11-22-2019

Statistical Methods and Validation Procedures for Spike Sorting

Michael Thomas
Portland State University

Statistical Methods and Validation Procedures for Spike Sorting

by
Michael Thomas

An undergraduate honors thesis submitted in partial fufillment of the
requirements for the degree of
Bachelor of Science
in
University Honors
and
Mathematics

Thesis Adviser
Daniel Taylor-Rodriguez Ph.D.

Portland State University
2019

1 Abstract

Spike sorting is the process of converting a recording of the electrical activity generated by neurons firing in the brain, into a representation of the timings for each distinct neuron's firing. This representation of neuronal firings, when paired with data on the subject's perceptions or actions during the recording, can then be used to map patterns of brain activity to specific stimuli or behaviors. Here we will examine the evolution of mathematical techniques employed to tackle this problem, as well as examine a few of the still open questions.

2 Introduction & Background

The goal of generating a transcript of interactions between neurons in the brain is known as spike sorting. The process involves converting the electrophysiological data recorded from electrodes placed intracellularly, into a timeline representing the spike times for each neuron, known as a spike train. Such a transcript would be beneficial to diagnosing abnormal brain activity, research in brain-computer interfaces and prosthetics, and generally to researchers and practitioners in the neurological and cognitive sciences. Applications in prosthetics are of the most demanding targets, as a viable spike sorting process for use in these devices would need to be real time (referred to as 'online' spike sorting); a criteria we will keep in mind when evaluating the methodologies below.

The mathematical presumption underpinning all spike sorting models is the black box abstraction of the neuron, most notably that of Hodgkin and Huxley (Hodgkin et al., 1952). A neuron's cell membrane may depolarize, and then re-polarize to a baseline potential as a result of ions being let in and out through voltage gated channels, this whole phenomenon is known as an action potential. Hodgkin and Huxley modeled the propagation of action potentials by representing neural cell membranes as collections of discrete electrical components. While there is no analytical solution to the model, it proposes a rational basis for understanding the neuron as a black box obeying some consistent set of rules: 1) that there exist a limit cycle between the voltage and the potassium gate in a neuron, that is the cell's negative before positive wave is governed by the movement of physical ions in the cell, so we know there is some refractory period where new potentials may not be produced; and 2) there is a stochastically predictable 'characteristic spike' that each particular neuron will exhibit in it's action potential upon

firing. From this characterization of the neuron, the simplistic model for spike sorting would expect a single neuron's firing pattern to behave as a Poisson process–a method that charaterizes the structure underlying the times at which events occur.

The recordings used are generated by placing electrodes in the brain and recording the voltage generated by neurons during action potentials– the depolarization of an axon causing a voltage fluctuation. This action potential is the 'spike', and the objective is to classify or sort each spike in order to identify which neuron generated it. Recalling the second principle of our black box neuron, this is possible as an individual neuron, under *most* circumstances, produces a characteristic waveform when it fires. That is, the shape of the wave when plotting amplitude against time, for the same neuron, is the same (Rey et al., 2015). This assumption does break down during a 'burst', or a period of high excitement, when a neuron may fire rapidly in succession, and the resultant waveform recorded may be unrecognizable when compared to the non-bursting wave produced by the same cell.

Detecting these waveforms in the noisy extracellular recordings generally means first filtering out the activity of nearby cells, or local field potential (LFP), from the signal. This is typically done by using a bandpass filter between 300Hz and 5-8kHz (ibid.). Once filtered, the most straightforward detection algorithm simply detects spike peaks as threshold crossings above some threshold voltage estimated to be above background noise.

Once spikes are detected, the aim is to use the characteristics of the spikes to sort them into classes according to which neurons produced them. Since we believe that there is a distinct, prototypical waveform produced by each neuron during its action potential, if we consider the properties of the wave (i.e. the maximum amplitude, the inter-spike spacing, the length, etc), by analyzing these features we should be able to perform the desired sorting. It is at this stage of feature extraction where modern algorithms start to show real differences.

