E and 3E. This allows us to flexibly probe the statistical dynamics of precise spiking in novel ways.

For concreteness we consider only datasets that are arranged like those in Figure 1A — neuron $\times$ time or trial $\times$ time — but this concreteness is easily relaxed. Figure 3 contains schematic illustrations of the key concepts. Define the **fine-precision data** to be

$\mathbf{X} \stackrel{\text{def}}{=} (X_{it}; i=1, \dots, N; t=1, \dots, T)$, where X_{it} is the number of spikes from spike train i in time bin t . We assume that the time bins are all the same size, that they are a non-overlapping partition of time, that each spike is assigned to exactly one time bin, and that they are small enough so that there is at most one spike in each time bin, making $\mathbf{X}$ an $N \times T$ binary matrix.

Consider the set of possible neuron-label / time-bin pairs, or equivalently, row-index / column-index pairs for the fine-precision data, namely,

$\mathcal{F} \stackrel{\text{def}}{=} \{(i, t); i=1, \dots, N; t=1, \dots, T\}$. A **space-time bin**, say ω , is any nonempty subset of

$\mathcal{F}$. For example, we could consider the space-time bin $\omega = \{(1, 8), (1, 9), (1, 10), (1, 11), (1, 12), (2, 9), (2, 10), (2, 11)\}$ which includes time bins 8–11 and neurons 1 and 2, but not all combinations of these (cf., ω_5 in Figure 3). In general, space-time bins need not be contiguous, so space-time bins such as $\omega = \{(3, 5), (11, 2)\}$ are perfectly acceptable. The only space-time bins used in the fine-precision data are the singleton space-time bins of the form $\omega = \{(i, t)\}$, which we call the **fine space-time bins**. Although we are using spatial terminology to refer to the row-indices in $\mathbf{X}$, these need not index space in any meaningful way. If the rows correspond to repeated trials separated in time, then the row-indices are themselves temporal in nature.

For any space-time bin ω we use $\mathbf{X}_\omega$ to denote the pattern of fine-precision spiking within ω , that is,

$$\mathbf{X}_\omega \stackrel{\text{def}}{=} (X_{it} : (i, t) \in \omega),$$

and we can summarize this pattern of spiking with the total spike count in ω via

$$C_\omega \stackrel{\text{def}}{=} \sum_{(i,t) \in \omega} X_{it}.$$

This summary obscures the pattern of spiking in $\mathbf{X}_\omega$, in effect, uniformly blurring the fine-precision data over ω . (Compare $\mathbf{X}_{\omega_5}$ with C_{ω_5} in Figure 3.)

A **space-time binning**, say Ω , is any collection of space-time bins that completely covers $\mathcal{F}$, namely, $\Omega = \{\omega_1, \dots, \omega_m\}$ such that $\cup_i \omega_i = \mathcal{F}$. A **space-time partition** is a space-time binning that is also a partition of $\mathcal{F}$, meaning that any two space-time bins in the binning, Ω , are disjoint, i.e., $\omega, \omega' \in \Omega$ with $\omega \neq \omega'$ implies $\omega \cap \omega' = \emptyset$. If Ω is a space-time partition, then $\mathbf{X}$ is equivalent to $(\mathbf{X}_\omega : \omega \in \Omega)$.

For any fixed space-time binning, Ω , we define the **coarsened data**

$$\mathbf{C}^\Omega \stackrel{\text{def}}{=} (C_\omega : \omega \in \Omega),$$

which contains the spike-count summaries of $\mathbf{X}$ for every space-time bin in Ω . If Ω is a space-time partition, then the coarsened data is simply a blurring of the fine precision data over the space-time bins in Ω (Figure 3B). If we are given a dataset $\mathbf{x}$ based on fine space-time bins, then we use $C_\omega(\mathbf{x})$ and $\mathbf{C}^\Omega(\mathbf{x})$ to denote its summary over ω and its **coarsening** or **coarsened version** over Ω , respectively. Sometimes it is useful to emphasize that $C_\omega(\mathbf{x})$ only depends on $\mathbf{x}_\omega$ by writing $C_\omega(\mathbf{x}_\omega)$.

This paper heavily concentrates on a few simple classes of space-time binnings and the corresponding coarsenings. For fixed positive integers a and b , define the set of temporal indices

$$W_b^\Delta \stackrel{\text{def}}{=} \{(b-1)\Delta+1, \dots, b\Delta\},$$

which is a set of Δ consecutive indices ending at $b\Delta$. Define the space-time partition

$$\text{TIME-}\Delta \stackrel{\text{def}}{=} \{\{i\} \times W_b^\Delta : i=1, \dots, N; b=1, \dots, T^\Delta\},$$

which are the space-time bins consisting of Δ -length disjoint blocks of fine space-time bins for individual neurons. $T^\Delta \stackrel{\text{def}}{=} T/\Delta$ is the number of these blocks, which we conveniently assume to be an integer. (This assumption can be easily relaxed, for example, by making the final block shorter.) For example, $\mathbf{C}^{\text{TIME-}\Delta}$ is identically $\mathbf{X}$ and $\mathbf{C}^{\text{TIME-}1}$ is exactly what one would obtain instead of $\mathbf{X}$ if the original fine space-time bins were Δ times longer (in time). $\mathbf{C}^{\text{TIME-}1}$ is a pure temporal coarsening in that the label of each spike is preserved, but the temporal resolution is blurred (Figure 3C). It is the temporal coarsening used by interval jitter. $\mathbf{C}^{\text{TIME-}T}$ is equivalent to the row sums of $\mathbf{X}$.

Similarly, for a fixed positive integer Γ define the space-time partition

$$\text{ALL-}\Gamma \stackrel{\text{def}}{=} \{\{1, \dots, N\} \times W_b^\Gamma : b=1, \dots, T^\Gamma\},$$

which are the space-time bins consisting of Γ -length disjoint blocks of fine space-time bins for all neurons simultaneously. $\mathbf{C}^{\text{ALL-}\Gamma}$ only contains the total number of spikes in each of these blocks summed over all neurons. $\mathbf{C}^{\text{ALL-}1}$ is a pure spatial coarsening in that the precise time of each spike is preserved, but the label of each spike is removed. $\mathbf{C}^{\text{ALL-}1}$ is equivalent to the column sums of $\mathbf{X}$. For $\Gamma > 1$, $\mathbf{C}^{\text{ALL-}\Gamma}$ is a spatio-temporal coarsening, blurring both the temporal resolution of spikes and obscuring the spike label (Figure 3D). $\mathbf{C}^{\text{ALL-}T}$ is a single number reporting the total number of spikes in $\mathbf{X}$. Okun et al. (2012) refers to $\mathbf{C}^{\text{ALL-}\Gamma}$ as the “population rate” using time bins of length Γ . If the columns of $\mathbf{X}$ index repeated trials instead of distinct neurons, then $\mathbf{C}^{\text{ALL-}\Gamma}$ is the PSTH in Γ -length bins.

We can combine $\text{TIME-}\Delta$ and $\text{ALL-}\Gamma$ to create the space-time binning

$$\text{TA-}(\Delta, \Gamma) \stackrel{\text{def}}{=} \text{TIME-}\Delta \cup \text{ALL-}\Gamma,$$

which is not a partition. The coarsening $\mathbf{C}^{\text{TA-}(\Delta, \Gamma)}$ simultaneously provides the information in $\mathbf{C}^{\text{TIME-}\Delta}$ and $\mathbf{C}^{\text{ALL-}\Gamma}$ (Figure 3E). For example, the special case of $\mathbf{C}^{\text{TA-}(T, 1)}$ gives the row sums and the columns sums of $\mathbf{X}$. Algorithms for working with $\text{TA-}(\Delta, \Gamma)$ are an important contribution of this paper. These algorithms build upon recent innovations for sampling uniformly from the set of binary matrices with specified row and column sums (Miller and Harrison, 2013; Harrison and Miller, 2013).

Finally, we remark that there is nothing special about using a two-dimensional matrix representation for $\mathbf{X}$ and the fine space-time bins. More generally, we require a collection of finest resolution bins $\mathcal{F}$ (which can be indexed in any way that happens to be experimentally relevant), the original binary data $\mathbf{X}$ recorded in these fine bins, and a collection of subsets Ω that covers $\mathcal{F}$, along with the corresponding coarsening $\mathbf{C}^\Omega$.

3 Resampling algorithms that preserve coarsened data

Fix a space-time binning Ω and coarsened data $\mathbf{c} = \mathbf{C}(\mathbf{X}) = \mathbf{C}^\Omega(\mathbf{X})$ for some observed data $\mathbf{X}$. We are interested in algorithms that can create surrogate datasets that have the same coarsening $\mathbf{c}$, but that are otherwise maximally random. In particular, define the conditional probability mass function (pmf)

$$f_0(\mathbf{x}|\mathbf{c}) \stackrel{\text{def}}{=} \frac{1}{\eta(\mathbf{c})} \mathbb{1}\{\mathbf{C}(\mathbf{x})=\mathbf{c}\}, \quad (2)$$

for all $\mathbf{x}$ and $\mathbf{c}$, where $\eta(\mathbf{c})$ is a normalization constant chosen so that each $f_0(\cdot|\mathbf{c})$ is a valid pmf, namely,

$$\eta(\mathbf{c}) \stackrel{\text{def}}{=} \#\{\mathbf{x}:\mathbf{C}(\mathbf{x})=\mathbf{c}\}.$$

In words, $f_0(\cdot|\mathbf{c})$ is the uniform (or maximum entropy) distribution over all possible spike trains that have the same Ω -coarsening as $\mathbf{X}$. We are interested in algorithms that can sample from $f_0(\cdot|\mathbf{c})$. Sometime we write f_0^Ω to emphasize the dependence on Ω .

If Ω is a space-time partition, such as TIME- or ALL- Γ , then generating a random observation $\mathbf{Y}$ from $f_0(\cdot|\mathbf{c})$ is easy. In short, spikes are perturbed uniformly (or jittered) throughout their respective space-time bins (multiple spikes in the same fine space-time bin are prohibited). This procedure preserves the coarsening, but is otherwise maximally random. More formally, to generate $\mathbf{Y}_\omega$ we place c_ω spikes uniformly at random in the fine

space-time bins contained in ω . If there are $|\omega|$ bins in ω , then there are $\binom{|\omega|}{c_\omega}$ distinct ways

to choose $\mathbf{Y}_\omega$, where $\binom{a}{b} \stackrel{\text{def}}{=} a!/((b!(a-b)!))$. Since the space-time bins do not overlap, we independently repeat this procedure for each $\omega \in \Omega$ until $\mathbf{Y}$ is fully constructed; see Corollary 8 in the Monte Carlo Appendix. The number of distinct $\mathbf{Y}$ is

$$\eta(\mathbf{c}) = \prod_{\omega \in \Omega} \binom{|\omega|}{c_\omega}.$$

This formula is only valid for space-time partitions, not arbitrary space-time binnings. To generate multiple observations, say $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$, we independently repeat this process n times.

If Ω is not a partition, meaning that some of the space-time bins overlap, then the situation is much more complicated. Generating a random observation from $f_0(\cdot|c)$ may not be practical. In these cases, it may be possible to use more sophisticated techniques, such as importance sampling or Markov chain Monte Carlo (MCMC). Further discussion of these approaches can be found in the Monte Carlo Appendix.

The only non-partition space-time binning that we use in this paper is TA- $(\cdot, \Gamma)$ for $\cdot = k\Gamma$ for some integer $k > 1$. The Monte Carlo Appendix describes algorithms for generating resamples for this space-time binning, and Figure 1E shows some resampled data. These algorithms are among the key contributions of the paper. Our current software can efficiently generate exact uniform samples for the special case of $\Gamma = 1$, particularly for small $\cdot$ or for datasets with sparse spiking. The case $\Gamma = 1$ includes as a subproblem the ability to generate uniform samples from the set of binary matrices with specified row and column sums, which is a celebrated problem in statistics, combinatorics, and theoretical computer science. The case of general Γ is even more challenging, and we have developed importance sampling software that has excellent performance.

4 Spatio-temporal conditional inference

The remainder of the paper develops a statistical framework for using the resampling algorithms to carefully probe the nature of precise spike timing. In this section we draw connections to conditional inference and in Section 5 we specialize to conditional hypothesis testing. The resampling algorithms above can be used to construct p-values for these hypothesis tests. In Section 6 we present a sequential hypothesis testing framework that, in a certain sense, allows the data to determine the type of coarsening. And in Section 8 we draw connections to parametric log-linear models, or maximum entropy models, which are popular contemporary models for spike train data.

Suppose p is a pmf for the data $\mathbf{X}$. Then for any space-time binning Ω and the corresponding coarsening $\mathbf{C} = \mathbf{C}^\Omega$ we can always factor p as

$$\begin{aligned} p(\mathbf{x}) &\stackrel{\text{def}}{=} \mathbb{P}(\mathbf{X}=\mathbf{x}) = \mathbb{P}(\mathbf{X}=\mathbf{x}, \mathbf{C}(\mathbf{X})=\mathbf{C}(\mathbf{x})) \\ &= \underbrace{\mathbb{P}(\mathbf{C}(\mathbf{X})=\mathbf{C}(\mathbf{x}))}_{\text{coarse spiking } (g)} \underbrace{\mathbb{P}(\mathbf{X}=\mathbf{x}|\mathbf{C}(\mathbf{X})=\mathbf{C}(\mathbf{x}))}_{\text{precise spiking given coarse } (f)} \quad (3) \\ &\stackrel{\text{def}}{=} g(\mathbf{C}(\mathbf{x}))f(\mathbf{x}|\mathbf{C}(\mathbf{x})), \end{aligned}$$

where $\mathbb{P}$ denotes probability. The second equality holds because $\mathbf{C}$ is a function of $\mathbf{X}$. The function g is the pmf for the coarsened data $\mathbf{C}$ and the function f is the conditional pmf for the fine data $\mathbf{X}$ given the coarse data $\mathbf{C}$. These functions can be derived directly from p via

$$g(c) = \sum_{\mathbf{x}: C(\mathbf{x})=c} p(\mathbf{x}) \quad \text{and} \quad f(\mathbf{x}|c) = \frac{p(\mathbf{x})}{g(c)} \mathbb{1}\{C(\mathbf{x})=c\}, \quad (4)$$

where we take $0/0 = 0$ and where $\mathbb{1}\{A\}$ denotes the indicator function of the event A . Both f and g depend on the choice of Ω , whereas p does not. Sometimes we write f^Ω and g^Ω for clarity.

For real data, p is not known and we must use statistical techniques to learn about p from the data. Our uncertainty about p extends to uncertainty about both f and g , and in many cases we must use the data to resolve uncertainty about both. There are certain scientific questions, however, that pertain primarily to one of f or g , but not both. For questions about precise spiking, it is particularly useful to focus only on f for an appropriately chosen space-time binning. If we restrict statistical inference to f only, which is called **conditional inference** (Reid, 1995), then we can, in essence, allow g to be completely nonparametric and also avoid estimating it from data altogether. The resulting inferences will be robust to the distribution of the coarsened data, including distributions that have nonstationarity, trial-to-trial variability, complex dependencies, etc. Note that the choice of Ω is a modeling choice, just like any other, and should be based on sensible neuroscientific assumptions.

This robustness of STCI to complexities in the distribution of the coarsened data is one of its strong benefits. There is a growing recognition among investigators that coarse-temporal spiking dynamics result from many unknown or unobserved influences, such as signals originating in sensory or feedback systems, or changes in the extracellular environment, that might vary irregularly in time and across trials and that could affect multiple neurons in unanticipated ways (Brody, 1998; Grün et al., 2002; Ventura et al., 2005b; Amarasingham et al., 2006; Churchland et al., 2007; Staude and Rotter, 2009; Louis et al., 2010; Cohen and Kohn, 2011; Amarasingham et al., 2012; Kelly and Kass, 2012). For some applications, STCI provides an attractive method for building models that can allow for this complexity.

