

Traditional waveform based spike sorting yields biased rate code estimates

Valérie Ventura¹

Department of Statistics and Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Avenue, Baker Hall 132, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, March 6, 2009 (received for review June 12, 2008)

Much of neuroscience has to do with relating neural activity and behavior or environment. One common measure of this relationship is the firing rates of neurons as functions of behavioral or environmental parameters, often called tuning functions and receptive fields. Firing rates are estimated from the spike trains of neurons recorded by electrodes implanted in the brain. Individual neurons' spike trains are not typically readily available, because the signal collected at an electrode is often a mixture of activities from different neurons and noise. Extracting individual neurons' spike trains from voltage signals, which is known as spike sorting, is one of the most important data analysis problems in neuroscience, because it has to be undertaken prior to any analysis of neurophysiological data in which more than one neuron is believed to be recorded on a single electrode. All current spike-sorting methods consist of clustering the characteristic spike waveforms of neurons. The sequence of first spike sorting based on waveforms, then estimating tuning functions, has long been the accepted way to proceed. Here, we argue that the covariates that modulate tuning functions also contain information about spike identities, and that if tuning information is ignored for spike sorting, the resulting tuning function estimates are biased and inconsistent, unless spikes can be classified with perfect accuracy. This means, for example, that the commonly used peristimulus time histogram is a biased estimate of the firing rate of a neuron that is not perfectly isolated. We further argue that the correct conceptual way to view the problem out is to note that spike sorting provides information about rate estimation and vice versa, so that the two relationships should be considered simultaneously rather than sequentially. Indeed we show that when spike sorting and tuning-curve estimation are performed in parallel, unbiased estimates of tuning curves can be recovered even from imperfectly sorted neurons.

clustering | encoding | tuning properties | inconsistent | tuning information

Much of neuroscience has to do with relating neural activity and behavior: how does the brain use its neurons to produce sensory integration, motor coordination, learning, emotions, etc., and how do neurons encode parameters associated with these behaviors? Such questions have been investigated by recording brain activity during behavioral tasks to uncover associations between the two. Different aspects of brain activity are captured by different tools like functional MRI, PET, magnetoencephalography, etc. Here, we consider the spike trains of neurons provided by electrodes implanted in the brain, that is, the sequence of times at which neurons fire action potentials, or spikes. The modulation of neurons' spiking rates by behavioral or environmental covariates is widely accepted to be one way that neurons encode information about these covariates. One common measure of association between behavior and neural activity is therefore the firing rates of neurons as functions of covariates of interest, often called tuning functions or receptive fields.

The spike trains needed to calculate tuning functions are typically obtained from extra-cellular electrodes, that is, from electrodes that are positioned outside of the neurons in the tissue. The signal collected at such electrodes is typically a mixture of the activities of nearby neurons and noise, from which individual neurons' spike trains must be extracted. This extraction process is

known as spike sorting. It is one of the most important data analysis problems in neuroscience, because it has to be undertaken prior to any analysis of neurophysiological data in which more than one neuron is believed to be recorded on a single electrode.

Spike sorting is a clustering problem: neurons produce spikes that have distinct, reproducible waveforms, so that the spikes recorded at an electrode can be clustered into homogeneous groups, which presumably correspond to different neurons. Clustering techniques for spike sorting are many and range from nonparametric approaches such as k-means (1), neural networks (2), to likelihood and Bayesian model-based clustering using mixtures of distributions (3). More extensive references are provided in refs. 4 and 5.

The sequence of first spike sorting based on waveforms, then estimating tuning functions, has long been the accepted way to proceed. However, the covariates c that modulate the neurons' firing also contain information about spike identities. To see this, consider an electrode that records two neurons. Imagine that one neuron spikes only when $c_1 < c < c_2$, and the other only when $c_2 < c < c_3$, so that they never spike together. A spike recorded at the electrode will be assigned to one of the two neurons based on features of its waveform, perhaps in error if waveform clusters overlap. But we cannot make a mistake if we use c for spike sorting. Indeed, if $c_1 < c < c_2$ when a spike is detected at the electrode, then the spike must have been produced by neuron 1. If $c_2 < c < c_3$, then it is necessarily neuron 2 that spiked.

Because they ignore the information in rate modulating covariates c , current spike sorters are suboptimal. But what is more troubling is that their misclassification rates are functions of c , so that spikes are not misclassified at random. This is intuitively problematic if the goal is to estimate how neurons are modulated by c . Our first contribution is a proof that tuning functions estimated from spikes sorted based on waveforms are biased and inconsistent, unless spikes can be classified with perfect accuracy. Our second contribution is the formulation of a clustering approach that incorporates tuning information, and which yields unbiased tuning function estimates.

1. Background and Results

Consider an electrode that records I waveform generators. Generators are either neurons, or sources of noise such as static discharges, fluctuations in the local field potential, etc. When the bandpassed voltage of the electrode exceeds a chosen threshold at time t , we record a snippet of measurements a_t , which may correspond to a real spike, to noise, or to some combination of spikes and noise. Note that for simplicity and without loss of generality, all generators and their waveforms may be referred to as neurons and spikes in the rest of the article, unless noted otherwise. The

Author contributions: V.V. designed research, performed research, and wrote the paper.

The author declares no conflict of interest.