STCI can be used for general statistical inference. In the remainder of this paper, however, we focus on a specific application that connects directly to the resampling algorithms developed in the previous section. In particular, we develop tools for testing the null hypothesis that f is uniform and for using the data to find Ω for which this null hypothesis is true.

5 Testing for fine-precision spiking

Fix a space-time binning Ω of interest. Beginning with Equation (3), we want to test the null hypothesis that $f = f_0$, where f_0 is the uniform conditional pmf from Equation (2):

$$H_0: f(\mathbf{x}|c) = f_0(\mathbf{x}|c) \quad (5)$$

for all $\mathbf{x}$ and c . For clarity, we sometimes write H_0^Ω .

For small time bins, so that $\mathbf{X}$ has binary entries, the null hypothesis H_0 captures the intuitive notation that the coarsened data $\mathbf{C}$ contains all of the interesting spiking dynamics. Under the

null, the conditional arrangement of spikes given the coarsened data is the uniform distribution, which is maximally random and structureless. (Technically speaking, H_0 seems most appropriate when Ω has a subset that is itself a space-time partition, as is the case with all of the examples here.) For the special case of a pure temporal coarsening, Amarasingham et al. (2012) argue that the corresponding null hypothesis, $H_0^{\text{TIME}-\Delta}$, is a reasonable statistical formulation of the scientific hypothesis that there is no precise spike timing, which they call the **interval jitter** null hypothesis. The parameter Δ controls the qualitative distinction between coarse and fine time scales.

For a given test statistic S , a p-value for the null hypothesis is the fraction of those datasets with the same coarsening as the original data that also give a value of the test statistic at least as extreme as the original data. Let $\mathbf{Y}$ be a random dataset that satisfies the null hypothesis H_0 . Formally, the p-value is $\alpha(\mathbf{X}) = \alpha^{\Omega, S}(\mathbf{X})$ defined by

$$\begin{aligned}\alpha(\mathbf{x}) &\stackrel{\text{def}}{=} \mathbb{P}(S(\mathbf{Y}) \geq S(\mathbf{x}) | \mathbf{C}(\mathbf{Y}) = \mathbf{C}(\mathbf{x})) \\ &= \sum_{\mathbf{y}} f_0(\mathbf{y} | \mathbf{C}(\mathbf{x})) \mathbb{1}\{S(\mathbf{y}) \geq S(\mathbf{x})\} \\ &= \frac{\#\{\mathbf{y}: \mathbf{C}(\mathbf{y}) = \mathbf{C}(\mathbf{x}) \text{ and } S(\mathbf{y}) \geq S(\mathbf{x})\}}{\#\{\mathbf{y}: \mathbf{C}(\mathbf{y}) = \mathbf{C}(\mathbf{x})\}}.\end{aligned}\quad (6)$$

The choice of S depends on the experimental context; see Section 9. Notice that g , the distribution of $\mathbf{C}$, never appears in this hypothesis test. It does not need to be modeled or estimated from the data. This striking invariance to coarse spiking is one of the key features of the conditional inference approach.

For some special choices of S and Ω the value of $\alpha(\mathbf{x})$ can be computed exactly (Amarasingham, 2004; Harrison, 2005, 2013). More commonly, the value of $\alpha(\mathbf{x})$ will need to be approximated. Suppose that we have an algorithm that can generate independent and identically distributed (iid) observations, $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$, from $f_0(\cdot | \mathbf{C}(\mathbf{x}))$. In other words, we can obtain a random sample from the conditional distribution of the fine precision data given the observed coarse precision data under the null hypothesis that the conditional distribution is uniform. A simple Monte Carlo p-value is

$$\hat{\alpha}(\mathbf{x}) \stackrel{\text{def}}{=} \hat{\alpha}_n^{\Omega, S}(\mathbf{x}; \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}) \stackrel{\text{def}}{=} \frac{1 + \sum_{k=1}^n \mathbb{1}\{S(\mathbf{Y}^{(k)}) \geq S(\mathbf{x})\}}{1 + n}.\quad (7)$$

This is the fraction of the Monte Carlo random sample, including the observation $\mathbf{x}$, for which the test statistic is at least as large as it is for $\mathbf{x}$. For large n , $\hat{\alpha}(\mathbf{x})$ will usually be an excellent approximation of $\alpha(\mathbf{x})$. Furthermore, $\hat{\alpha}(\hat{\mathbf{X}})$ is itself a valid p-value, meaning that it can be used to construct hypothesis tests that appropriately bound the type I error rate. Formally, if H_0 is true and $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}$ is a random sample from $f_0(\cdot | \mathbf{C}(\mathbf{X}))$, then

$$\mathbb{P}(\hat{\alpha}(\mathbf{X}; \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}) \leq \varepsilon) \leq \varepsilon$$

for all $\varepsilon > 0$; see Davison and Hinkley (1997); Harrison (2012, 2013). The Monte Carlo Appendix describes how to construct p-values for more complex Monte Carlo sampling schemes.

Example 1a: Temporal precision and pairwise synchrony — exact p-values—

Let the fine precision data, $\mathbf{X}$, be the bottom two spike trains in Figure 4 (the two neurons with the highest firing rates; roughly 10 spikes/s each) over a 2 minute observation interval centered at seizure onset using 1 ms duration fine space-time bins (so that $N = 2$ and $T = 120000$). We will use the space-time partition TIME-100 , which corresponds to interval jitter, and we will use test statistic $S(\mathbf{x}) \stackrel{\text{def}}{=} \sum_t x_{1t}x_{2t}$, which counts the number of precise (1 ms) spike alignments between the two spike trains. In this case, there are $S(\mathbf{X}) = 57$ alignments. When $\tau = 100$ ms, there are $n(\mathbf{C}(\mathbf{X})) = 4.842 \times 10^{4058}$ different possible arrangements of spikes that preserve the original observed coarsening. Of these, 1.346×10^{4057} have 57 or more alignments, giving a p-value of $\alpha(\mathbf{X}) = (1.346 \times 10^{4057}) / (4.842 \times 10^{4058}) = 0.0278$ from Equation (6). See Harrison (2013) for details about how these numbers were computed exactly. If we were trying to control the type I error rate at level 0.05, then we could safely reject the null hypothesis.

The scientific interpretation of a rejection of $H_0^{\text{TIME-100}}$ is that the spike trains appear to have structure on time scales of roughly 100 ms or less. Even though we used a test statistic that measured 1 ms precision synchrony, we do not conclude that spikes can be precisely timed to 1 ms. The primary interpretation of a hypothesis test is driven by the null hypothesis, not the test statistic. For STCI, the precision being tested is the precision of the space-time binning that was used to define the coarsened data — in this case, a pure temporal coarsening with 100 ms bins. To further illustrate this important point, we can reduce τ to 25 ms and test $H_0^{\text{TIME-25}}$ using the same test statistic. Now there are 3.835×10^{3122} possible arrangements that preserve the coarsening (fewer arrangements than before, because the finer coarsening imposes more constraints), of which 2.534×10^{3122} have 57 or more precise alignments, giving a p-value of 0.661. The 25 ms temporally coarsened data can easily explain the occurrence of 57 precise synchronies without needing to postulate the existence of even more precise spike timing. Consequently, we do not have enough evidence to conclude that spikes are timed with precision below 25 ms. Since we tested two different null hypotheses in this example, we should control for multiple tests; see Section 6.3.

Example 1b: Temporal precision and pairwise synchrony — Monte Carlo p-values—Here we continue Example 1a, but use Monte Carlo methods to approximate the p-values. This is not necessary, since we know the exact p-values, but we include it for comparison. We used $n = 10000$ and obtained Monte Carlo p-values of $\hat{\alpha}^{\text{TIME-100}}(\mathbf{X}) = 0.0265$ and $\hat{\alpha}^{\text{TIME-25}}(\mathbf{X}) = 0.654$ using Equation (7), closely matching the target p-values that were computed exactly in Example 1a. Figure 5 has more details.

Example 2a: Temporal precision and pairwise synchrony — all pairs—This example is similar to Example 1a–b, except that we will use all $N = 84$ neurons and use the test statistic

$$S(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{\binom{N}{2}} \sum_{1 \leq i < j \leq N} \frac{\sum_t (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)}{\sqrt{\sum_t (x_{it} - \bar{x}_i)^2} \sqrt{\sum_t (x_{jt} - \bar{x}_j)^2}}, \text{ where } \bar{x}_i \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T x_{it},$$

which is the average over all pairs of spike trains of the correlation coefficient (Pearson's r) between the pair of spike trains. (It is not our intention to suggest S as a good or even appropriate test statistic for some purpose. S was chosen for this example because it is simple and easily recognizable as a measure of precise correlation across the network.) We know of no efficient way to compute the p-value exactly. But it is trivial to compute a Monte Carlo p-value. Analogous to Example 1b, we used a Monte Carlo sample size of $n = 10000$ for each of $\Omega = \text{TIME-100}$ and $\Omega = \text{TIME-25}$, obtaining p-values of $\alpha(\mathbf{X}) = 0.0001$ in each case. (In each case the original dataset gave the most extreme test statistic among all of the resampled datasets.) Unlike Example 1a–b, now we can conclude that the pattern of spiking has structure on time scales of 25 ms or less.

Example 2b: Shared influences on precision—Example 2a suggested sub-25 ms spiking structure in the seizure data, in the sense that the conditional distribution $f^{\text{TIME-25}}(\cdot | \mathbf{C}(\mathbf{X}))$ is not uniform. But how is it structured? One explanation for this structure, particularly in light of the strong vertical banding observed in the rasters (Figure 4), is that the firing rates of all neurons in this recorded subpopulation are influenced by a common cause, perhaps a shared excitation or release of inhibition, or perhaps the collective dynamics of the local network create a rapid, shared coordination of spiking across all neurons. If so, this common cause may manifest itself within $f^{\text{TIME-25}}$ primarily by additionally structuring the total number of spikes (i.e., the population rate) on sub-25 ms time scales, irrespective of neuron identity. In other words, it may primarily influence $\mathbf{C}^{\text{ALL-}\Gamma}$ for some $\Gamma < 25$, in addition to the structure already present in $\mathbf{C}^{\text{TIME-25}}$. As described in more detail in Sections 6–7 we can use $H_0^{\text{TA}-(\Delta, \Gamma)}$ to test this conjecture. (Our experimental questions do not correspond to tests of $H_0^{\text{ALL-}\Gamma}$, which we can easily reject for the seizure data, because $H_0^{\text{ALL-}\Gamma}$ does not account for the heterogeneity in spiking across neurons.)

$H_0^{\text{TA}-(\Delta, \Gamma)}$ for $1 < \Gamma < 25$ has complicated structure, and it is difficult to create efficient algorithms for generating Monte Carlo observations. The Monte Carlo Appendix describes an algorithm that generates Monte Carlo observations from a more tractable probability distribution, say Q , that is close to the target uniform distribution. These observations can be appropriately reweighted according to Equation (22) to provide accurate p-value approximations $\hat{\alpha}_Q$ that are themselves valid p-values (as was the case for $\hat{\alpha}$). We used this importance sampling procedure with $\Delta = 25$ ms and $\Gamma = 5$ ms to generate $n = 10000$ observations from Q , obtaining a Monte Carlo p-value of $\hat{\alpha}_Q(\mathbf{X}) = 0.132$. Once we have accounted for the individual 25 ms coarse-temporal firing rates ($\mathbf{C}^{\text{TIME-25}}$) and the shared 5 ms fine-temporal population rate ($\mathbf{C}^{\text{ALL-5}}$), then the spiking patterns appear otherwise structureless, at least from the perspective of our precise correlation test statistic.

6 Exploring different types of fine-precision spiking

In this section we partially address the issue of using the data to determine the space-time binning Ω . For example, suppose we want to use the data to determine the jitter window width for interval jitter, which, in our notation, corresponds to choosing the value of Δ for TIME- Δ . The first thing to note is that there is not a “correct” or “true” value of Δ without additional assumptions, since the decomposition in Equation (3) is valid for any Δ , i.e., valid for any $C = C^{\text{TIME-}\Delta}$. If we assume that f is uniform, however, then Equation (3) will only be valid for certain Δ , depending on the distribution of the data, and we can use the data to learn which Δ do and do not make the assumption valid. We are, in effect, using the data to discover the precision of neural spiking.

The set-up is again hypothesis testing, but instead of testing a single null hypothesis, say $H_0^{\text{TIME-}\Delta}$, we are testing a family of null hypotheses, say $(H_0^{\text{TIME-}\Delta} : \Delta = 2, 4, 8, \dots, 64)$, in order to determine which ones are true and false. The purpose of this section is to describe a sequential testing procedure that is much more powerful than individually testing all of the null hypotheses and generically controlling for multiple comparisons. We introduce the key terminology in Section 6.1, describe the procedure in Section 6.2, and address multiple testing in Section 6.3. Returning to the interval jitter illustration, we are not trying to find the right Δ so that we can then use interval jitter with this Δ , but rather, we are using interval jitter with many different Δ to determine which Δ seem to match the data and which do not.

6.1 Refinements of space-time binnings

Consider two space-time binnings Ω and Ω' . We say that Ω' is a **refinement** of Ω if every space-time bin in Ω can be decomposed as a partition of space-time bins from Ω' . (Formally, for each $\omega \in \Omega$, there exists $k \geq 1$ and $\omega'_1, \dots, \omega'_k \in \Omega'$ such that $\bigcup_{j=1}^k \omega'_j = \omega$ and such that $\omega'_j \cap \omega'_\ell = \emptyset$ for $j \neq \ell$.) The simplest case of a refinement is when $\Omega \subseteq \Omega'$, because then for each space-time bin in Ω , we can simply choose the identical bin in Ω' (i.e., the trivial partition into a single piece).

The following relationships hold among our special space-time binnings. Let $j, k \geq 1$ be positive integers. ALL- Γ is a refinement of ALL- $j\Gamma$. TIME- Δ is a refinement of TIME- j , of ALL- k , and of TA- (j, k) . TA- (j, Γ) is a refinement of TIME- j , of ALL- $k\Gamma$, and of TA- $(j, k\Gamma)$.

If Ω' is a refinement of Ω , then Ω' has smaller space-time bins. If the null hypothesis H_0^Ω is true, meaning that the distribution of the data appears structureless within the space-time bins of Ω , then it should also be structureless in the smaller bins of Ω' . In particular, $H_0^{\Omega'}$ should be true. This reasoning is correct as the next theorem shows.

Theorem 1: Suppose that Ω and Ω' are space-time binnings, Ω' is a refinement of Ω , and H_0^Ω is true. Then $H_0^{\Omega'}$ is true.

Corollary 2: Suppose that Ω and Ω'' are arbitrary space-time binnings (one need not be a refinement of the other) and H_0^Ω is true. Then $H_0^{\Omega \cup \Omega''}$ is true.

A proof of Theorem 1 can be found in Lemma 10 in the Mathematical Appendix. Corollary 2 follows from the Theorem by defining $\Omega' = \Omega \cup \Omega''$, which is always a refinement of Ω , since $\Omega \subseteq \Omega \cup \Omega''$. These results are particularly useful in the next section where we discuss successive testing.

When testing H_0^Ω , if we fail to reject, then there is no need to further test $H_0^{\Omega'}$ for any refinement Ω' of Ω . We can invoke Theorem 1 and fail to reject $H_0^{\Omega'}$ as well. This does not mean that the p-values of a test of $H_0^{\Omega'}$ will be larger than those of H_0^Ω , although in practice this is often the case. The tests may have different power, and it is possible that an actual test of $H_0^{\Omega'}$ would lead to a rejection even though a test of H_0^Ω did not. This is rare, however, and it is usually not worth the penalty of performing an additional test (see the discussion of multiple testing adjustments in Section 9). Similarly, if we reject $H_0^{\Omega'}$, then we can invoke Theorem 1 and reject H_0^Ω for all those Ω for which Ω' is a refinement of Ω . We do not need to perform the tests nor to correct for them with multiple testing adjustments. These additional rejections do not come from multiple tests, but from the logical structure of the null hypotheses as described in Theorem 1. Figure 6 illustrates the logical structure of the null hypotheses.