¹E-mail: vventura@stat.cmu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0901771106/DCSupplemental.

methods described here apply in general so the recorded waveforms a_t can also be reduced to sets of features such as principal components (PCs) (6) or wavelet coefficients (7). We denote by \mathcal{X} the set of I -dimensional binary vectors $x = (x_1, \dots, x_I)$ that give all $(2^I - 1)$ possible subsets of the I generators being active alone or approximately together to produce a suprathreshold event at the electrode. For example, when $I = 2$, x can take $2^I - 1 = 3$ values, $(1, 0) \equiv 10$, $(0, 1) \equiv 01$ or $(1, 1) \equiv 11$, which corresponds to generators 1 and 2 being active alone and together. Finally, we denote by c_t the value at time t of the covariates thought to modulate the neurons' firing rates, $\lambda_i(c_t)$, $i = 1, \dots, I$.

1.1. Background. Traditional spike sorting using waveforms. Neurons fire spikes that have characteristic waveforms, but, because the voltage of an electrode is noisy, recorded waveforms do not match their true waveforms exactly, but rather arise from distributions f_x centered around them. Either implicitly or explicitly, all spike-sorting methods assume that suprathreshold measurements a_t originate from a mixture distribution

$$f(a) = \sum_{x \in \mathcal{X}} \pi_x f_x(a), \quad [1]$$

where π_x is the proportion of events produced by generator combination x so that $\sum_{x \in \mathcal{X}} \pi_x = 1$. Eq. 1 makes no assumption. It simply states that, given a suprathreshold event at the electrode, the probability that it was produced by x is π_x , and if so, its measurement a arises from f_x . Eq. 1 can be visualized by plots of the data. For example, overlaying raw voltage measurements or plotting their PCs against one another will reveal, more or less clearly, clusters that are each a random sample of a component distribution f_x . Spike sorting effectively consists of separating these clusters, so that all spikes within each domain can be assumed to have originated from the same f_x .

Many methods exist to find cluster boundaries (4), the simplest being to draw them by hand. Another is to make explicit use of Eq. 1. First, Bayes rule yields the probability that a suprathreshold event with measurement a was produced by combination $x \in \mathcal{X}$,

$$P(x | a) = \frac{\pi_x f_x(a)}{f(a)}, \quad [2]$$

where the denominator is Eq. 1 and the numerator are its summands. The event is then assigned to the combination x^* that maximizes $P(x | a)$,

$$x^*(a) = \arg \max_{x \in \mathcal{X}} P(x | a), \quad [3]$$

with corresponding allocation for generator $i = 1, \dots, I$ the i th component of $x^*(a)$. Although it is not immediately obvious, this procedure also consists of drawing cluster boundaries. For example, the boundaries implied by Eq. 3 are linear or quadratic when the f_x are Gaussians with equal or unequal covariance matrices (8). This is illustrated in Fig. 1B.

Eq. 3 is known as the optimal classification rule, because it yields the lowest spike misclassification rate (8). The catch is that π_x and f_x must be known. How close to optimal this classification rule is in practice depends, therefore, on the validity of the models we select for f_x and on how well we can estimate them from data. At present, f_x are most commonly assumed to be Gaussians (3), although t distributions might be more suitable (9, 10).

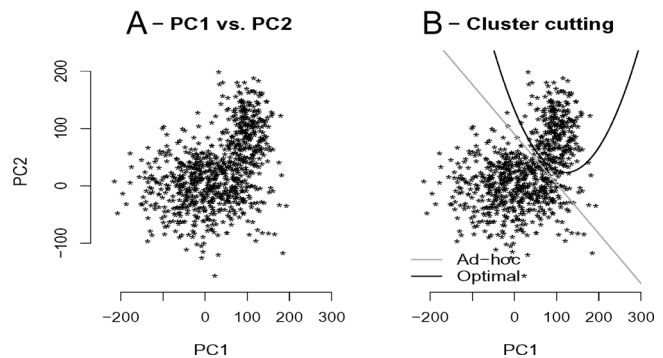


Fig. 1. Spike sorting, toy example. Two neurons are recorded by an electrode. Neuron 1 spikes only when $0 < c < 3$, neuron 2 only when $3 < c < 6$. The first two PCs of their waveforms are simulated from bivariate Gaussians: $f_{10} = N((0, 0), \begin{pmatrix} 70^2 & 0 \\ 0 & 40^2 \end{pmatrix})$, $f_{01} = N(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 30^2 & 0 \\ 0 & 40^2 \end{pmatrix})$, and $f_{11} \equiv 0$ since neurons do not spike together. (A) Plot of the two PCs. Each dot represents a spike. We see what could be two overlapping elliptical shaped clusters, or a single peanut-shaped cluster. (B) Cluster boundaries drawn by hand assuming there are two clusters (straight line), and by applying Eq. 3 assuming that f_x are Gaussians.

Soft or probabilistic spike assignments are seldom used for spike sorting, but we consider them because they are crucial to our results. The portion of an event with measurement a we allocate to generator i is

$$y_i^{\text{soft}} = \sum_{x: x_i=1} P(x | a), \quad [4]$$

where the sum is over all combinations x that code for generator i being active ($x_i = 1$). For example, when $I = 2$, if $P(x | a) = 0.1, 0.3$ and 0.6 for $x = 10, 01$, and 11 , respectively, soft assignments allocate $0.1 + 0.6 = 0.7$ and $0.3 + 0.6 = 0.9$ spike to generator 1 and 2, respectively, whereas hard assignments (Eq. 3) allocate one full spike to both.