6.2 Testing successive refinements

Consider testing H_0^Ω . If we fail to reject, then there is no need to test $H_0^{\Omega'}$ for any refinement Ω' of Ω (Theorem 1). The data appear structureless beyond the structure present in $\mathcal{C}^\Omega$ and there is no need to continue looking for even finer structure. On the other hand, if we reject H_0^Ω , suggesting that there is finer structure in the data, we may wish to explore further and test $H_0^{\Omega'}$ for some refinement Ω' of Ω . This section discusses some equivalent ways of thinking about this type of successive testing.

The most direct way to think about testing successive refinements is as follows. We have a sequence of successively finer space-time binnings $\Omega_1, \Omega_2, \dots, \Omega_r$, where each Ω_{j+1} is a refinement of Ω_j . Beginning with Ω_1 , which is the coarsest binning of interest, we sequentially test each $H_0^{\Omega_j}$ until we fail to reject, and then we stop. If Ω_r is sufficiently fine — for example, if each $\{(i, t)\} \in \Omega_r$, so that $H_0^{\Omega_r}$ is trivially true — then we will reject when $j = r$, if not sooner. The last j for which we reject, say j^* , reveals the finest structure that we were able to detect. It is this rejection of $H_0^{\Omega_{j^*}}$ that drives the experimental conclusions. The rejections of $H_0^{\Omega_1}, \dots, H_0^{\Omega_{j^*-1}}$ are now logically implied by the rejection of $H_0^{\Omega_{j^*}}$ (Theorem 1) and do not add any additional scientific or statistical information.

This method of successive testing — starting with the coarsest binning and stopping with the first failure to reject — automatically controls for multiple tests, in that the sequence of null

hypotheses, $H_0^{\Omega_1}, \dots, H_0^{\Omega_r}$, behaves like a single hypothesis test. Section 6.3 discusses this feature in more detail.

Another equivalent way to think about testing successive refinements applies the conditional inference perspective in Equation (3) to successive conditional distributions of p . This was the rationale that we used to motivate Example 2b above. As before, we begin with the coarsest binning of interest, Ω^1 . If we reject $H_0^{\Omega_1}$, then we believe f^{Ω_1} is not uniform. To further probe how f^{Ω_1} is not uniform, we can choose a new binning Ω'' and apply the same style of factorization from Equation (3) to $f^{\Omega_1}(\cdot|C^{\Omega_1})$ rather than to $p(\cdot)$. Formally,

$$\begin{aligned} f^{\Omega_1}(x|C^{\Omega_1}) &\stackrel{\text{def}}{=} \mathbb{P}(X=x|C^{\Omega_1}(X)=C^{\Omega_1}(x)) \stackrel{\text{def}}{=} \mathbb{P}(X|C^{\Omega_1}) \\ &= \mathbb{P}(X, C^{\Omega''}|C^{\Omega_1}) = \underbrace{\mathbb{P}(C^{\Omega''}|C^{\Omega_1})}_{\substack{\text{new coarse} \\ \text{spiking in } f^{\Omega_1}}} \underbrace{\mathbb{P}(X|C^{\Omega''}, C^{\Omega_1})}_{\substack{\text{new fine spiking} \\ \text{in } f^{\Omega_1}}}. \end{aligned} \quad (8)$$

Our hypothesis testing perspective suggests we should now test to see if $\mathbb{P}(X|C^{\Omega''}, C^{\Omega_1})$ is uniform in order to explore whether the information in $C^{\Omega''}$ can account for the non-uniformity originally detected in f^{Ω_1} . Notice, however, that $(C^{\Omega''}, C^{\Omega_1})$ is equivalent to $C^{\Omega'' \cup \Omega_1}$, so testing if $\mathbb{P}(X|C^{\Omega''}, C^{\Omega_1})$ is uniform is equivalent to testing if $\mathbb{P}(X|C^{\Omega'' \cup \Omega_1})$ is uniform, which is exactly a test of $H_0^{\Omega'' \cup \Omega_1}$. Furthermore, by defining $\Omega_2 = \Omega'' \cup \Omega_1$, we see that we are testing $H_0^{\Omega_2}$, where Ω_2 is a refinement of Ω_1 , thus recovering the original framework of testing successive refinements. (This procedure can obviously be iterated to include $\Omega_3, \Omega_4, \dots$, each time incorporating new binnings into the conditioning, say Ω''', Ω'''' ,)

To summarize, we can view the sequential procedure described here either as testing successively finer binnings or as testing successively narrower conditional distributions. In the first view Ω_2 is just some binning that we chose that happens to be finer than Ω_1 . In the second view Ω_2 is constructed via $\Omega'' \cup \Omega_1$, where Ω'' is a new type of binning that perhaps reveals structure that cannot be disambiguated with Ω_1 alone. This brings up two questions: (i) If Ω'' is itself a refinement of Ω_1 , is there a difference between using $\Omega_2 = \Omega''$ and $\Omega_2 = \Omega'' \cup \Omega_1$? (ii) Are there non-informative choices of Ω_2 or Ω'' ? Both questions are resolved by the next theorem.

Theorem 3: Suppose that Ω and Ω' are space-time binnings and that Ω' is a refinement of Ω . Then

$$f^{\Omega \cup \Omega'}(x|C^{\Omega \cup \Omega'}(x)) = f^{\Omega'}(x|C^{\Omega'}(x))$$

for all x . In particular, $H_0^{\Omega \cup \Omega'}$ is true if and only if $H_0^{\Omega'}$ is true.

The intuitive reason for this theorem is that $\mathbf{C}^\Omega$ can be uniquely determined from $\mathbf{C}^{\Omega'}$, so combining $\mathbf{C}^\Omega$ and $\mathbf{C}^{\Omega'}$ into $\mathbf{C}^{\Omega \cup \Omega'}$ provides no additional conditioning information beyond what was already in $\mathbf{C}^{\Omega'}$. To see that $\mathbf{C}^\Omega$ can be determined from $\mathbf{C}^{\Omega'}$, note that for each $\omega \in \Omega$ we can express

$$C_\omega(\mathbf{x}) = \sum_{(i,t) \in \omega} x_{it} = \sum_{j=1}^k \sum_{(i,t) \in \omega'_j} x_{it} = \sum_{j=1}^k C_{\omega'_j}(\mathbf{x}), \quad (9)$$

where $\{\omega'_1, \dots, \omega'_k\}$ is a partition of ω chosen from Ω' . A proof of Theorem 3 can be found in Lemma 11 in the Mathematical Appendix.

Returning to our two questions, in light of Theorems 1 and 3 we see that (i) if Ω'' is itself a refinement of Ω_1 , then defining $\Omega_2 = \Omega''$ is equivalent to defining $\Omega_2 = \Omega'' \cup \Omega_1$; and (ii) it is not useful to choose Ω_2 so that Ω_1 is also a refinement of Ω_2 (meaning they are each refinements of the other), nor is it useful to choose Ω'' to be coarser than Ω_1 (which again makes Ω_1 a refinement of Ω_2), because then testing $H_0^{\Omega_2}$ is equivalent to testing $H_0^{\Omega_1}$.

6.3 Adjusted p-values for testing successive refinements

When testing multiple hypotheses, one should appropriately adjust the p-values. Section 9 cautions about testing multiple hypotheses without proper adjustment of p-values in general contexts. In this section, we consider multiple testing in the special case of testing successive refinements, for which the logical structure of the null hypotheses can be used to provide adjusted p-values that are much less conservative than more generic approaches, such as Bonferroni corrections.

Consider two space-time binnings Ω and Ω' with p-values $\alpha(\mathbf{X})$ and $\alpha'(\mathbf{X})$ for the respective null hypotheses H_0^Ω and $H_0^{\Omega'}$, and further suppose that Ω' is a refinement of Ω . Suppose that we are testing at level ε , so that we only reject when a p-value is less than or equal to ε . The usual Bonferroni correction would be to change the level to $\varepsilon/2$, or equivalently to test at level ε using the adjusted p-values $2\alpha(\mathbf{X})$ and $2\alpha'(\mathbf{X})$. In the sequential testing framework from the previous section, however, we only test $H_0^{\Omega'}$ after a successful rejection of H_0^Ω . So the only way to reject $H_0^{\Omega'}$ is to have both $\alpha(\mathbf{X}) \leq \varepsilon$ and $\alpha'(\mathbf{X}) \leq \varepsilon$. An equivalent way to view this is to test at level ε using the original p-value for H_0^Ω and an adjusted p-value of $\max\{\alpha(\mathbf{X}), \alpha'(\mathbf{X})\}$ for $H_0^{\Omega'}$. It turns out that this provides the same guarantees as the Bonferroni correction, because of the logical structure of the null hypotheses described in Theorem 1.

Theorem 4: Suppose $\Omega_1, \dots, \Omega_r$ is a sequence of space-time binnings such that Ω_j is a refinement of Ω_{j-1} for each $j = 2, \dots, r$. Suppose also that $\alpha_j(\mathbf{X})$ is a valid p-value for testing $H_0^{\Omega_j}$ for each $j = 1, \dots, r$, meaning that $\mathbb{P}(\alpha_j(\mathbf{X}) \leq \varepsilon) \leq \varepsilon$ for all $\varepsilon \in [0, 1]$ whenever $H_0^{\Omega_j}$ is true. Define

$$\alpha_j^+(\mathbf{x}) \stackrel{\text{def}}{=} \max_{k=1, \dots, j} \alpha_k(\mathbf{x}).$$

Then $\alpha_j^+(\mathbf{X})$ is an adjusted p-value for testing $H_0^{\Omega_j}$ that controls the family-wise error rate for testing $H_0^{\Omega_1}, \dots, H_0^{\Omega_r}$, meaning

$$\mathbb{P}(\text{any false rejections}) = \mathbb{P}(\text{there exists a } j \text{ with } H_0^{\Omega_j} \text{ true and } \alpha_j^+(\mathbf{X}) \leq \varepsilon) \leq \varepsilon$$

for all $\varepsilon > 0$.

This is a special case of the closed testing procedure of Marcus et al. (1976). For completeness, Lemma 12 in the Mathematical Appendix contains a simple proof. The theorem only requires $\alpha_j(\mathbf{X})$ to be a valid p-value, so either of the p-values in Equations (6) or (7) will work. Rejecting those hypotheses for which the corresponding $\alpha_j^+(\mathbf{X}) \leq \varepsilon$ is equivalent to the sequential testing procedure in the previous section using level ε .

When the adjusted p-values given in Theorem 4 are used for testing successive refinements, then the sequence of hypotheses $H_0^{\Omega_1}, \dots, H_0^{\Omega_r}$ behaves like a single hypothesis test from the perspective of further multiple testing adjustments. For example, suppose $\Omega_1, \dots, \Omega_r$ is a sequence of successive refinements with adjusted p-values $\alpha_1^+, \dots, \alpha_r^+$, and similarly, $\Omega_1, \dots, \Omega_s$ is a sequence of successive refinements with adjusted p-values $\bar{\alpha}_1^+, \dots, \bar{\alpha}_s^+$. To simultaneously test both sequences, we can use the adjusted p-values $2\alpha_1^+, \dots, 2\alpha_r^+$ and $2\bar{\alpha}_1^+, \dots, 2\bar{\alpha}_s^+$, even though there are $r + s$ different hypothesis tests. These adjustments provide the same guarantees as the naive Bonferroni correction, namely, they control the familywise error rate, or the probability of making even a single false rejection. But they are much more powerful than the Bonferroni correction, which makes an overly conservative adjustment by dividing the level by $r + s$. We do not know if there are further improvements for the case when the sequences are interrelated, for example, if some Ω_j is a refinement of some Ω_k .

Examples 1c and 2c — Adjusted p-values—Example 1a uses the space-time binning TIME-100 and then a refinement TIME-25. The respective p-values are $\alpha^{\text{TIME-100}}(\mathbf{X}) = 0.0278$ and $\alpha^{\text{TIME-25}}(\mathbf{X}) = 0.661$. Using Theorem 4 to adjust for multiple hypothesis tests gives $\alpha^{\text{TIME-100},+}(\mathbf{X}) = 0.0278$ and $\alpha^{\text{TIME-25},+}(\mathbf{X}) = 0.661$, which are identical to the original p-values, because the originals were already increasing in value. The increased power for this type of multiple testing adjustment is clear. Using α^+ and a level of 0.05 we can continue to reject $H_0^{\text{TIME-100}}$ even after adjusting for multiple tests. On the other hand, a standard Bonferroni correction, which provides the same type I error guarantees as Theorem 4, would simply double the p-values in this case, so that $H_0^{\text{TIME-100}}$ is no longer rejected at level 0.05. Example 1b is identical. Theorem 4 can be used with any valid p-values, including valid Monte Carlo p-values.

Examples 2a–b use the space-time binnings TIME-100, TIME-25, and TA-(25, 5), which is a sequence of successive refinements. The respective p-values are $\alpha^{\text{TIME-100}}(\mathbf{X}) = 0.0001$, $\alpha^{\text{TIME-25}}(\mathbf{X}) = 0.0001$, and $\hat{\alpha}_Q^{\text{TA-(25,5)}}(\mathbf{X}) = 0.132$. These are each valid p-values and, as with Example 1, adjusting them with Theorem 4 does not change their values, because they are already increasing (technically, non-decreasing) along the sequence of successive refinements. If we use a level of 0.05 and reject $H_0^{\text{TIME-100}}$ and $H_0^{\text{TIME-25}}$, then our chance of making even a single type I error is at most 5%.

7 Special cases

In this section we provide some intuitive discussion of the space-time binnings TIME- , ALL- Γ , and TA-(, Γ), which are used in Figure 1 and several of the examples. These special cases are also discussed in Section 8.2 in the context of maximum entropy models, or log-linear models. TIME- corresponds to interval jitter and is useful for investigating precise spike timing in the context of arbitrary, but slower modulation in spiking rates. Rapid fluctuations in population rate (Okun et al., 2012) are a potential source of precise spike timing. Population rate is measured by $C^{\text{ALL-}\Gamma}$. Combining TIME- and ALL- Γ gives TA-(, Γ), which can be used to investigate additional types of precise spiking in the context of both slower modulations in spiking rates and potentially rapid modulations in population rate.

7.1 TIME-

Using $\Omega = \text{TIME-}$ is called interval jitter. Interval jitter is extensively reviewed in Amarasingham et al. (2012). Since it is based on a pure temporal coarsening, it can be used to probe the temporal precision of spiking, and in particular, for testing for the existence of fine temporal structure in the pattern of spiking. The parameter controls the *qualitative* distinction between fine and coarse time scales in the interpretation of interval jitter. Since “time scale” is not a precisely defined term, the exact numerical value of should not be over-interpreted. As mentioned in Section 6.1, TIME- is a refinement of TIME- j for any positive integer j , so if $H_0^{\text{TIME-}j\Delta}$ is true, so is $H_0^{\text{TIME-}\Delta}$. Intuitively speaking, if there is no structure to the pattern of spiking within large time bins, then there can be no structure within smaller time bins. Testing a sequence of successive refinements such as TIME- 2^r , TIME- 2^{r-1} , ..., TIME-2 can be used to find the finest precision of spiking supported by the data. Figure 1D shows resamples under $H_0^{\text{TIME-}\Delta}$. Examples 1a–c and 2a above use TIME-

7.2 ALL- Γ

Using $\Omega = \text{ALL-}\Gamma$ conditions inference on a spatial and temporal coarsening of the data. Like TIME- Γ , spike times are blurred over Γ -length windows, but unlike TIME- Γ , the identities of spikes (meaning which spike train they originally belonged to) are not preserved. Since TIME- Γ is a refinement of ALL- Γ , a rejection of $H_0^{\text{TIME-}\Gamma}$ implies a rejection of $H_0^{\text{ALL-}\Gamma}$. Similarly, a failure to reject $H_0^{\text{ALL-}\Gamma}$ is a recommendation against testing $H_0^{\text{TIME-}\Gamma}$. Figure 1B–C shows resamples under $H_0^{\text{ALL-}\Gamma}$.