Traditional estimation of tuning curves. Estimating the neurons' tuning curves $\lambda_i(c)$ involves choosing a model for $\lambda_i(c)$, and regressing the neurons' spike trains $\mathbf{y}_i = (y_{it}, t = 1, \dots, T)$ on $\mathbf{c} = (c_t, t = 1, \dots, T)$. Models for $\lambda_i(c)$ can be parametric, such as cosine functions for motor cortex data, or non-parametric, such as spline smoothers or step functions, which are commonly used to obtain peristimulus time histograms (PSTHs). Different types of regressions are appropriate for different situations. Hard assignments spike trains are binary, so a binary, e.g., logistic, regression should be used. If spike trains are first binned and resulting spike counts regressed on the covariate, as is often done to obtain a PSTH, a Poisson regression should be used. Soft spike trains $\mathbf{y}_i = \mathbf{y}_i^{\text{soft}}$ take values in $[0, 1]$, so binary regression is no longer appropriate. In that case we apply a transformation to $\mathbf{y}_i^{\text{soft}}$, which maps $[0, 1]$ onto $[-\infty, +\infty]$, e.g., logit or probit transformations, so that the transformed $\mathbf{y}_i^{\text{soft}}$ becomes amenable to ordinary regression.

Spike sorting incorporating tuning information. The inputs to any clustering procedure are vectors of features that characterize the data, e.g., waveform features in the context of spike sorting. The simplest such procedure relies on a plot of these inputs to cut clusters by hand. The same inputs yield estimates of π_x and f_x when clustering is based on Eq. 3. This was illustrated in Fig. 1B.

The information provided by the modulation of neurons by covariates c can be incorporated for spike sorting by supplementing the feature vectors with the values of c concurrent with suprathreshold events. These augmented vectors can then be used as inputs to any clustering procedure. The simplest case is illustrated in Fig. 2: features are plotted against each other and clusters can be cut by the naked eye. Just as different waveforms help identify neurons, so do the dimensions of c that modulate neurons

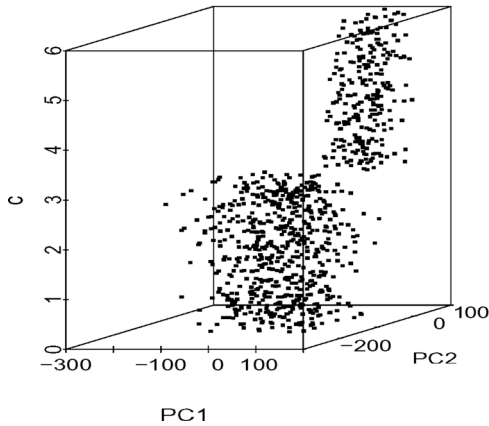
PC1 vs. PC2 vs. c 

Fig. 2. Same data as in Fig. 1, but the PCs are now plotted against the rate-modulating covariate c . The two clusters that previously overlapped (Fig. 1) are now separated along the c axis. Note that better cluster separation can sometimes be achieved by using more PCs. However, the information in covariates about spikes' identities is independent of waveform information, and can be useful for spike sorting however much information waveforms provide.

most differently. In the extreme case when tuning curves share no common support, as in Fig. 2, clusters are separated perfectly. In practice, tuning curves often have common support, so clusters will overlap. In that case misclassified spikes are unavoidable, so we prefer a model-based spike sorter that minimizes misclassifications, as follows. We begin by writing the distribution of the waveforms, as in Eq. 1, but this time we condition on c . This gives

$$f(a | c) = \sum_{x \in \mathcal{X}} \pi_x(c) f_x(a | c), \quad [5]$$

where $\pi_x(c)$ is the probability that, given a suprathreshold event at the electrode when the covariates take value c , generator combination x gave rise to that event, and $f_x(a | c)$ is the distribution of its measurement a . Waveforms are characteristic of the neurons who produce them and do not depend on covariates, so we can reduce $f_x(a | c)$ to the same $f_x(a)$ used in Eq. 1.[†] We then use Bayes rule to calculate the probability that a suprathreshold event with waveform a , detected at the electrode when the covariate has value c , was generated by $x \in \mathcal{X}$,

$$P(x | a, c) = \frac{\pi_x(c) f_x(a)}{f(a | c)}. \quad [6]$$

A hard assignment allocates this spike to combination

$$x^*(a | c) = \arg \max_{x \in \mathcal{X}} P(x | a, c), \quad [7]$$

with corresponding allocation for generator i the i th component, $y_i^{\text{hard}} = x_i^*(a | c)$, while a soft assignment allocates

$$y_i^{\text{soft}} = \sum_{x: x_i=1} P(x | a, c) \quad [8]$$

to generator i , $i = 1, \dots, I$.

This approach is optimal, but requires that Eq. 5 be estimated. This involves choosing models for $f_x(a)$ and $\lambda_i(c)$, and a model

[†] Note that in Eqs. 1 and 5, we could let $f_x(a)$ depend on spiking history, to account for waveform nonstationarities such as spike amplitude decays after short interspike intervals.

for neuron dependencies, which dictates how the $\pi_x(c)$'s relate to the $\lambda_i(c)$'s. To see this, assume that composite, substantially different, waveforms are recorded whenever two generators spike within γ ms of one another. Then the probabilities that generator j contaminates a spike from generator i , $j \neq i$, are $2\gamma\lambda_j(c)$ and $1 - 2\gamma\lambda_j(c)$, respectively, with λ_j expressed in spikes per millisecond. Therefore, if an electrode records $I = 2$ independent generators, the probabilities $\pi_{10}(c)$, $\pi_{01}(c)$, and $\pi_{11}(c)$ that generators spike alone or together are proportional to $\lambda_1(c)[1 - 2\gamma\lambda_2(c)]$, $\lambda_2(c)[1 - 2\gamma\lambda_1(c)]$, and $2\gamma\lambda_1(c)\lambda_2(c)$, respectively. Similar expressions can be derived for larger I (11). If generators are dependent, $\pi_x(c) = \pi_x(c | \mathcal{H})$ are expressed as above, but the λ_i 's now depend on some aspects of spike train histories \mathcal{H} (ref. 12, and references therein). For example, if $I = 2$ neurons cannot spike within s_0 ms of each other, we could reduce spiking history to $\mathcal{H} = \{s_1, s_2\}$, with s_i the time elapsed since neuron i last spiked, and set $\lambda_i(c | s_1, s_2) = \lambda_i^*(c)$ if $s_j > s_0$ and 0 otherwise, $i = 1, 2$, $j \neq i$. Then given a suprathreshold electrode event at t , this would imply $\pi_{10}(c_t | 0, s_2) \propto \lambda_1^*(c_t)$, $\pi_{01}(c_t | s_1, 0) \propto \lambda_2^*(c_t)$, and $\pi_{11}(c_t | 0, 0) = 0$, which matches the intuition that, if neurons cannot spike together, the probability that the spike at t was fired by neuron i is proportional to its rate.

Models for f_x , λ_i , and joint spiking are the same assumptions needed to first spike sort based on Eq. 1, then estimate tuning curves, although they must now be specified all at once rather than sequentially. Note that neurons are typically, or implicitly, assumed to be independent for traditional spike sorting. Similarly, relationships between neurons are typically ignored to estimate tuning curves, unless they are of primary interest (13). A default assumption of independence is still an assumption.

The next step is to estimate Eq. 5. This might first seem impossible, because the $\pi_x(c)$'s depend on the yet unknown $\lambda_i(c)$'s. But, just as the information in waveforms can be harnessed to estimate f_x and π_x in Eq. 1 (3), so the information in the times of suprathreshold events can be harnessed to estimate $\pi_x(c)$, and therefore $\lambda_i(c)$, in Eq. 5 (11). The algorithm in ref. 11 does just that, under the assumptions that f_x are Gaussians, neurons are independent, and spike independently of the past; it accommodates parametric and nonparametric models for λ_i , and can be easily extended to allow other choices for f_x , such as t distributions. This algorithm is an exact expectation-maximization (EM) algorithm of the same type as the algorithm in ref. 3, which outputs the maximum likelihood estimates of f_x and $\pi_x(c)$. Because this algorithm is maximum likelihood based, tools for model and variable selection are readily available: the number of neurons recorded by an electrode can be determined by penalized likelihood (AIC, BIC), and models for f_x , λ_i , and the variables c that significantly modulate spiking rates can be chosen via likelihood ratio tests (12). Several of these issues are illustrated in supporting information (SI) Appendix, and more details are in ref. 11.

With Eq. 5 estimated, suprathreshold events are sorted and tuning curves estimated, as described in the previous section. Note that the estimates of λ_i , $i = 1, \dots, I$, obtained as part of the estimation of Eq. 5 correspond to the estimates obtained by regressing the soft spike trains in Eq. 8 on the covariates c . The proposed spike sorter effectively performs spike sorting and tuning function estimation simultaneously rather than sequentially.

1.2. Results. However basic or sophisticated, regressing a response variable y on covariates c always achieves the same goal: it provides an estimate of $E(Y | c)$, that is an estimate of how y varies as a function of c on average. In our context, where y_i is the spike train of neuron i after spike sorting, we regress y_i on c to estimate its tuning curve $\lambda_i(c)$. This regression therefore makes sense only if $E(Y_i | c) = \lambda_i(c)$.

Theorem 1.1. *Hard assignment spike trains from ad hoc and optimal spike sorters in Eqs. 3 and 7 are such that*

$$E(Y_i^{\text{hard}} | c) \neq \lambda_i(c),$$

unless spikes can be classified with no errors. Hence, hard spike trains do not yield consistent tuning curve estimates unless waveform clusters are perfectly separated.

An estimate is inconsistent if it is systematically biased, and if the bias does not disappear as the sample size increases. In practice, the more the waveform clusters will overlap, the more severe the bias will be, especially if neurons have very different tuning curves, since c then carries substantial information about spikes' identities that is ignored for sorting spikes. Note that tuning-curve estimates are biased even if neurons are not tuned to c . To see that, imagine that an electrode records $I = 2$ neurons, and that the firing rate of neuron 1 is large enough compared with that of neuron 2 so that $\pi_{10}f_{10}(a) > \pi_{01}f_{01}(a)$ and $\pi_{10}f_{10}(a) > \pi_{11}f_{11}(a)$ for all a . Then according to Eq. 3, all spikes recorded at the electrode will be assigned to neuron 1, so that the firing rate estimate of neuron 2 will be zero.