The interpretation of $H_0^{\text{ALL}-\Gamma}$ depends on whether the different spike trains index different neurons or different trials. We first discuss the case of different neurons so that $C^{\text{ALL}-\Gamma}$ measures the time-varying population rate measured in Γ -length bins. The null hypothesis $H_0^{\text{ALL}-\Gamma}$ states that modulations in population rate can explain all of the structure in the data. In many experiments, recorded neurons differ markedly in their overall activity levels and in their responses to experimental or behavioral manipulations, suggesting that $H_0^{\text{ALL}-\Gamma}$ is false for any choice of Γ . A rejection of $H_0^{\text{ALL}-\Gamma}$ is of little experimental interest in this case. For certain test statistics, however, a failure to reject $H_0^{\text{ALL}-\Gamma}$ can be revealing, suggesting that the test statistic and/or the data have little power for distinguishing the differential responses of different neurons.

If the different spike trains correspond to different trials, then $C^{\text{ALL}-\Gamma}$ is the PSTH measured in bins of length Γ . The null hypothesis $H_0^{\text{ALL}-\Gamma}$ implies that this PSTH captures all of the statistical structure in the data. There are no additional dependencies within or across trials. Many experiments show long-range dependencies that are not consistent with $H_0^{\text{ALL}-\Gamma}$ for any Γ . A common example is when the total spike counts per trial are either more or less variable than they should be under $H_0^{\text{ALL}-\Gamma}$. See Example 3 below. Excess variability in this case is often called *trial-to-trial variability*. In situations where trial-to-trial variability is quite common or expected, then a rejection of $H_0^{\text{ALL}-\Gamma}$ may be of little interest. As with the case of multiple neurons, sometimes a failure to reject $H_0^{\text{ALL}-\Gamma}$ is even more revealing.

7.3 TA-(Δ, Γ)

Using $\Omega = \text{TA}-(\Delta, \Gamma)$ conditions both on the differential time-varying spiking rates of individual neurons measured in Δ -length bins and the shared time-varying population rate (for multiple neurons) or PSTH (for multiple trials) measured in Γ -length bins. One way to think about $H_0^{\text{TA}-(\Delta, \Gamma)}$ is to think about applying $H_0^{\text{ALL}-\Gamma}$ to a conditional distribution that already accounts for much of the differential activity across neurons and/or trials, namely, $f^{\text{TIME}-\Delta}$. For example, suppose we reject $H_0^{\text{TIME}-\Delta}$, suggesting that the temporal precision of spiking is less than Δ . We may want to further probe the nature of this precise spiking. One approach, of course, is to test $H_0^{\text{TIME}-\Gamma}$ for some $\Gamma < \Delta$, which simply probes the data at an even finer precision. An intermediate step, however, is to first test whether the fine precision spiking is consistent with $H_0^{\text{ALL}-\Gamma}$. As described more fully in Section 6.2, the way to carry-out this successive testing is to simply test the original data using $H_0^{\text{TA}-(\Delta, \Gamma)}$, which asks whether the fine precision spiking can be explained by the additional information in $C^{\text{ALL}-\Gamma}$, beyond what was already explained by $C^{\text{TIME}-\Gamma}$. Figure 1E shows resamples under $H_0^{\text{TA}-(\Delta, \Gamma)}$.

The interpretation of $H_0^{\text{TA}-(\Delta, \Gamma)}$ for $\Gamma < \Delta$ derives from this successive testing perspective of the previous paragraph. There are potentially many types of precise spike timing, some of which primarily influence $C^{\text{ALL}-\Gamma}$. Recall that $C^{\text{ALL}-\Gamma}$ is the the overall activity of the recorded ensemble, or the population rate, when different spike trains correspond to

different neurons, and it is the PSTH when different spike trains correspond to different trials. If a test of $H_0^{\text{TIME}-\Delta}$ — interval jitter — detects precise spike timing, then it may be detecting fine temporal structure in $C^{\text{ALL}-\Gamma}$. Testing $H_0^{\text{TA}-(\Delta,\Gamma)}$ can help the investigator decide if processes influencing only $C^{\text{ALL}-\Gamma}$ are sufficient to explain the precise timing detected by interval jitter. A failure to reject $H_0^{\text{TA}-(\Delta,\Gamma)}$ means that the temporal coarsening, $C^{\text{TIME}-}$, and the ensemble activity, $C^{\text{ALL}-}$ (or the PSTH for repeated trials), can account for the patterns of spiking in $\mathbf{X}$. A rejection of this null hypothesis suggests the existence of fine temporal processes beyond those that influence only coarse temporal dynamics and total ensemble activity (or PSTH). Examples 2b–c above use $\text{TA}-(\Delta, \Gamma)$.

$H_0^{\text{TA}-(\Delta,\Gamma)}$ is easiest to use when $\Delta = k\Gamma$ for some $k > 1$, so that the boundaries of the space-time bins used for the temporal coarsening, namely those bins in $\text{TIME}-k\Gamma$, align with the boundaries of the bins used for the finer spatio-temporal coarsening in $\text{ALL}-\Gamma$. This has important computation advantages, as described in the Monte Carlo Appendix. Testing $H_0^{\text{TA}-(\Delta,\Gamma)}$ for Γ is of little interest, since, intuitively speaking, there is no additional information in $C^{\text{ALL}-\Gamma}$ beyond that in $C^{\text{TIME}-}$ when Γ is small. $C^{\text{ALL}-\Gamma}$ is more blurred, so to speak, than $C^{\text{TIME}-}$ when Γ is small. This is formally true when $\Gamma = k^{-1}$ for some integer $k \geq 1$, in which case $H_0^{\text{TA}-(\Delta,k\Delta)}$ is equivalent to $H_0^{\text{TIME}-\Delta}$; see Theorem 3.

8 Maximum entropy models

This section connects the null hypothesis H_0^Ω to more classical, parametric statistical models of spike trains, namely, log-linear models (Agresti, 2007; Amari and Nagaoka, 2000), sometimes called maximum entropy models. Log-linear models are popular contemporary models for spike train data (Gerstein et al., 1989; Martignon et al., 1995; Vaadia et al., 1995; Martignon et al., 2000; Amari et al., 2003; Gütig et al., 2003; Schneidman et al., 2006; Shlens et al., 2006; Montani et al., 2009; Roudi et al., 2009; Truccolo et al., 2010; Kass et al., 2011; Long II and Carmena, 2011; Shimazaki et al., 2012). We begin with the abstract framework and then use special cases to build intuition.

8.1 Coarsened data as sufficient statistics in log-linear models

Suppose that the distribution of the data can be factored as

$$p(\mathbf{x}|\boldsymbol{\beta}) \stackrel{\text{def}}{=} \prod_{t=1}^T \prod_{i=1}^N \beta_{it}^{x_{it}} (1-\beta_{it})^{1-x_{it}}, \quad (10)$$

for all $\mathbf{x}$ and for some $\boldsymbol{\beta} = (\beta_{it} : i = 1, \dots, N; t = 1, \dots, T) \in (0, 1)^{N \times T}$. This model says that, given $\boldsymbol{\beta}$, all time bins across all spike trains are mutually independent Bernoulli random variables with the probability of a spike in time bin t of spike train i given by β_{it} . This inhomogeneous Bernoulli process model is the discrete-time equivalent of an inhomogeneous Poisson process model. The Bernoulli process formulation in Equation (10) can be equivalently written as a **log-linear model**, or a maximum entropy model, by transforming $\boldsymbol{\beta}$ into $\boldsymbol{\theta}$ via

$$\theta_{it} \stackrel{\text{def}}{=} \text{logit } \beta_{it} \stackrel{\text{def}}{=} \log \frac{\beta_{it}}{1-\beta_{it}}, \quad (11)$$

so that

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp \left(\sum_{i=1}^N \sum_{t=1}^T \theta_{it} x_{it} \right), \quad (12)$$

where $\kappa(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{\mathbf{x}} \exp(\sum_i \sum_t \theta_{it} x_{it}) = \prod_t \prod_i (1 + \exp(\theta_{it}))$ is the normalization constant. The models in Equations (10) and (12) are equivalent, only parameterized differently.

In this paper it does not matter whether we conceptualize $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$ as a collection of deterministic parameters or as a random process. When the parameters are random, then the resulting model on $\mathbf{X}$ (obtained after averaging over the random parameters) is a mixture of inhomogeneous Bernoulli processes, which is the discrete-time equivalent of a Cox process or a doubly-stochastic Poisson process. The terms log-linear model and maximum entropy model are most frequently used when the parameters are deterministic.

For a space-time binning, Ω , define the null hypothesis

$$\tilde{H}_0^\Omega: \theta_{it} = \sum_{\omega \in \Omega: (i,t) \in \omega} \tilde{\theta}_\omega \quad \text{for all } i, t, \text{ and some } \tilde{\boldsymbol{\theta}} = (\tilde{\theta}_\omega \in \mathbb{R}: \omega \in \Omega) \quad (13)$$

Instead of explicitly specifying each component of $\boldsymbol{\theta}$, we specify a smaller set of parameters in $\tilde{\boldsymbol{\theta}}$ (one parameter per space-time bin) and use these to determine $\boldsymbol{\theta}$ through the constraints specified in the null hypothesis in Equation (13). One way to interpret this null hypothesis is that each space-time bin, ω , contributes a constant, additive effect to each of its log-linear model parameters, i.e., to those θ_{it} with $(i, t) \in \omega$. Under $\tilde{H}_0^\Omega$, the log-linear model in Equation (12) becomes

$$p^\Omega(\mathbf{x}|\tilde{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} \frac{1}{\tilde{\kappa}^\Omega(\tilde{\boldsymbol{\theta}})} \exp \left(\sum_{\omega \in \Omega} \tilde{\theta}_\omega \sum_{(i,t) \in \omega} x_{it} \right) = \frac{1}{\tilde{\kappa}^\Omega(\tilde{\boldsymbol{\theta}})} \exp \left(\sum_{\omega \in \Omega} \tilde{\theta}_\omega C_\omega(\mathbf{x}) \right), \quad (14)$$

where $\tilde{\kappa}^\Omega$ is a normalization constant. Using statistical terminology, the null hypothesis is that the coarsened data are the **sufficient statistics** of the log-linear model.

Our STCI framework can be used to test $\tilde{H}_0^\Omega$.

Theorem 5: Assume the modeling approach in Equations (10)–(12), and assume $\tilde{H}_0^\Omega$ is true. Then H_0^Ω is true.

In particular, a rejection of H_0^Ω using STCI implies a rejection of $\tilde{H}_0^\Omega$. A formal proof of Theorem 5 can be found in Lemma 9 in the Mathematical Appendix.

It is important to note that H_0^Ω and $\tilde{H}_0^\Omega$ are not equivalent. A rejection of H_0^Ω is both a rejection of $\tilde{H}_0^\Omega$ and *a rejection of a much larger class of models*. Log-linear models satisfying $\tilde{H}_0^\Omega$ are not the only class of models that satisfy H_0^Ω . STCI is much more general. In particular, STCI makes no assumptions at all about g^Ω , the distribution of the coarsened data $\mathbf{C}^\Omega$, whereas, $\tilde{H}_0^\Omega$ contains many implicit assumptions about g^Ω . Figure 6 provides a schematic illustration.

Even though H_0^Ω and $\tilde{H}_0^\Omega$ are not equivalent, everything in Section 6 about refinements continues to hold if H_0 is replaced by $\tilde{H}_0$, the key result being

Theorem 6: Suppose that Ω and Ω' are space-time binnings, that Ω' is a refinement of Ω , and that $\tilde{H}_0^\Omega$ is true. Then $\tilde{H}_0^{\Omega'}$ is true.

This theorem is easy to prove directly from Equation (13) and the definition of a refinement. Complete details are in the Mathematical Appendix.

$\tilde{H}_0^\Omega$ is easier to state and interpret in the special case when Ω is a space-time partition, in particular,

$$\tilde{H}_0^\Omega: \theta_{it} = \tilde{\theta}_\omega \quad \text{for all } i, t, \omega \text{ with } (i, t) \in \omega \text{ and some } \tilde{\theta} \text{ (when } \Omega \text{ is a partition).}$$

The null hypothesis can be equivalently formulated in terms of spiking probabilities as

$$\tilde{H}_0^\Omega: \beta_{it} = \tilde{\beta}_\omega \quad \text{for all } i, t, \omega \text{ with } (i, t) \in \omega \text{ and some } \tilde{\beta} \text{ (when } \Omega \text{ is a partition).}$$

For a partition, the null hypothesis is that the probability of spiking given θ (or β) is constant within each space-time bin of the partition Ω . If we conceptualize spike trains as independent inhomogeneous Bernoulli processes (or mixtures of such processes), then the null hypothesis that the spiking probabilities do not vary within the space-time bins of Ω can be tested by our conditional inference framework. Although it seems unlikely that these null hypotheses are exactly true, as long as these parameters vary little within the space-time bins, then this piecewise constant assumption will be an excellent approximation.

The point of this section was to show that STCI can be used to test hypotheses about whether the parameters of a first-order log-linear model have certain properties relating to space-time binnings. A rejection of these hypotheses can be interpreted as a statement about the parameters or as a statement about the underlying modeling assumptions in Equation (10) or (12). This latter interpretation is common. For example, the first-order log-linear

model in Equation (12) could be tested against a higher-order model that allows for interactions between spikes. (The basic model in Equation (10) or (12) is universal if the parameters are unconstrained and cannot be tested directly without some type of additional constraints, such as those provided by $\tilde{H}_0^\Omega$. For example, choosing each β_{it} close to the corresponding x_{it} , i.e., a high probability of spiking where we observe a spike and a low probability where we do not, gives a choice of β that always provides a good fit to the data.) Conditional inference has a long history within statistics for constructing exact tests of log-linear models (Agresti, 2001; Reid, 1995). See Gütig et al. (2003) for a neurostatistics example that illustrates how to use conditional inference, at least in principle, to test for higher-order interactions in a temporally stationary g th-order log-linear model.

8.2 Parameterizations for special cases

In this section we specialize $\tilde{H}_0^\Omega$ to the space-time binnings from Section 7. Recall that

$$\begin{aligned}\text{TIME}-\Delta &\stackrel{\text{def}}{=} \{\{i\} \times W_b^\Delta : i=1, \dots, N; b=1, \dots, T^\Delta\}, \\ \text{ALL}-\Gamma &\stackrel{\text{def}}{=} \{\{1, \dots, N\} \times W_d^\Gamma : d=1, \dots, T^\Gamma\}, \\ \text{TA}-(\Delta, \Gamma) &\stackrel{\text{def}}{=} \text{TIME}-\Delta \cup \text{ALL}-\Gamma.\end{aligned}$$

For these space-time binnings, the corresponding log-linear model null hypotheses are

$$\tilde{H}_0^{\text{TIME}-\Delta} : \theta_{it} = \nu_{ib} \quad \text{for the } b \text{ with } t \in W_b^\Delta, \quad (15)$$

$$\tilde{H}_0^{\text{ALL}-\Gamma} : \theta_{it} = \xi_d \quad \text{for the } d \text{ with } t \in W_d^\Gamma, \quad (16)$$

$$\tilde{H}_0^{\text{TA}-(\Delta, \Gamma)} : \theta_{it} = \nu_{ib} + \xi_d \quad \text{for the } b \text{ and } d \text{ with } t \in W_b^\Delta \cap W_d^\Gamma, \quad (17)$$

for all i and t and for some $\nu \in \mathbb{R}^{N \times T}$ and $\xi \in \mathbb{R}^{1 \times T^\Gamma}$. Each of these null hypotheses presupposes the basic log-linear model framework of Equations (10)–(12). For example, $\tilde{H}_0^{\text{TA}-(\Delta, \Gamma)}$ corresponds to the log-linear model

$$p^{\text{TA}-(\Delta, \Gamma)}(x | \nu, \xi) = \frac{1}{\tilde{\kappa}^{\text{TA}-(\Delta, \Gamma)}(\nu, \xi)} \exp \left(\sum_{b=1}^{T^\Delta} \sum_{i=1}^N \nu_{ib} \sum_{t \in W_b^\Delta} x_{it} + \sum_{d=1}^{T^\Gamma} \xi_d \sum_{i=1}^N \sum_{t \in W_d^\Gamma} x_{it} \right).$$

Within this framework, the null hypotheses control the precision of spiking by putting constraints on the allowed variation of the model parameters, θ , or equivalently, by putting constraints on the allowed variation of the probabilities of spiking, β .

For $\tilde{H}_0^{\text{TIME}-\Delta}$ in Equation (15), θ can vary arbitrarily across spike trains and across coarse Δ -length bins (the jitter intervals), but cannot vary within these coarse bins. The values within the coarse bins are specified by ν . Translating the null hypothesis into a statement about the spiking probabilities gives the equivalent formulation

$$\tilde{H}_0^{\text{TIME}-\Delta} : \beta_{it} = \mu_{ib} \quad \text{for the } b \text{ with } t \in W_b^\Delta, \quad (18)$$

or all i and t and for some $\mu \in (0, 1)^{N \times T}$. Loosely speaking, each spike train's firing rate is piecewise constant with Δ -length pieces, which should be an excellent approximation to an even larger class of models with firing rates that vary slowly on Δ -length time scales. The ν in Equation (15) is related to the μ in Equation (18) via $\nu_{ib} = \text{logit}^{-1}(\mu_{ib})$.

For $\tilde{H}_0^{\text{ALL}-\Gamma}$ in Equation (15), θ can vary arbitrarily across coarse Γ -length bins (the jitter intervals), but cannot vary within these coarse bins and cannot vary across spike trains. The common value of θ within the coarse bins is specified by ξ . Translating the null hypothesis into a statement about the spiking probabilities gives the equivalent formulation

$$\tilde{H}_0^{\text{ALL}-\Gamma} : \beta_{it} = \tau_d \quad \text{for the } d \text{ with } t \in W_d^\Gamma, \quad (19)$$

for all i and t and for some $\tau \in (0, 1)^{1 \times T^\Gamma}$. The ξ in Equation (16) is related to the τ in Equation (19) via $\xi_d = \text{logit}^{-1}(\tau_d)$. The difference between $\tilde{H}_0^{\text{ALL}-\Gamma}$ and $\tilde{H}_0^{\text{TIME}-\Gamma}$ is that, for a given slice of time, the former requires all spike trains to share the same probability of spiking, whereas the latter allows each spike train to have a different probability of spiking.

For $\tilde{H}_0^{\text{TA}-(\Delta, \Gamma)}$ in Equation (17), each spike train has two components that control the probability of spiking. The first is a shared component (ξ) that varies on Γ -length time scales, analogous to $\tilde{H}_0^{\text{ALL}-\Gamma}$. Usually we imagine that Γ is small so that this shared component allows for rapid changes in firing rate, as long as the changes are shared across all spike trains. The second component (ν) varies on Δ -length time scales, but is not shared across spike trains, analogous to $\tilde{H}_0^{\text{TIME}-\Delta}$. The only interesting case is when $\Delta > \Gamma$, so that the components of individual variation cannot change as rapidly as the shared-variation component, otherwise we can subsume the shared variation into the individual variation. Transforming the constraints on θ into equivalent constraints on β gives

$$\tilde{H}_0^{\text{TA}-(\Delta, \Gamma)} : \beta_{it} = \frac{\lambda_{ib} \gamma_d}{1 + \lambda_{ib} \gamma_d} \quad \text{for the } b \text{ and } d \text{ with } t \in W_b^\Delta \cap W_d^\Gamma, \quad (20)$$

for all i and t and for some $\lambda \in (0, \infty)^{N \times T}$ and some $\gamma \in (0, \infty)^{1 \times T^\Gamma}$, where the parameters in Equations (17) and (20) are related by $\lambda_{ib} = \exp(\nu_{ib})$ and $\gamma_d = \exp(\xi_d)$. At first glance, the constraints on the spiking probabilities in Equation (20) are much less intuitive than the equivalent constraints on the log-linear model parameters in Equation (17). For some

intuition, consider the case where the entries of β are small (close to zero), as one might expect for small time bins. In this case, the constraints in Equation (20) become

$$\beta_{it} \approx \lambda_{ib} \gamma_d,$$

so we can interpret the original constraints as roughly multiplicative: the spiking probabilities consist of slowly varying components, which can be different for each spike train and can vary arbitrarily across Γ -coarse time bins, multiplied by a shared, perhaps rapidly varying component, which can vary arbitrarily across finer Γ -coarse time bins, but is shared by all spike trains. Shared multiplicative variation has been considered by several authors as a statistical model for correlation between spike trains (Ventura et al., 2005a; Kelly and Kass, 2012) and is also typical in generalized linear models for conditional intensity functions where correlations are explained by neuronal interactions and common sensory stimuli and behavioral covariate inputs (Truccolo et al., 2005; Pillow et al., 2008; Truccolo et al., 2010). (As a formal constraint, multiplicative variation is cleaner when applied to the intensities of continuous time point processes, since the intensities are not constrained to take values between zero and one like the spiking probabilities of a discrete-time zero-one process. That the spiking probabilities can never exceed one partly explains the unusual appearance of Equation (20).)

Example 3: Trial-to-trial variability—The trials on the top right panel of Figure 1 from monkey visual cortex are ordered from lowest to highest spike count, and the gradient of increasing firing rate is apparent. Under what types of models would such variation in spike counts be typical? For example, can the variation be explained by an inhomogeneous Bernoulli process model without trial-to-trial variability, meaning that each X_{it} is an independent Bernoulli random variable with probability of spiking given by an unknown parameter τ_t that can vary arbitrarily in time (t), but is shared across all trials (i)? This class of models is exactly $\tilde{H}_0^{\text{ALL}-1}$; see Equation (19).

We tested $H_0^{\text{ALL}-1}$ using the data from the top right panel of Figure 1 and using the test statistic

$$S(\mathbf{x}) \stackrel{\text{def}}{=} \sum_i \left(\sum_t x_{it} \right)^2,$$

which is the sum of the squared spike counts from each trial (equivalent to using the variance of spike counts across trials, in this case). The Monte Carlo p-value using $n = 10000$ was $\alpha(\hat{\mathbf{X}}) = 0.0001$. Since a rejection of $H_0^{\text{ALL}-1}$ implies a rejection of $\tilde{H}_0^{\text{ALL}-1}$ (Theorem 5), we can conclude that, from the perspective of inhomogeneous Bernoulli process models, the monkey data exhibits trial-to-trial variability. (The notion of “trial-to-trial variability” can and has been used within the context of many different models, not just the inhomogeneous Bernoulli process model in this example.)

9 Statistical and experimental interpretations

Modern neurophysiologists require a suite of statistical tools to interrogate increasingly larger and more complex datasets. STCI is one such tool that can be used to probe the structure of neural spiking patterns, perhaps revealing clues about collective neural dynamics. As with all statistical models and methods, however, proper use requires a careful understanding of the method's assumptions and limitations. In this section we discuss some of the issues that can arise when using our hypothesis testing framework. As mentioned later in the discussion, STCI is not limited to hypothesis testing, and some of the concerns discussed here may be alleviated in future work that extends the framework to more complete inferential settings.

A rejection of the null hypothesis is a rejection that f in Equation (3) is uniform. In particular, a small p-value, $\alpha(\mathbf{X})$, is statistical evidence that the pattern of precise spiking in the observed data, $\mathbf{X}$, is more likely than the pattern of precise spiking in other hypothetical datasets having the same coarsening, $\mathbf{C}(\mathbf{X})$. As with all statistical procedures, relating this statistical evidence to a given scientific problem requires great care, and investigators might reasonably disagree about the appropriateness of certain experimental conclusions. The most benign conclusion from a rejection of the null hypothesis is that there are processes that structure the precision of spiking beyond what is implied by $\mathbf{C}$. The nature of these processes and details about their biological or scientific importance are left unspecified. In principle, the choice of test statistic, S , provides some additional evidence about the nature of these processes. For example, if S measures precise synchrony, then a rejection of the null hypothesis further suggests that these processes promote patterns of spiking with more precise synchrony over those with less precise synchrony. Again, the nature of these processes and whether synchrony plays some biological role are not addressed by the hypothesis test. Any number of processes that contribute to fine precision spiking, including processes seemingly unrelated to synchrony, might promote patterns with more precise synchrony by happenstance.

One of the most common classes of processes that might contribute to fine precision spiking are those that influence the fine-temporal autocorrelation of spike trains, such as processes that create refractory periods and bursting. Some aspects of the short absolute refractory period are automatically enforced by our framework since it prevents multiple spikes in a fine space-time bin. Other types of self-excitation or self-inhibition are not included in any of our null hypotheses, however, and it is always possible that a rejection of the null hypothesis results from fine-temporal autocorrelation structure. This is not an incorrect rejection, since spiking patterns are indeed finely structured, but in many contexts it is not what an investigator intends. That our hypothesis testing framework cannot explicitly rule out fine-temporal autocorrelation structure as the source of a rejection is one of its main limitations. This limitation can be alleviated somewhat by using test statistics that are robust to violations of uniformity in the fine-temporal autocorrelation structure, and by using complementary methods, such as pattern jitter (Harrison and Geman, 2009), to verify that the observed autocorrelation structure is not a source of concern. Nevertheless, our hypothesis testing framework is likely inappropriate in settings with strong fine-temporal

autocorrelation structure for which the investigator wishes to explore other types of fine-precision spiking patterns. We revisit this point below when we discuss future work.

A failure to reject the null hypothesis means that there is a lack of statistical evidence that f is not uniform. Often this is interpreted experimentally as a lack of evidence for processes that create fine precision spiking. This lack of evidence could mean that there are no such processes or it could mean that the processes exist, but the hypothesis test is not powerful enough to detect them. Lack of power can result from many overlapping reasons, including (i) the processes create a weak signal, (ii) the test statistic, S , is inappropriate for detecting the signal created by these processes, (iii) there is insufficient data to detect these processes, or (iv) conditioning on C does not leave enough residual variability in spiking in order to have a meaningful hypothesis test, which we call *over-conditioning*. An extreme case of over-conditioning is when the value of $C(X)$ completely determines X , so that there are no other hypothetical datasets with the same coarsening as the original data, that is, $\eta(C(X)) = 1$. In this case $\alpha(X) = 1$ and we cannot reject the null hypothesis.

It is frequently the case that one performs many hypothesis tests, because, for example, there are multiple experimental conditions that are analyzed separately, or because one wishes to explore multiple space-time binnings, Ω , or multiple test statistics, S . There are two dangers that must be avoided here. The first is the well-known danger of multiple hypothesis testing: many hypothesis tests will lead to some small p-values even if all of the hypothesis tests are true. Multiple hypothesis tests should always be reported and the p-values corrected to ensure appropriate interpretations (Lehmann and Romano, 2005). For example, one can use the Bonferroni correction to ensure that the familywise error rate is controlled at the stated level. The second danger results from misinterpreting the pattern of rejected and accepted hypotheses as evidence for underlying changes in the processes that control precise spike timing. As we just noted, there are many potential causes for failing to reject a null hypothesis. A particular pattern of rejected hypotheses could result from any combination of these causes. For example, rejecting the null hypothesis in experimental condition A, but not in experimental condition B, is not always compelling evidence that there is more precise spiking in condition A than in condition B. It may be that the coarsened data, that is, the observed C , in condition B does not provide enough power for detecting precise spiking, even though the same processes governing precise spiking are at work in both conditions. Nevertheless, it seems to be common practice to scientifically interpret the pattern of accepted and rejected null hypotheses.

Our favorite usages of this type of hypothesis testing are (i) as an aid to model selection and (ii) as a check against model misspecification. In the first usage, one varies the space-time binning, Ω , in order to gain an understanding of the degree of spiking precision in a given system. This then informs the design of the precise spiking components of more complete statistical models for the data. For example, one might vary the parameter (the jitter window width) in tests of $H_0^{\text{TIME}-\Delta}$ (interval jitter) to understand the degree of temporal precision in spike timing that needs to be captured by a parametric statistical model.

In the second usage, one begins with a statistical procedure that results in some conclusion about the degree of spiking precision, and then uses the numerical output of this procedure

as the test statistic, S , for testing a null hypothesis, H_0^Ω , that cannot have the same degree of precision. A failure to reject is evidence of model misspecification. For example, consider using a pairwise log-linear model to estimate a parameter, say σ , that controls millisecond precision synchrony and that this estimate is large, meaning that there seems to be a lot of synchrony. As a sanity check against model misspecification and/or systematic estimation errors, the spikes are perturbed uniformly (jittered) using a pure temporal binning, say TIME- , and the entire estimation procedure is repeated. If the estimated value of σ now becomes much smaller, we are reassured that σ is capturing the effects of precise spike timing as intended. If, however, the estimated value of σ remains high, then this is strong evidence that σ is not measuring precise spike timing, since there can be no precise timing in a temporally jittered dataset, but rather, the high value of σ can be explained entirely by the coarse temporal spiking. In this latter case there is something wrong with our log-linear modeling approach. Usually this procedure would be repeated many times to compare the original estimate of σ to a collection of estimates obtained from artificial data. Formally, the estimator of σ becomes the test statistic, S , in our conditional inference hypothesis test of $H_0^{\text{TIME}-\Delta}$.

Finally, we repeat the well-known caution that statistical significance and scientific significance can be misaligned. Scientifically irrelevant deviations from a statistical null hypothesis may lead to tiny p-values, particularly in large datasets. Hence, hypothesis testing is best used in conjunction with carefully chosen test statistics that are tailored to sensible alternative hypotheses, along with other procedures designed to quantify the experimental relevance of any detected departure from the null hypothesis. In the absence of a complete statistical model for alternative hypotheses, Amarasingham et al. (2012) and Harrison et al. (2013) suggest choosing test statistics that have physiologically interpretable units (such as the number of synchronous spikes per second), and then quantifying the difference between the observed test statistic and its expected value under the null hypothesis. A negligible difference, regardless of whether the difference is statistically significant or not, may suggest that the detected amounts of precise spiking are of little physiological importance.

10 Discussion

This paper introduces STCI and specializes it for hypothesis tests concerning the spatio-temporal precision of neuronal spiking patterns. This work arose out of our attempts over many years to extend jitter-style resampling methods to more complex settings such as those exemplified in Figures 1 and 4. The intuition underlying the null hypothesis $H_0^{\text{TA}-(\Delta,1)}$, which combines interval jitter with a constraint that preserves the instantaneous population rate (for multiple neurons) or the unsmoothed PSTH (for multiple trials), has been of particular interest. Both ourselves and others (Furukawa and Middlebrooks, 2002; Harrison et al., 2007; Smith and Kohn, 2008) have proposed resampling algorithms targeting this intuition, but these algorithms are lacking in that they are not statistically well-formulated, they allow multiple spikes from the same neuron at the same time, and the accompanying resampling procedures are not appropriately uniform. A notable exception is the rasters marginals model in Okun et al. (2012), which can be interpreted as properly testing $H_0^{\text{TA}-(\Delta,1)}$ for the special case of $\Delta = T$ and restricted to test statistics that are invariant to

permutations of time bins. STCI improves upon and unifies all of these efforts within a common and much more general statistical framework.