Although we proved *Theorem 1.1* only for hard spike trains from model-based and ad hoc spike sorting, it will likely apply to all spike-sorting procedures that ignore covariate information.

Theorem 1.2. *Soft spike trains obtained from waveform and tuning based spike sorting in Eq. 8 are such that*

$$E(Y_i^{\text{soft}} | c) = \lambda_i(c).$$

Regressing c on such spike trains thus provides unbiased tuning-curve estimates.

Theorem 1.2 is valid regardless of how much the waveform clusters overlap. In practice, if the overlap is substantial, or if the sample size is small, tuning-curve estimates will have large variances, so they may not match the true curves closely. However they match the true curves on average, whatever the sample size.

Theorem 1.2 is unlikely to apply to soft spike trains from other procedures. For example, we prove in *SI Appendix* that the soft spike trains from traditional waveform based optimal spike sorting (Eq. 4) yield biased tuning-curve estimates.

1.3. Illustration. Spike sorting is a central issue for designing algorithms for neural prosthetic control, because spike trains are collected from chronically implanted electrodes. The following toy experiment is inspired by spike train decoding experiments. Other examples can be found in *SI Appendix*

Say that $I = 2$ motor cortex neurons are recorded by an electrode while a monkey traces a 2D circle over the course of 12 s, with hand position at time t , $x_t = 12 \cos(\pi t/12)$ and $y_t = 12 \sin(\pi t/12)$. The velocity amplitude remains constant on this path, so tuning curves can be expressed as functions of directional angle/tuning $d \in [0, 2\pi]$, where $d = \arctan(y/x)$. We assume that neurons spike independently according to Poisson processes with rates $\lambda_i(d) = \exp(2.7 + 2 \cos(d - d_i))$, $i = 1, 2$, and preferred directions $d_1 = 0$ and $d_2 = \pi/2$. These rates have the same profile and little common support, so $\pi_{10} = \pi_{01} = 49.5\%$ are equal, while $\pi_{11} = 1\%$ is small. Without loss of generality we use only one waveform PC for spike sorting, which we simulate from normal distributions with means and variances (6, 1) and (8, 1) for the two neurons ($x = 10$ and 01), respectively. The f_x overlap partially, so spike misclassification errors are unavoidable. We also assume that a single composite waveform is recorded whenever the two neurons spike within $\gamma = 1$ ms of one another, and we simulate the PCs of such waveforms from f_{11} , a normal distribution with mean and variance (10.5, 3).

We simulated the neurons' spike trains from this model during 50 loops of the circular trajectory, and combined them to create the

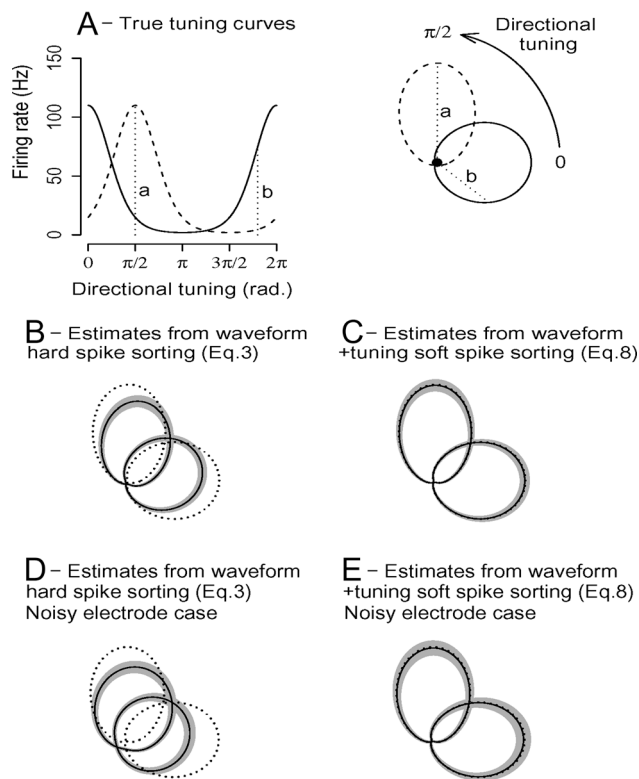


Fig. 3. True tuning curves $\lambda_i(d)$ of two simulated M1 neurons as functions of directional tuning d in Cartesian and circular coordinates, and mean estimates, along with 95% simulation bands. True curves are overlaid in dotted lines. (B and D) Waveform-based hard spike sorting in Eq. 3. (C and E) Waveform and tuning-based soft spike sorting in Eq. 8. (D and E) Same as B and C, but the electrode spike train is corrupted with noise. Only tuning-based soft spike sorting yields consistent estimates. This remains true when the electrode is noisy.

spike train of suprathreshold electrode events. We then simulated the PCs of these events from the f_x specified above. In practice, one would now specify models for the unknown f_x in Eq. 1, and for f_x , $\lambda_i(d)$, and $\pi_x(d)$ in Eq. 5, estimate these models from data, and only then spike sort. Instead we used the *true* f_x , $\lambda_i(d)$, and $\pi_x(d)$, which in practice would correspond to selecting the correct families of models and fitting them to a very large dataset. Our illustrations can thus be reproduced easily, while avoiding estimation issues that are not central to this article.[‡] With the data sorted, we fitted functions linear in $\cos(d)$ and $\sin(d)$ to the neurons' spike trains, the correct family of tuning curve models. We repeated this simulation 100 times so we could calculate the mean estimated firing rates and 95% simulation bands, within which fall 95% of firing rate estimates in repeated simulations. Fig. 3B and C shows true and mean estimated tuning curves with the 95% bands. As expected from *Theorem 1.1*, waveform-based spike sorting yields inconsistent estimates. This happens because spikes are not misclassified at random: when d is close to the preferred direction of neuron 2 (1), almost all spikes recorded at the electrode belong to neuron 2 (1), yet traditional spike sorting classifies them based on waveform information only. Hence, the misclassification rate is highest in the preferred directions of the neurons.