STCI is itself a special case of the general principle of conditional modeling and inference formalized in Equation (3), which is valid for arbitrary conditioning statistics C , not necessarily the space-time coarsenings described in Section 2. There are likely other families of statistics that would lead to other useful frameworks. Pattern jitter is a step in this direction (Harrison and Geman, 2009; Amarasingham et al., 2012). In this regard, STCI can be viewed as a template for developing other frameworks that, like the framework here, share common computational and statistical considerations among a family of restricted conditioning statistics. Although one could formulate an abstract theory that includes arbitrary conditioning statistics, there is little to say beyond Equations (3) and (5) and the Lemmas in the Mathematical Appendix, which are stated in abstract terms.

Finally, we would like to emphasize that conditional inference is not restricted to hypothesis testing (Reid, 1995). The primary appeal of conditional inference is that the methods are robust to the distribution g of the conditioning statistic C . This robustness, which is increasingly important as neuroscience generates ever larger and more complex datasets, is true for hypothesis testing, but also for other procedures like parameter estimation. By modeling the conditional distribution f with some statistical family, one can in principle estimate properties about precise spiking patterns, as opposed to just testing whether the patterns appear uniform. By moving away from the uniform distribution, one can also adjust for spiking patterns that are of little experimental interest, but that cannot be conditioned away, such as precise autocorrelation structure for STCI. In future work we plan to extend STCI beyond hypothesis testing, combining it with modern spike train models in order to robustly quantify the statistical structure of detected neuronal spiking patterns.

Acknowledgments

This work was supported in part by the National Science Foundation grants DMS-1007593 and DMS-1309004, by a National Institute of Mental Health (NIMH) grant R01MH102840, by a National Institute of Neurological Disorders and Stroke (NINDS) grant R01NS079533 and a NINDS K01 Career Award NS057389, by a National Institute of General Medical Sciences, National Institutes of Health grant IDeA P20GM103645, by the Defense Advanced Research Projects Agency contract FA8650-11-1-715, and by a U.S. Department of Veterans Affairs Merit Review Award. The authors thank Omar J. Ahmed, N. Stevenson Potter and Sydney S. Cash for help with the human epilepsy data collection, M. Smith and A. Kohn for sharing the monkey data in Figure 1, and C. Curto and two anonymous referees for helpful comments. Conversations and collaborations with S. Geman have strongly influenced our ideas about jitter and conditional inference.

References

- Abeles, M. Local Cortical Circuits: Studies of Brain Function. Springer; Berlin: 1982.
- Agresti A. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*. 2001; 20(17–18):2709–2722. [PubMed: 11523078]
- Agresti, A. An Introduction to Categorical Data Analysis. Vol. 423. Wiley-Interscience; 2007.
- Amarasingham, A. PhD thesis. Brown University Division of Applied Mathematics; 2004. Statistical Methods for the Assessment of Temporal Structure in the Activity of the Nervous System.
- Amarasingham A, Chen TL, Geman S, Harrison MT, Sheinberg DL. Spike count reliability and the Poisson hypothesis. *The Journal of Neuroscience*. 2006; 26:801–809. [PubMed: 16421300]
- Amarasingham A, Harrison MT, Hatsopolous N, Geman S. Conditional modeling and the jitter method of spike re-sampling. *Journal of Neurophysiology*. 2012; 107:517–531. [PubMed: 22031767]

- Amari, S.; Nagaoka, H. *Methods of Information Geometry*. Vol. 191. AMS Bookstore; 2000.
- Amari S, Nakahara H, Wu S, Sakai Y. Synchronous firing and higher-order interactions in neuron pool. *Neural Computation*. 2003; 15:127–142. [PubMed: 12590822]
- Besag J, Clifford P. Generalized Monte Carlo significance tests. *Biometrika*. 1989; 76(4):633–642.
- Brody CD. Slow covariations in neuronal resting potentials can lead to artefactually fast cross-correlations in their spike trains. *Journal of Neurophysiology*. 1998; 80:3345–3351. [PubMed: 9862930]
- Canfield ER, Greenhill C, McKay BD. Asymptotic enumeration of dense 0–1 matrices with specified line sums. *J Comb Theory A*. 2008; 115(1):32–66.
- Churchland MM, Yu BM, Sahani M, Shenoy KV. Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Current Opinion in Neurobiology*. 2007; 17:609–618. [PubMed: 18093826]
- Cohen MR, Kohn A. Measuring and interpreting neuronal correlations. *Nature Neuroscience*. 2011; 14:811–819.
- Davison, AC.; Hinkley, DV. *Bootstrap Methods and Their Application*. Cambridge University Press; Cambridge, UK: 1997.
- Engel A, Singer W. Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*. 2001; 5(1):16–25. [PubMed: 11164732]
- Frey, BJ. *Graphical Models for Machine Learning and Digital Communication*. MIT Press; Cambridge, MA: 1998.
- Fujisawa S, Amarasingham A, Harrison M, Buzsáki G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience*. 2008; 11:823–833.
- Furukawa S, Middlebrooks JC. Cortical representation of auditory space: Information-bearing features of spike patterns. *Journal of Neurophysiology*. 2002; 87:1749–1762. [PubMed: 11929896]
- Gale D. A theorem on flows in networks. *Pacific Journal of Mathematics*. 1957; 7:1073–1082.
- Gerstein GL, Bedenbaugh P, Aertsen AM. Neuronal assemblies. *IEEE Transactions on Biomedical Engineering*. 1989; 36:4–14. [PubMed: 2646211]
- Grün S, Diesmann M, Aertsen A. Unitary events in multiple single-neuron spiking activity: II. Nonstationary data. *Neural Computation*. 2002; 14:81–119. [PubMed: 11747535]
- Gütig R, Aertsen A, Rotter S. Analysis of higher-order neuronal interactions based on conditional inference. *Biological Cybernetics*. 2003; 88:352–359. [PubMed: 12750897]
- Haider B, McCormick DA. Rapid neocortical dynamics: Cellular and network mechanisms. *Neuron*. 2009; 62:171–189. [PubMed: 19409263]
- Harrison, MT. PhD thesis. Brown University Division of Applied Mathematics; 2005. *Discovering Compositional Structures*.
- Harrison MT. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*. 2012; 99:57–69. [PubMed: 23049134]
- Harrison MT. Accelerated spike resampling for accurate multiple testing controls. *Neural Computation*. 2013; 25:418–449. [PubMed: 23148410]
- Harrison MT, Amarasingham A, Geman S. Jitter methods for investigating spike train dependencies. *Computational and Systems Neuroscience Abstracts* III-17. 2007
- Harrison, MT.; Amarasingham, A.; Kass, RE. Statistical identification of synchronous spiking. In: Lorenzo, PD.; Victor, J., editors. *Spike Timing: Mechanisms and Function*. Taylor & Francis; 2013.
- Harrison MT, Geman S. A rate and history-preserving resampling algorithm for neural spike trains. *Neural Computation*. 2009; 21(5):1244–1258. [PubMed: 19018703]
- Harrison, MT.; Miller, JW. Importance sampling for weighted binary random matrices with specified margins. 2013. arXiv:1301.3928
- Kass RE, Kelly RC, Loh WL. Assessment of synchrony in multiple neural spike trains using loglinear point process models. *The Annals of Applied Statistics*. 2011; 5:1262–1292. [PubMed: 21837263]
- Kelly RC, Kass RE. A framework for evaluating pairwise and multiway synchrony among stimulus-driven neurons. *Neural Computation*. 2012; 24:2007–2032. [PubMed: 22509967]
- Lehmann, EL.; Romano, JP. *Testing Statistical Hypotheses*. Springer; 2005.

- Long JD II, Carmena JM. A statistical description of neural ensemble dynamics. *Frontiers in Computational Neuroscience*. 2011;5. [PubMed: 21369354]
- Louis S, Gerstein GL, Grün S, Diesmann M. Surrogate spike train generation through dithering in operational time. *Frontiers in Computational Neuroscience*. 2010;4. [PubMed: 20428513]
- Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976; 63:655–660.
- Martignon L, Deco G, Laskey K, Diamond M, Freiwald W, Vaadia E. Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*. 2000; 12:2621–2653. [PubMed: 11110130]
- Martignon L, Von Hasse H, Grün S, Aertsen A, Palm G. Detecting higher-order interactions among the spiking events in a group of neurons. *Biological Cybernetics*. 1995; 73:69–81. [PubMed: 7654851]
- Miller JW, Harrison MT. Exact sampling and counting for fixed-margin matrices. *Annals of Statistics*. 2013; 41:1569–1592.
- Montani F, Ince RA, Senatore R, Arabzadeh E, Diamond ME, Panzeri S. The impact of high-order interactions on the rate of synchronous discharge and information transmission in somatosensory cortex. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2009; 367:3297–3310.
- Netoff TI, Schiff SJ. Decreased neuronal synchronization during experimental seizures. *The Journal of Neuroscience*. 2002; 22(16):7297–7307. [PubMed: 12177225]
- Okun M, Yger P, Marguet SL, Gerard-Mercier F, Benucci A, Katzner S, Busse L, Carandini M, Harris KD. Population rate dynamics and multineuron firing patterns in sensory cortex. *The Journal of Neuroscience*. 2012; 32:17108–17119. [PubMed: 23197704]
- Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky E, Simoncelli EP. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*. 2008; 454:995–999. [PubMed: 18650810]
- Reid N. The roles of conditioning in inference. *Statistical Science*. 1995; 10:138–157.
- Roudi Y, Nirenberg S, Latham PE. Pairwise maximum entropy models for studying large biological systems: When they can work and when they can't. *PLoS Computational Biology*. 2009; 5:e1000380. [PubMed: 19424487]
- Ryser H. Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics*. 1957; 9:371–377.
- Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*. 2006; 440:1007–1012. [PubMed: 16625187]
- Shimazaki H, Amari S, Brown EN, Grün S. State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS Computational Biology*. 2012; 8(3):e1002385. [PubMed: 22412358]
- Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, Litke AM, Chichilnisky E. The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience*. 2006; 26:8254–8266. [PubMed: 16899720]
- Smith MA, Kohn A. Spatial and temporal scales of neuronal correlation in primary visual cortex. *The Journal of Neuroscience*. 2008; 28:12591–12603. [PubMed: 19036953]
- Staudte B, Rotter S. Higher-order correlations in non-stationary parallel spike trains: Statistical modeling and inference. *BMC Neuroscience*. 2009; 10(Suppl 1):P108.
- Truccolo W, Donoghue JA, Hochberg LR, Eskandar EN, Madsen JR, Anderson WS, Brown EN, Halgren E, Cash SS. Single-neuron dynamics in human focal epilepsy. *Nature Neuroscience*. 2011; 14:635–641.
- Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J Neurophysiol*. 2005; 93:1074–1089. [PubMed: 15356183]
- Truccolo W, Hochberg LR, Donoghue JP. Collective dynamics in human and monkey sensorimotor cortex: Predicting single neuron spikes. *Nature Neuroscience*. 2010; 13:105–111.
- Uhlhaas PJ, Singer W. Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology. *Neuron*. 2006; 52(1):155–168. [PubMed: 17015233]

- Vaadia E, Haalman I, Abeles M, Bergman H, Prut Y, Slovin H, Aertsen A. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature*. 1995; 373(6514):515–518. [PubMed: 7845462]
- Ventura V, Cai C, Kass RE. Statistical assessment of time-varying dependence between two neurons. *Journal of Neurophysiology*. 2005a; 94:2940–2947. [PubMed: 16160097]
- Ventura V, Cai C, Kass RE. Trial-to-trial variability and its effect on time-varying dependency between two neurons. *Journal of Neurophysiology*. 2005b; 94:2928–2939. [PubMed: 16160096]

A Monte Carlo Appendix

Monte Carlo p-values

Monte Carlo approximation of p-values is an important topic in statistics. Here we briefly mention three common approaches that can be used when the null hypothesis specifies a unique distribution P (possibly data dependent) under which to compute p-values. For STCI, P will be $f_0(\cdot|c)$ where c is the known coarsening of the original data. Conditional inference is one of the most common ways to reduce a large, composite null hypothesis to a single distribution; bootstrap hypothesis testing is another. The target p-value is $\alpha(X)$ defined by

$$\alpha(x) \stackrel{\text{def}}{=} \mathbb{P}(S(X) \geq S(x)) \stackrel{\text{def}}{=} \mathbb{E}[\mathbb{1}\{S(X) \geq S(x)\}],$$

where S is a test statistic and where the probability and expectation are over a random variable X that has distribution P . The p-value has the important property that

$$\mathbb{P}(\alpha(X) \leq \varepsilon) \leq \varepsilon \quad (21)$$

for all $\varepsilon \in [0, 1]$, whenever X has distribution P , so rejecting the null hypothesis when we observe $\alpha(X) \leq \varepsilon$ has at most ε chance of error (type I error). If $\alpha(X)$ cannot be computed exactly, then we can approximate it with some $\hat{\alpha}(X)$. We not only want $\hat{\alpha}(X) \approx \alpha(X)$, but we also want $\hat{\alpha}(X)$ to satisfy Equation (21) when the null hypothesis is true, which means that $\hat{\alpha}(X)$ is itself a **valid** p-value. Using valid p-values, as opposed to just approximations of p-values that are not themselves valid, is always important when applying multiple testing corrections, which are sensitive to small approximation errors, or when using importance sampling or MCMC, both of which can have large approximation errors that are difficult to diagnose.

The first method is **direct sampling**, which was described in Section 5. For direct sampling, we obtain iid observations $Y_1, \dots, Y_n$ from P and compute the approximate p-value

$$\hat{\alpha}(X) = \hat{\alpha}_n(X; Y_1, \dots, Y_n) = \frac{1 + \sum_{k=1}^n \mathbb{1}\{S(Y_k) \geq S(X)\}}{1 + n},$$

which is a valid p-value for any n and has $\hat{\alpha}_n(x) \rightarrow \alpha(x)$ as $n \rightarrow \infty$. For a discussion of $\hat{\alpha}$, see Davison and Hinkley (1997); Harrison (2012, 2013).

For importance sampling we obtain iid observations $Y_1, \dots, Y_n$ from a distribution Q whose support includes the support of P , meaning that any set A with $P(A) > 0$ also has $Q(A) > 0$. The approximate p-value is defined by

$$\hat{\alpha}_Q(X) = \hat{\alpha}_{Q,n}(X; Y_1, \dots, Y_n) = \frac{w(X) + \sum_{k=1}^n w(Y_k) \mathbb{1}\{S(Y_k) \geq S(X)\}}{w(X) + \sum_{k=1}^n w(Y_k)}, \quad (22)$$

where $w(y) = P(y)/Q(y)$ is called the importance weight function. This definition is for the discrete case where P and Q are pmfs, but it can be generalized to arbitrary distributions. Importance sampling is analogous to the direct sampling approximation, except that the Monte Carlo observations are re-weighted by their self-normalized importance weights. Note that the original observation is again included among the resamples and that its importance weight is included when computing the self-normalized importance weights. For the special case where $Q \equiv P$, that is, the special case of direct sampling, then $w \equiv 1$ and the importance sampling approximation is identical to the direct sampling approximation. In many cases, w only known up to constant of proportionality, meaning that we can compute a function $w(\tilde{y})$ with the known property that $w(\tilde{y}) = \kappa w(y)$ for all y , but for which the value of κ is unknown and cannot be efficiently computed. In this case we can still compute $\hat{\alpha}_Q$ by using w in place of w , because the unknown κ cancels in the numerator and denominator of Equation (22). As before $\hat{\alpha}_Q(X)$ is a valid p-value for any n and has $\hat{\alpha}_{Q,n}(x) \rightarrow \alpha(x)$ as $n \rightarrow \infty$. For a discussion of $\hat{\alpha}_Q$, see Harrison (2012, 2013).

For MCMC we first choose a random number D uniformly from the set $\{0, 1, \dots, n\}$ and then, given D , we independently obtain a sequence of $n + 1$ observations

$$Y_{-D}, \dots, Y_0, \dots, Y_{n-D} \\ \searrow_X$$

from an ergodic Markov chain whose unique stationary distribution is P , conditioned on the event $Y_0 = X$. This is usually done by starting the desired Markov chain at $Y_0 = X$ and running it both forward and backwards in time for the designated number of steps in each direction. The approximate p-value is defined by

$$\hat{\alpha}_{MC}(X) = \hat{\alpha}_{MC,n}(X; Y_{-D}, \dots, Y_{n-D}) = \frac{\sum_{k=-D}^{n-D} \mathbb{1}\{S(Y_k) \geq S(X)\}}{1+n}, \quad (23)$$

which is a valid p-value for any n and has $\hat{\alpha}_{MC,n}(x) \rightarrow \alpha(x)$ as $n \rightarrow \infty$. For a discussion of $\hat{\alpha}_{MC}$, see Besag and Clifford (1989).

Monte Carlo sampling for generic space-time binnings

In Section 5 we discuss sampling from $f_0^\Omega(\cdot|c)$ for a space-time partition. If Ω is not a partition, then the situation is much more complicated. In some cases, the next theorem can help simplify things.

Theorem 7

Suppose Ω is a space-time binning and Ω'' is a space-time partition with the property that each $\omega \in \Omega$ is a subset of some $\omega'' \in \Omega''$. If $\mathbf{X}$ satisfies H_0^Ω , then given $\mathbf{C}^\Omega(\mathbf{X})$, the random variables $(\mathbf{X}_{\omega''} : \omega'' \in \Omega'')$ are mutually conditionally independent with

$$\mathbb{P}(\mathbf{X}_{\omega''} = \mathbf{x}_{\omega''} | \mathbf{C}^\Omega(\mathbf{X}) = \mathbf{c}^\Omega) = \frac{\mathbb{1}\{C_\omega(\mathbf{x}_\omega) = c_\omega \text{ for all } \omega \in \Omega \text{ with } \omega \subseteq \omega''\}}{\#\{\mathbf{y}_{\omega''} : C_\omega(\mathbf{y}_\omega) = c_\omega \text{ for all } \omega \in \Omega \text{ with } \omega \subseteq \omega''\}}.$$

Remark—A key special case is $\Omega = \text{TA}-(k\Gamma, \Gamma)$ and $\Omega'' = \text{ALL}-k\Gamma$ for some $k \geq 1$.

Corollary 8

Suppose Ω is a space-time partition and that $\mathbf{X}$ satisfies H_0^Ω . Then the random variables $(\mathbf{X}_\omega : \omega \in \Omega)$ are mutually conditionally independent and each $\mathbf{X}_\omega$ is uniform subject to the value of $C_\omega(\mathbf{X}_\omega)$.

A proof of the theorem is in the Mathematical Appendix. The corollary follows by taking $\Omega'' = \Omega$ (in which case we must choose $\omega'' = \omega$). Note that we can always find a partition Ω'' satisfying the assumptions of Theorem 7 by taking $\Omega'' = \{\mathcal{F}\}$, i.e., Ω'' contains a single space-time bin that contains all possible fine space-time bins. This trivial choice of Ω'' is not interesting or useful, however. The theorem becomes most useful if Ω'' can be chosen to be a fine partition, so that each $\mathbf{X}_{\omega''}$ is a low-dimensional random variable.

Theorem 7 says that for generating a random observation $\mathbf{Y}$ from $f_0^\Omega(\cdot | \mathbf{c}^\Omega)$ we can independently sample $\mathbf{Y}_{\omega''}$ for each $\omega'' \in \Omega''$. Furthermore, the distribution of $\mathbf{Y}_{\omega''}$ is still uniform subject to the relevant constraints, namely, that $C_\omega(\mathbf{Y}_\omega) = c_\omega$ for each of the $\omega \subseteq \omega''$. It may still be a challenging problem to sample $\mathbf{Y}_{\omega''}$, particularly if ω'' is large and the $\omega \subseteq \omega''$ overlap in complicated ways.

Monte Carlo sampling for TA-(, 1)

Let $\Omega = \text{TA}-(, 1)$ and let $\Omega'' = \text{ALL}-$ so that the assumptions of Theorem 7 are satisfied.

To generate a random observation $\mathbf{Y}$ from $f_0^\Omega(\cdot | \mathbf{c}^\Omega)$, we can independently generate each $\mathbf{Y}_{\omega''}$ uniformly subject to its respective constraints. So we need only describe how to generate $\mathbf{Y}_{\omega''}$.

A space-time bin $\omega'' \in \Omega''$ consists of a large $N \times$ rectangle of fine space-time bins containing all N spike trains over a -length interval of time, so that $\mathbf{Y}_{\omega''}$ is an $N \times$ binary matrix. The collection of measurements $(C_\omega(\mathbf{Y}) : \omega \in \Omega, \omega \cup \omega'')$ correspond to the sequences of row sums and column sums of $\mathbf{Y}_{\omega''}$. Generating a random $\mathbf{Y}_{\omega''}$ is equivalent to generating an $N \times$ binary matrix selected uniformly at random from the set of binary matrices with specified row and column sums. The particular choice of row and column sums is specified in $\mathbf{c}^\Omega$, and for conditional inference this will be the relevant row and column sums taken from the original data.

We have reduced the problem of sampling from $f_0^{\text{TA}-(\Delta,1)}$ to the sub-problem of uniform sampling from binary matrices with specified row and column sums. This sub-problem is a classical and well-studied problem in statistics, combinatorics, and theoretical computer science. The algorithm from Miller and Harrison (2013) can be used for direct sampling. It is practical for small $N \times$ or very sparse data. For human seizure data, which tends to be sparse, we have used it successfully on large datasets where $N \approx 100$ and 25.

The algorithm from Harrison and Miller (2013) can be used for importance sampling. It is faster and scales to much larger problems. In many cases it is nearly indistinguishable from the ideal case of direct sampling. Equation (22) must be used for p-values, so we will briefly discuss how to compute the importance weight function w . When using importance sampling, each $N \times$ submatrix $Y_{\omega'}$ will have two associated probability distributions: $P_{\omega'}$, which is the target uniform distribution subject to the constraints, and $Q_{\omega'}$, which is the importance sampling proposal distribution that is used to generate resamples. The overall importance weight function needed for Equation (22) is

$$w(\mathbf{y}) \stackrel{\text{def}}{=} \frac{\prod_{\omega'' \in \Omega''} P_{\omega''}(\mathbf{y}_{\omega''})}{\prod_{\omega'' \in \Omega''} Q_{\omega''}(\mathbf{y}_{\omega''})} \prod_{\omega'' \in \Omega''} \frac{\{C_{\omega}(\mathbf{y}_{\omega}) = c_{\omega} \text{ for all } \omega \in \Omega \text{ with } \omega \subseteq \omega''\}}{\kappa_{\omega''} Q_{\omega''}(\mathbf{y}_{\omega''})} \quad (24)$$

$$= \frac{\{\mathbf{y} \text{ satisfies the constraints in } \mathbf{c}^{\Omega}\}}{\kappa \prod_{\omega'' \in \Omega''} Q_{\omega''}(\mathbf{y}_{\omega''})}$$

where $\kappa_{\omega''}$ is the denominator from the distribution in Theorem 7 and $\kappa \stackrel{\text{def}}{=} \prod_{\omega''} \kappa_{\omega''}$. The importance sampling algorithm in Harrison and Miller (2013) will compute the value of $Q_{\omega''}(Y_{\omega''})$ for each sampled submatrix, as well as, the value of $Q_{\omega''}(X_{\omega''})$ for each submatrix of the original data. It always returns a $\mathbf{Y}$ that satisfies the stated constraints, so the indicator function in the final expression for w will always evaluate to one. The unknown value of κ can be ignored since it cancels in the numerator and denominator of Equation (22).

The blocked Gibbs sampling algorithm from Besag and Clifford (1989) can be used for MCMC, in which case Equation (23) should be used for p-values. The Gibbs sampling algorithm is the same backwards as it is forwards, as is typical for Gibbs sampling algorithms designed for the uniform distribution. Direct sampling and importance sampling are generally preferable to MCMC, however the advantage of MCMC in this case is that the sampling algorithm described in Besag and Clifford (1989) is easy to implement and does not require custom software. It is run simultaneously on each of the submatrices indexed by ω' , and the results are concatenated together.

Monte Carlo sampling for TA-(, Γ)

Let $\Omega = \text{TA}-(, \Gamma)$ for $= k\Gamma$ for some integer $k > 1$ and let $\Omega'' = \text{ALL}-$ so that the assumptions of Theorem 7 are satisfied. The special case $\Gamma = 1$ is treated in the previous section. As in that section, to generate a random observation $\mathbf{Y}$ from $f_0^{\Omega}(\cdot | \mathbf{c}^{\Omega})$, we can independently generate each $Y_{\omega''}$ uniformly subject to its respective constraints. So we need only describe how to generate $Y_{\omega''}$. As before, $Y_{\omega''}$ is an $N \times$ binary matrix and the relevant constraints from $\mathbf{C}$ specify the row sums of $Y_{\omega''}$. Let us denote these row sums as the $N \times 1$

vector $\mathbf{a}$. The constraints also specify the total sum of the matrix over $N \times \Gamma$ blocks. There are k of these blocks and we will use the $1 \times k$ vector $\mathbf{d}$ to denote the sum over each block. When $\Gamma = 1$ (so that $k = \Delta$) these are just the column sums, which we addressed in the previous section. When $\Gamma > 1$, we can conceptualize the constraints as specifying a set of possible column sums, namely,

$$\mathcal{M} \stackrel{\text{def}}{=} \{\mathbf{b} \in \{0, 1, \dots, N\}^\Delta : \sum_{j=(\ell-1)\Gamma+1}^{\ell\Gamma} b_j = d_\ell, \ell=1, \dots, k\}.$$

To sample $\mathbf{Y}_{\omega'}$ we can first sample the vector of column sums from $\mathcal{M}$ and then sample $\mathbf{Y}_{\omega'}$ given these column sums. This latter problem was discussed in the previous section, so we need only address how to sample a column sum from $\mathcal{M}$.

Since $\mathbf{Y}_{\omega'}$ is chosen uniformly, the distribution over column sums in $\mathcal{M}$ will weight each $\mathbf{b} \in \mathcal{M}$ according to how many binary matrices (with row sums of $\mathbf{a}$) have column sums of $\mathbf{b}$. Computing these weights for each element of $\mathcal{M}$ and then sampling exactly is computationally prohibitive. Instead we develop an importance sampling algorithm. Define the pmf q over $\{0, 1, \dots, N\}$ as

$$q(\mathbf{b}) \stackrel{\text{def}}{=} F(\mathbf{b}) \left[\prod_{\ell=1}^k \mathbb{1}\{\sum_{j=1}^{\ell\Gamma} b_j = \sum_{i=1}^{\ell} d_i\} \right] \left[\prod_{j=1}^{\Delta} \mathbb{1}\{\sum_{i=1}^j b_i \leq \sum_{i=1}^j a_i^*\} \right], \quad (25)$$

where $\mathbf{a}^*$ is the conjugate sequence of $\mathbf{a}$ defined by $a_i^* = \#\{j: a_j \geq i\}$, and where $F(\mathbf{b})$ is the following combinatorial approximation for the number of binary matrices with row sums $\mathbf{a}$ and column sums $\mathbf{b}$ (Canfield et al., 2008):

$$H(\mathbf{b}) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^N \binom{\Delta}{a_i} \prod_{j=1}^{\Delta} \binom{N}{b_j}}{\binom{N\Delta}{D}} \exp \left(-\frac{1}{2} \left(1 - \frac{\sum_{i=1}^N (a_i - \frac{D}{N\Delta})^2}{D(1 - \frac{D}{N\Delta})} \right) \left(1 - \frac{\sum_{j=1}^{\Delta} (b_j - \frac{D}{N\Delta})^2}{D(1 - \frac{D}{N\Delta})} \right) \right)$$

for $D \stackrel{\text{def}}{=} \sum_i a_i = \sum_{\ell} d_{\ell} = \sum_j b_j = \sum_i a_i^*$, which is known and enforced by the first set of indicator functions in Equation (25). It is straightforward to verify that $F(\mathbf{b})$ factors as a product $\prod_j F_j(b_j)$ over the support of q as in (Harrison and Miller, 2013). Consequently, we can use dynamic programming (cf Frey (1998); Harrison and Geman (2009); Harrison and Miller (2013)) to efficiently and exactly sample from q . The first set of indicator functions in Equation (25) prohibit exactly those column sums that are not in $\mathcal{M}$ and the second set of indicator functions prohibit some (but perhaps not all) impossible column sums (meaning that there are no binary matrices with these columns sums and with row sums $\mathbf{a}$); see Gale (1957); Ryser (1957); Harrison and Miller (2013). If $F(\mathbf{b})$ was exactly the number of binary matrices with row sums $\mathbf{a}$ and column sums $\mathbf{b}$, then q would give a direct sample from the

distribution of column sums of $\mathbf{Y}_{\omega'}$. Since F is only an approximation (and since F is strictly positive), we can use q as an importance sampling proposal distribution.

To generate a random observation from $\mathbf{Y}_{\omega'}$, we first sample the column sums, say $\mathbf{b}$, from q in Equation (25) and evaluate $q(\mathbf{b})$ (needed later for the importance weight). Then we use the importance sampling algorithm described in the previous section to sample $\mathbf{y}_{\omega'}$ from a pmf $Q_{\omega'}$ for the row sums $\mathbf{a}$ and column sums $\mathbf{b}$. (If there are no such matrices with these row and column sums, then the importance weight is zero and we can stop.) The total probability is $q(\mathbf{b})Q_{\omega'}(\mathbf{y}_{\omega'})$, which is used in place of $Q_{\omega'}(\mathbf{y}_{\omega'})$ in Equation (24) to get the final importance weight. (The values of $\kappa_{\omega'}$ and κ are different in this case, but they are never used. We could in principle use direct sampling for $Q_{\omega'}$ instead of importance sampling, but it would be much slower computationally.) Additional random observations from $\mathbf{Y}_{\omega'}$ repeat this process, beginning with a new sequence of column sums sampled from q . A Matlab implementation of this algorithm is available on the authors' websites.

B Mathematical Appendix

Lemma 9

Let Z be a discrete random variable with distribution $p(z) = \mathbb{P}(Z = z)$ and let V be any function of Z . If $p(z) = h(V(z))$ for all z and some function h , then the conditional distribution of Z given V is uniform, meaning

$$\mathbb{P}(Z=z|V(Z)=v) = \frac{\mathbb{1}\{V(z)=v\}}{\#\{a:V(a)=v\}}$$

for all z and all v with $\mathbb{P}(V(Z) = v) > 0$.

Remark—We can apply the lemma with $Z = \mathbf{X}$ and $V = \mathbf{C}^\Omega$ to prove Theorem 5. That $p^\Omega(\mathbf{x}|\tilde{\boldsymbol{\theta}})$ depends on $\mathbf{x}$ only through $\mathbf{C}(\mathbf{x})$ is apparent by inspection of Equation (14). Hence, the lemma implies

$$f^\Omega(\mathbf{x}|\mathbf{c}, \tilde{\boldsymbol{\theta}}) = \mathbb{P}(\mathbf{X}=\mathbf{x}|\mathbf{C}^\Omega(\mathbf{X})=\mathbf{c}, \tilde{\boldsymbol{\theta}}) = f_0^\Omega(\mathbf{x}|\mathbf{c})$$

regardless of $\tilde{\boldsymbol{\theta}}$. If $\tilde{\boldsymbol{\theta}}$ is deterministic, then $f(\mathbf{x}|\mathbf{c}) = f(\mathbf{x}|\mathbf{c}, \tilde{\boldsymbol{\theta}}) = f_0(\mathbf{x}|\mathbf{c})$. If $\tilde{\boldsymbol{\theta}}$ is random, then $f(\mathbf{x}|\mathbf{c}) = \mathbb{E}[f(\mathbf{x}|\mathbf{c}, \tilde{\boldsymbol{\theta}})] = \mathbb{E}[f_0(\mathbf{x}|\mathbf{c})] = f_0(\mathbf{x}|\mathbf{c})$, where in this case $\mathbb{E}$ denotes the expected value over $\tilde{\boldsymbol{\theta}}$. In either case, $f \equiv f_0$.