Chronically implanted electrodes with fixed-depth assignments cannot be placed strategically to minimize noise. Noise also tends to increase with time as scar tissue forms around the electrodes.

[‡] Visually indistinguishable results were obtained by actually estimating Eqs. 1 and 5 by using the algorithms in ref. 11.

In that case many suprathreshold electrode events will be noise. To illustrate this, assume that the noise on the previous electrode exceeds the threshold at a constant rate of 158 Hz, which corresponds to normally distributed noise with mean 0, SD equal to the threshold, and voltage sampled every millisecond. The electrode now records $I = 3$ generators, the third being a noise source. We simulated noise events and combined them with the previous electrode spike train. The proportion of pure noise events versus real spikes is $\pi_{001} = 66\%$ to 34% , and a significant proportion of real spikes are corrupted by noise. We simulated the PCs of pure-noise events from f_{001} , a normal distribution with mean 0 and SD 5, and assumed for simplicity that the waveform PCs of clean and noise-corrupted spikes arise from the same distributions, that is we assumed $f_{100} = f_{101}$, $f_{010} = f_{011}$, and $f_{110} = f_{111}$. This choice is not particularly realistic but does not affect our results. We spike sorted the data and estimated tuning curves; they are shown in Fig. 3D and E. The presence of a noise cluster overlapping with the clusters of real spikes results in further bias in the tuning-curve estimates when spike sorting ignores tuning information, whereas soft assignments in Eq. 8 still yield unbiased estimates.

2. Discussion

The standard paradigm in neuroscience is to perform spike sorting first, and then analyze the relationship between the spikes and the putative stimuli. We proved that when spike sorting is conducted without considering the covariates that modulate neurons spiking, estimates of their tuning functions are not consistent. We further argued that the correct conceptual way to view the problem out is to note that spike sorting provides information about rate estimation and vice versa. As a consequence, spike sorting and tuning curve estimates should be performed simultaneously rather than sequentially.

Spike sorting is a clustering problem. The traditional approach is to cluster vectors of waveform features that characterize suprathreshold events. We suggest to supplement these vectors with the covariates thought to modulate neurons spiking. Any clustering method can then be applied to these augmented vectors, although some adjustments might be needed since covariates are not physical measurements like waveforms. In particular, we showed how to adjust model-based automatic spike sorting, and proved that the resulting tuning curve estimates were unbiased. From a statistical view point, the proposed method consists of modeling the waveform measurements as a covariate varying mixture of distributions, whose mixture proportions are themselves mixtures of the unknown rates of point processes. In practice, this approach requires that models be chosen for the distributions of waveforms and joint firing rate model, and that an algorithm be available to estimate them. Such an algorithm was developed in ref. 11, under the assumption that neurons are independent of the past and of other neurons, and the waveform distributions f_x are Gaussian. The more general case is under construction, and will incorporate non-Poisson spiking behavior and nonstationarity of waveforms due to, for example, refractory periods and neurons bursting, as in ref. 14. As for modeling assumptions, they are the very same ones needed for sequential model-based spike sorting and tuning-curve estimation, no more, no less. But one legitimate concern is that the covariate-dependent and soft sorting method suggested as an alternative to pure waveform hard sorting represents a significant shift from current practices. Furthermore, waveform measurements must be saved so they can be included in the estimation of tuning properties, which is far more cumbersome than the current practice of spike sorting just once and being done with it.

Is it worth changing current practices? Tuning-curve bias can produce erroneous scientific conclusions, as illustrated by the examples in *SI Appendix*. But there will be situations where the size of the bias is not large enough, in an absolute sense, to cause concern. Additionally, there are other sources of bias and variability

that might be of greater magnitude. Such sources include the quality of chosen models for waveforms and tuning functions, and the size of the sample available for estimation of classification rules, which not only determines their variances, but also impacts how well the estimating algorithm will converge. In decoding experiments, we showed that tuning-curve estimates may be biased, but they will be consistently so if the same spike-sorting method is used for encoding and decoding. Hence, decoded trajectories will not themselves be biased. However, Fig. 3B and D suggests that tuning-curve estimates are less modulated than the true curves, which should translate into loss of efficiency. Because decoding uses many electrodes at once, the resulting aggregate effect might be substantial. More generally, the size of the bias might be of concern in studies that report results aggregated across neurons.

Determining when to implement the suggested method will require an extensive study, and the development of diagnostic tools, which is beyond the scope of this article. Our main intention here was to bring awareness to the conceptual flaw of waveform-based spike sorting, and to propose a solution.

3. Methods

Proof of Theorem 1.2: For time bins where a spike is recorded at the electrode, which we denote by $Z = 1$, soft spike assignments are such that

$$E(Y_i^{\text{soft}} | Z = 1, c) = E\left(\frac{\sum_{x:x_i=1} \pi_x(c) f_x(A)}{f(A|c)}\right)$$

with expectation with respect to the true waveform distribution. Given c , that distribution is Eq. 5, not Eq. 1. Hence,

$$\begin{aligned} E(Y_i^{\text{soft}} | Z = 1, c) &= \int \left(\frac{\sum_{x:x_i=1} \pi_x(c) f_x(a)}{f(a|c)}\right) f(a|c) da \\ &= \int \sum_{x:x_i=1} \pi_x(c) f_x(a) da = \sum_{x:x_i=1} \pi_x(c) \end{aligned}$$

since densities integrate to one. This summation is over neuron combinations x that have $x_i = 1$, hence, it is the probability that neuron i spiked, given $Z = 1$. Letting Y_i denote the true neuron spike train, we therefore have $E(Y_i^{\text{soft}} | Z = 1, c) = P(Y_i = 1 | Z = 1, c) = \frac{P(Y_i=1, Z=1|c)}{P(Z=1|c)} = \frac{P(Y_i=1|c)}{P(Z=1|c)}$, since $Y_i = 1$ implies $Z = 1$. When no spike is detected at the electrode ($Z = 0$), we set $Y_i^{\text{soft}} = 0$, so that trivially $E(Y_i^{\text{soft}} | Z = 0, c) = 0$ for all c . Then, unconditionally, $E(Y_i^{\text{soft}} | c) = E(Y_i^{\text{soft}} | Z = 1, c)P(Z = 1 | c) + E(Y_i^{\text{soft}} | Z = 0, c)P(Z = 0 | c) = P(Y_i = 1 | c)$, which is the firing rate $\lambda_i(c)$ of neuron i , expressed in units of spikes per the duration of time bins used to discretize the EST. Q.E.D.

Proof of Theorem 1.1: As above we have $E(Y_i^{\text{hard}} | c) = E(Y_i^{\text{hard}} | Z = 1, c)P(Z = 1 | c)$, which will reduce to $\lambda_i(c)$ iff $E(Y_i^{\text{hard}} | c, Z = 1) = P(Y_i = 1 | c)/P(Z = 1 | c)$. Without loss of generality we set $i = 1$, and for simplicity we work with scalar waveform measurements a , and treat the case of $I = 2$ neurons. The proof does extend generally but becomes very cumbersome. Then $E(Y_1^{\text{hard}} | c, Z = 1) = P(Y_1^{\text{hard}} = 1 | c, Z = 1) = 1 - P(Y_1^{\text{hard}} = 0 | c, Z = 1)$, which, for hard spike assignments Eq. 3, simplifies to

$$1 - P[\pi_{01} f_{01}(A) > \pi_{10} f_{10}(A) \ \& \ \pi_{01} f_{01}(A) > \pi_{11} f_{11}(A)] = 1 - \int_{a \in \mathcal{A}} f(a|c) da, \quad [9]$$

where $\mathcal{A} = \{a : \pi_{01} f_{01}(a) > \pi_{10} f_{10}(a) \ \& \ \pi_{01} f_{01}(a) > \pi_{11} f_{11}(a)\}$. The form of \mathcal{A} depends on the configuration of the f_x , $x \in \mathcal{X}$. The six possibilities are shown schematically in Fig. 4. For configuration (Fig. 4A), we have

$$E(Y_1^{\text{hard}} | c, Z = 1) = 1 - \int_{a_1}^{a_2} f(a|c) da = 1 - \sum_{x \in \mathcal{X}} \pi_x(c) [F_x(a_2) - F_x(a_1)], \quad [10]$$

where F_x is the cumulative distribution function of f_x , a_1 is such that $\pi_{01} f_{01}(a_1) = \pi_{10} f_{10}(a_1)$, and a_2 is such that $\pi_{01} f_{01}(a_2) = \pi_{11} f_{11}(a_2)$. Eq. 10 reduces to the required $\sum_{x:x_1=1} \pi_x(c)$ for all c iff $F_{01}(a_2) - F_{01}(a_1) \rightarrow 1$, $F_{10}(a_2) - F_{10}(a_1) \rightarrow 0$, and $F_{11}(a_2) - F_{11}(a_1) \rightarrow 0$. Because a_1 and a_2 are constrained by $\pi_{01} f_{01}(a_1) = \pi_{10} f_{10}(a_1)$ and $\pi_{01} f_{01}(a_2) = \pi_{11} f_{11}(a_2)$, the three limits are obtained only as all the f_x pull away from one another, so that none shares a common support. Conversely, f_x disjoint implies that a_1 and a_2 are such that $F_{10}(a_1) = 1$, $F_{01}(a_1) = F_{11}(a_1) = 0$, $F_{01}(a_2) = F_{10}(a_2) = 1$, and $F_{11}(a_2) = 0$ (this can be seen from Fig. 4A by letting the distributions spread apart), so that $F_{01}(a_2) - F_{01}(a_1) = 1$ and $F_{10}(a_2) - F_{10}(a_1) = F_{11}(a_2) - F_{11}(a_1) = 0$. Then $E(Y_1^{\text{hard}} | c, Z = 1) = 1 - \pi_{01}(c)$, which is the probability that neuron 2 does not spike, which therefore equals $\pi_{10}(c) + \pi_{11}(c) = \sum_{x:x_1=1} \pi_x(c)$, as required,