Proof—Under the assumptions of the lemma.

$$\begin{aligned} \mathbb{P}(Z=z|V(Z)=v) &= \frac{p(z) \cdot \mathbb{1}\{V(z)=v\}}{\sum_a p(a) \cdot \mathbb{1}\{V(a)=v\}} = \frac{h(V(z)) \cdot \mathbb{1}\{V(z)=v\}}{\sum_a h(V(a)) \cdot \mathbb{1}\{V(a)=v\}} \\ &= \frac{h(v) \cdot \mathbb{1}\{V(z)=v\}}{\sum_a h(v) \cdot \mathbb{1}\{V(a)=v\}} = \frac{\mathbb{1}\{V(z)=v\}}{\sum_a \mathbb{1}\{V(a)=v\}} = \frac{\mathbb{1}\{V(z)=v\}}{\#\{a:V(a)=v\}}. \end{aligned}$$

Lemma 10

Suppose that V and Y are functions of a discrete random variable Z , and that Y is also a function of V , meaning that we can express $Y(z) = b(V(z))$ for all z and for some function b . If the conditional distribution of Z given Y is uniform, meaning

$$\mathbb{P}(Z=z|Y(Z)=y) = \frac{\mathbb{1}\{Y(z)=y\}}{\#\{a:Y(a)=y\}}$$

for all z and y , defining $0/0 = 0$, then the conditional distribution of Z given V is also uniform, namely,

$$\mathbb{P}(Z=z|V(Z)=v) = \frac{\mathbb{1}\{V(z)=v\}}{\#\{a:V(a)=v\}}$$

for all z and v .

Remark—We can apply the lemma with $Z = \mathbf{X}$, $Y = \mathbf{C}^\Omega$, and $V = \mathbf{C}^{\Omega'}$ for a refinement Ω' of Ω to prove Theorem 1. See Equation (9) for a demonstration that $\mathbf{C}^\Omega$ is a function of $\mathbf{C}^{\Omega'}$.

Proof—For any z we have

$$\begin{aligned} p(z) &= \mathbb{P}(Z=z) = \mathbb{P}(Y(Z)=Y(z))\mathbb{P}(Z=z|Y(Z)=Y(z)) \\ &= \mathbb{P}(Y(Z)=Y(z)) \frac{\#\{a:Y(a)=Y(z)\}}{\#\{a:Y(a)=Y(z)\}} = \frac{\mathbb{P}(Y(Z)=Y(z))}{\#\{a:Y(a)=Y(z)\}}. \end{aligned}$$

The second equality follows since Y is a function of Z , the third uses the assumption that Z is conditionally uniform given Y , and the fourth uses that assumption that Y is a function of V . This final expression for $p(z)$ depends on z only through the value of $V(z)$, so Lemma 9 implies that Z is also conditionally uniform given V .

Lemma 11

Suppose that V and Y are functions of a discrete random variable Z , and that Y is also a function of V , meaning that we can express $Y(z) = b(V(z))$ for all z and for some function b . Then

$$\mathbb{P}(Z=z|V(Z)=v, Y(Z)=y) = \mathbb{P}(Z=z|V(Z)=v)$$

for all z, v, y such that $\mathbb{P}(V(Z)=v, Y(Z)=y) > 0$.

Remark—We can apply the lemma with $Z = \mathbf{X}$, $Y = \mathbf{C}^\Omega$, and $V = \mathbf{C}^{\Omega'}$ for a refinement Ω' of Ω to prove Theorem 3. See Equation (9) for a demonstration that $\mathbf{C}^\Omega$ is a function of $\mathbf{C}^{\Omega'}$.

Proof—Since $\mathbb{P}(V(Z) = v, Y(Z) = y) > 0$, there is at least one z with $V(z) = v$ and $y = Y(z) = b(V(z)) = b(v)$, so in this case

$$\{z: V(z)=v, Y(z)=y\} = \{z: V(z)=v, b(V(z))=b(v)\} = \{z: V(z)=v\},$$

and the conditional probabilities are formally equivalent as stated in the lemma.

Lemma 12

Suppose that $H_0^1, H_0^2, \dots$ is a sequence of null hypotheses with the property that, for each j , H_0^{j+1} must be true whenever H_0^j is true. For each j , let $\alpha_j(Z)$ be a valid p -value for H_0^j , meaning that $\mathbb{P}(\alpha_j(Z) \leq \varepsilon) \leq \varepsilon$ for all $\varepsilon \in [0, 1]$ whenever H_0^j is true. Defining $\alpha_j^+(Z) = \max_{k \leq j} \alpha_k(Z)$ and $J = \{j: H_0^j \text{ is true}\}$, we have

$$\mathbb{P}(\text{there are any false rejections}) = \mathbb{P}(\text{there exists } j \in J \text{ with } \alpha_j^+(Z) \leq \varepsilon) \leq \varepsilon$$

for all $\varepsilon \in [0, 1]$.

Remark—See Marcus et al. (1976) for a more general theory. The sequence of null hypotheses can be either finite or infinite. J is not random, but for any choice of distribution for Z , there will be some set J of true null hypotheses, and the lemma bounds the probability that any of the corresponding maximal p -values are small (which would lead to a false rejection). Taking $Z = X$ and $H_0^j = H_0^{\Omega_j}$ proves Theorem 4. Note that Theorem 1 shows that $H_0^{\Omega_{j+1}}$ must be true whenever $H_0^{\Omega_j}$ is true.

Proof—If J is empty, there is nothing to prove, otherwise, define $j^* \stackrel{\text{def}}{=} \min\{j: j \in J\}$ so that by the assumed structure of the null hypotheses $J = \{j^*, j^* + 1, \dots\}$. Since α_j^+ is increasing in j , we have $\alpha_{j^*}^+ = \min_{j \in J} \alpha_j^+$ and

$$\begin{aligned} \mathbb{P}(\text{there exists } j \in J \text{ with } \alpha_j^+(Z) \leq \varepsilon) &= \mathbb{P}(\min_{j \in J} \alpha_j^+(Z) \leq \varepsilon) = \mathbb{P}(\alpha_{j^*}^+(Z) \leq \varepsilon) \\ &\leq \mathbb{P}(\alpha_{j^*}(Z) \leq \varepsilon) \leq \varepsilon. \end{aligned}$$

Proof of Theorem 6

For each $\omega \in \Omega$, let $A_\omega \subseteq \Omega'$ be a partition of ω chosen from Ω'' . That is, $\cup\{\omega' : \omega' \in A_\omega\} = \omega$ and elements of A_ω are pairwise disjoint. For each $\omega' \in \Omega'$ define

$$\tilde{\theta}_{\omega'} \stackrel{\text{def}}{=} \sum_{\omega \in \Omega} \tilde{\theta}_\omega \mathbb{1}\{\omega' \in A_\omega\},$$

which will be zero if ω' is not used in any of the $A_{\omega'}$. The parameters $\tilde{\theta}$ come from $\tilde{H}_0^\Omega$ in Equation (13). $\tilde{H}_0^{\Omega'}$ is true using the above $\tilde{\theta}$, as shown by this calculation:

$$\begin{aligned} \sum_{\omega' \in \Omega': (i,t) \in \omega'} \tilde{\theta}_{\omega'} &= \sum_{\omega' \in \Omega'} \sum_{\omega \in \Omega} \tilde{\theta}_\omega \mathbb{1}\{\omega' \in A_\omega\} \mathbb{1}\{(i,t) \in \omega'\} \\ &= \sum_{\omega \in \Omega} \tilde{\theta}_\omega \sum_{\omega' \in \Omega'} \mathbb{1}\{(i,t) \in \omega'\} \mathbb{1}\{\omega' \in A_\omega\} \\ &= \sum_{\omega \in \Omega} \tilde{\theta}_\omega \mathbb{1}\{(i,t) \in \omega\} = \sum_{\omega \in \Omega: (i,t) \in \omega} \tilde{\theta}_\omega = \theta_{it}. \end{aligned}$$

Proof of Theorem 7

Note that for each $\omega \in \Omega$ there is a unique $\omega'' \in \Omega''$ with $\omega \subseteq \omega''$. Note also that $\mathbf{X}$ is equivalent to $(\mathbf{X}_{\omega''} : \omega'' \in \Omega'')$ and that $\mathbf{X}_{\omega'}$ refers to completely different entries of $\mathbf{X}$ for different choices of $\omega' \in \Omega''$. Therefore, we can express

$$\begin{aligned} \mathbb{1}\{\mathbf{C}^\Omega(\mathbf{x}) = \mathbf{c}\} &= \prod_{\omega \in \Omega} \mathbb{1}\{C_\omega(\mathbf{x}_\omega) = c_\omega\} = \prod_{\omega'' \in \Omega''} \prod_{\omega \in \Omega: \omega \subseteq \omega''} \mathbb{1}\{C_\omega(\mathbf{x}_\omega) = c_\omega\} \\ &= \prod_{\omega'' \in \Omega''} \underbrace{\mathbb{1}\{C_\omega(\mathbf{x}_\omega) = c_\omega \text{ for all } \omega \in \Omega \text{ with } \omega \subseteq \omega''\}}_{\text{depends only on } \mathbf{x}_{\omega''}}. \end{aligned}$$

Combining this with Equation (5) shows that the conditional distribution of $\mathbf{X}$ given $\mathbf{C}^\Omega$ factors into a product of terms over $\omega'' \in \Omega''$, each term of which depends on the non-overlapping set of variables $\mathbf{X}_{\omega''} = (X_{it} : (i,t) \in \omega'')$. Hence, each $\mathbf{X}_{\omega''}$ is conditionally independent from the others. It also shows that the conditional distribution of each $\mathbf{X}_{\omega''}$ is still uniform subject to the relevant constraints.

Figure 1.

Different types of spike resampling. Left: 84 simultaneously recorded neurons from human middle temporal cortex during a seizure, approximately 10 seconds after seizure onset. See Figure 4 for additional details. Right: 100 repeated trials (ordered by total spike count) of a single neuron recorded from area V1 of the visual cortex of an anesthetized monkey viewing a sinusoidal drifting grating (courtesy of Matt Smith and Adam Kohn) (Smith and Kohn, 2008). A: Original data (X). B–E: Samples from the null hypothesis (H_0^Ω) that fine-precision spiking is otherwise unstructured given the observed coarse-precision spiking (C^Ω) for different choices of the definition of coarse-precision (Ω). The title above each figure indicates which type of coarse-precision (Ω) is used. See Section 2 for details and notation. Although Panels A and E seem quite similar, the exact pattern of spiking in Panel E was

selected from approximately 10^{95} (left) and 10^{2259} (right) different possible patterns of spiking that preserve the respective C^Ω .

Fine-Precision Spiking

Coarse-Precision Spiking

Figure 2.

Conditional inference intuition (schematic diagram). Top: Original fine-precision data for two neurons. Spikes are black bars. Time bins (not shown) are 1 ms in width. Four of the five spikes from the second neuron are precisely synchronous (occur in the same 1 ms time bin) with spikes from the first neuron. Bottom: Coarse-precision data — identical underlying data as the top, but measured using 4 ms time bins (demarcated with light gray bars). Coarse time-bins with spikes have black rectangles; the height of the rectangle corresponds to the number of spikes. The amount of precise synchrony is largely obscured in the coarse-precision data. Even if we are given complete knowledge of $\text{Prob}(\text{coarse spiking})$ in this case, i.e., we somehow know exactly the distribution of coarse-temporal spiking, then we are likely to still have great uncertainty about any processes that influence precise spike timing, i.e., about processes that largely affect $\text{Prob}(\text{fine spiking} \mid \text{coarse spiking})$. On the other hand, almost by definition, coarse temporal processes have little or no effect on $\text{Prob}(\text{fine spiking} \mid \text{coarse spiking})$, which, in this case, controls only the precise arrangement of spikes subject to the coarse-temporal firing rates and which is influenced primarily by the fine temporal processes of interest. Statistical investigations of precise spiking can therefore focus on the conditional distribution of the precise arrangement of spikes in the top panel given the coarsely measured spiking in the bottom panel.

Figure 3.

Schematic diagram of the spatio-temporal coarsening notation. The left column shows different space-time binnings and the right column shows the corresponding coarsening of the data. (We are using “time” to describe the horizontal axis and “space” to describe the vertical axis, although the semantic interpretations of the axes are mathematically irrelevant.) Panel A shows the original data, i.e., the fine space-time bins, $\mathcal{F}$, and the fine-precision data, X , with black indicating one (spike) and white indicating zero (no spike). The space-time binnings in Panels A–D are partitions, which are easy to visual. Panel B (left) shows a generic space-time partition with 7 bins, outlined in dark gray. (The partition is chosen first, before observing the data. The unusually shaped bins are simply for illustrative

purposes.) The fine space-time bins are faintly visible for comparison. Panel B (right) shows the corresponding coarsening, which is simply the spike count in each space-time bin. Shading indicates the spiking rates (spike count divided by space-time bin size; higher rates are darker) in each bin. This provides a useful visualization of the information contained in a coarsening. The pattern of spiking in a space-time bin is otherwise obscured. (Compare the pattern, $X_{\omega 5}$, to the coarsening, $C_{\omega 5}$.) Panel C shows an example of pure temporal coarsening, namely, TIME- for $\Gamma = 4$. Panel D shows the spatio-temporal coarsening ALL- Γ for $\Gamma = 2$. Panel E shows a space-time binning that is not a partition, namely, TA-(4, 2), obtained by the combination of the coarsenings in C and D.

Figure 4.

Multi-scaled spatio-temporal spiking patterns. 84 simultaneously recorded neurons from human middle temporal cortex during a seizure, which begins with the large burst of spiking on the top left. Neurons (vertical axis) are ordered by total spike count over a larger observation interval. Dots indicate the times (horizontal axis) of recorded action potentials (spikes). The three panels show different temporal intervals to illustrate multiple scales of structure in the spiking: 20 seconds (s), 4 s, and 300 milliseconds (ms), from top to bottom. Data was collected at Rhode Island Hospital, Brown Medical School, under local Institutional Review Board approved research protocol.

Figure 5.

Histograms of test statistic values computed over Monte Carlo spike trains for Example 1b.

Top: Using $\Omega = 100$ ms for $\Omega = \text{TIME-}$ we generated $n = 10000$ observations from $f_0(\cdot | C(X))$ and computed the test statistic of the number of 1 ms synchronous pairs, S , for each resampled dataset. The 10001 values of this test statistic (10000 Monte Carlo values and 1 from the original data) are shown in the histogram. $S(X) = 57$ for the original data; shown with dotted line. The fraction of the histogram (shaded darker gray) for values that equal or exceed the observed test statistic is the Monte Carlo p-value. Bottom: Identical, except for $\Omega = 25$ ms.

Figure 6.

Logical structure of the null hypotheses. The box contains the set of all probability distributions for the data. Ω and Ω' are space-time binnings and Ω' is a refinement of Ω , meaning it has finer space-time bins. The probability distributions that are consistent with $H_0^{\Omega'}$ are within the large circle. Within this subset of probability distributions is an even smaller subset consistent with H_0^{Ω} , as shown with the small circle. If H_0^{Ω} is true, meaning that the probability distribution governing the data is within the small circle, then that distribution is necessarily within the large circle, meaning that $H_0^{\Omega'}$ is also true. This relationship is stated formally in Theorem 1. The diagonal ellipse shows the structure of the null hypotheses when specialized to log-linear models, or maximum entropy models, as described in Section 8 below. The ellipse contains those distributions that are consistent with $\tilde{H}_0^{\Omega'}$. The intersection of $\tilde{H}_0^{\Omega'}$ and H_0^{Ω} is exactly $\tilde{H}_0^{\Omega}$. The relationships regarding log-linear models are stated formally in Theorems 5–6.