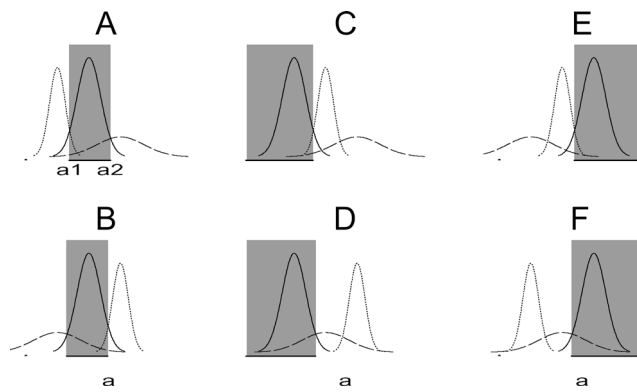


Fig. 4. Various configurations for $\pi_x f_x(a)$, $x \in \mathcal{X}$ [dotted is for neuron 1 ($x = 10$), solid for neuron 2 ($x = 01$), and large dashed for joint spikes ($x = 11$)]. The shaded areas indicate the set \mathcal{A} of values of a such that $\pi_{01} f_{01}(a) > \pi_x f_x(a)$, $x \neq 01$.

since $\sum_{x \in \mathcal{X}} \pi_x(c) = 1$. The other 5 configurations in Fig. 4 work out similarly. Therefore the common procedure of regressing Y_i^{hard} on c_t yields inconsistent firing rate estimates, unless the f_x are disjoint.

- Salganicoff M, Sarna M, Sax L, Gerstein GL (1988) Unsupervised waveform classification for multi-neuron recordings: A real-time, software-based system I Algorithms and implementation. *J Neurosci Methods* 25:181–7.
- Ohberg F, Johansson H, Bergenheim M, Pedersen J, Djupsjobacka M (1996) A neural network approach to real-time spike discrimination during simultaneous recording from several multi-unit nerve filaments. *J Neurosci Methods* 64:181–7.
- Lewicki MS (1994) Bayesian modeling and classification of neural signals. *Neural Comput* 6:1005–1030.
- Lewicki MS (1998) A review of methods for spike sorting: The detection and classification of neural action potential. *Network: Comput Neural Syst* 9:R53–R78.
- Quiroga RQ (2007) Spike sorting. *Scholarpedia* 2:3583.
- Glaser EM, Marks WB (1968) On-line separation of interleaved neuronal pulse sequences. *Data Acquisition Process Biol Med* 5:137–156.
- Letellier JC, Weber PP (2000) Spike sorting based on discrete wavelet transform coefficients. *J Neurosci Methods* 101:93–106.
- Wasserman L (2004) *All of Statistics* (Springer, New York), Chap 22.
- Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsaki G (2000) Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J Neurophysiol* 84:401–414.
- Shoham S, Fellows MR, Normann RA (2003) Robust, automatic spike sorting using mixtures of multivariate t-distributions. *J Neurosci Methods* 127:111–122.
- Ventura V (2009) Automatic spike sorting using tuning information. *Neural Comput*, in press.
- Kass RE, Ventura V, Brown EN (2005) Statistical issues in the analysis of neuronal data. *J Neurophysiol* 94:8–25.
- Okatan M, Wilson MA, Brown EN (2005) Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Comput* 17:1927–1961.
- Pouzat C, Delescluse M, Voit P, Diebolt J (2004) Improved spike-sorting by modeling firing statistics and burst-dependent spike amplitude attenuation: A Markov chain Monte Carlo approach. *J Neurophysiol* 91:2910–2928.

To prove that hard assignments Eq. 7 yield inconsistent estimates, all we do is rewrite the proof above with π_x replaced by $\pi_x(c)$. Then a_1 and a_2 also depend on c , which leaves the outcome of the proof unchanged.

Because all spike sorters assume, at least implicitly, that the data arise from the mixture distribution in Eq. 1, the same proof can be used to show that they yield inconsistent tuning-curve estimates, with the only change the domain of integration \mathcal{A} in Eq. 9, which must now describe formally the spike assignment rules of these spike sorters. For example, consider the spike sorter that consists of cutting clusters with the naked eye. A realistic way to do that is to assign a spike with waveform a to the neuron combination x whose mean waveform distribution is closest to a . If f_{01} and f_{10} are normal distributions with equal variance–covariance matrices, this means assigning to neuron 1 all waveforms in the set $\mathcal{A} = \{a : f_{10}(a) > f_{01}(a)\}$. Substituting this \mathcal{A} in Eq. 9 does not simplify it to the desired $\sum_{x \in \mathcal{X}} \pi_x(c)$, and thus yields the same conclusion that resulting tuning curves are inconsistent. One can proceed similarly for other spike sorters, although it might be difficult to formalize \mathcal{A} . However, despite lacking a general formal proof, we believe that it is highly unlikely for any waveform-based spike sorter to produce consistent tuning curves when waveform clusters overlap.

ACKNOWLEDGMENTS. I thank three reviewers for detailed and constructive comments, and the associate editor for support. This work was supported by National Institutes of Health Grants 2R01MH064537 and 1R01EB005847.