Amplitude-based methods (Sarah Gibson et al., 2008) are fast and straightforward to implement, and are used to cluster spikes based on their peak amplitude. These methods are prone to many shortcomings, as it is not necessarily the case that two neurons' waveforms have distinct peak amplitudes. For this reason, amplitude based methods were quickly forsaken for methods reliant on dimensionality reduction as a first step. In these techniques Principal Component Analysis (PCA) or Wavelets are used to decompose either the filtered signal, or some features obtained from it, into a smaller number of dimensions, before clustering (ibid.). PCA, among the most frequently used methods for dimensionality reduction, generates an orthogonal basis of

principal components, ranked by their variance. That is, if $x(t)$ denotes the spikes amplitude in time, and $c_i(t)$ is the principal component weight vector for the i-th with $i \in [1, T]$, summing over all time samples $t \in T$ yeilds the spike "score", or i-th principal component:

$$s_i = \sum_t c_i(t)x(t)$$

Under this process only the first 3-5 components, i, are generally retained as they are thought to capture most of the variational information necessary to sort the spikes into their classes, more specifically 3-5 components generally capture 45-80% of the variation in the data (Adamos et al., 2008). Ordering the components by variance, however, does not necessarily imply the most discriminating components will be selected. Since the high amplitudes are the first components, lower frequency components may not be well represented by this analysis (Rey et al., 2015).

Instead of PCA, Quiroga et al., proposed using Wavelets for feature extraction, as the wavelet transform produces a time-frequency representation of the signal, and then sorting using a process referred to as superparamagnetic clustering. Furthermore, rather than relying on variance, they select their wavelet coefficients using a Kolmogorov-Smirnov test against a normal distribution to select their components under the assumption that features with a multimodal distribution will behave as more informative classifiers– section 3 explains this approach in further detail. While Quiroga's superparamagnetic clustering process produce good results, and is argued to be more robust to non-Gaussian LFP (caused by electrode or neuron drift) than methods using PCA for dimensionality reduction, their work was performed on a single electrode and the wavelet transform becomes increasingly expensive to calculate online for multi-electrode arrays (MEAs).

Other methods of producing features exist and have included peak amplitude, inter-spike intervals, independent component analysis, the nonlinear-energy operator, and others. (Sarah Gibson et al., 2008; Rey et al., 2015)

The remainder of work here will focus on the clustering stage, and validation approaches for these algorithms. Beginning in the early 2000s, we will examine how the state of the art came to be where it is today as a result of external pressures to sort larger data faster, and more accurately. We will look at where certain approaches have fallen short, and where they have been iterated on for improvements. And finally, we examine what strategies are assailable for accessing the accuracy of models, when there is little to no ground truth data available for their validation.

3 Clustering

Regardless of how the data has been pre-processed, clustering the detected spikes by their shape is the ultimate goal of the sorting process. There has been immense growth in this area, which can largely be characterized into three generations of research.

3.1 Early Methods

Early methods of clustering were mostly manual. The feature space was presented graphically to a researcher who would then divide it, but this is error-prone, infeasible on dense electrode arrays, and obviously impossible to scale to an online method. Density based methods where the earliest attempts at an automatic process.

3.1.1 Density based methods

Density based clustering methods arise early and have remained prominent– in a review from pre-2000 Lewicki (Lewicki, 1998) names 3 approaches employing k-means clustering. These methods minimize the within-cluster sum of squares between the features, effectively partitioning the space into Voronoi cells. That is, first a random collection of data points are taken as the cell centers, and the other points are sorted based on their distance to these centers. At each step the centers of each cell are recalculated based on their members until no further update is possible. Such a partition implies sorting each spike into a set (the voronoi cell) with a center such that all points in the set are closer to their set's center than any other's. So if $\vec{x}$ is a feature vector for some spike, k-means iterates over that data in 2 steps. First, with x as some feature, and $K = k_1, \ldots, k_n$ a set of cluster centers, x is assigned to a cluster K_i, if its mean m_i has the smallest Euclidean distance to x. Second, once all points have been assigned to a cluster, then the updated cluster centers are recalculated with:

$$m_i = \frac{1}{|K_i|} \sum_{x_j \in K_i} x_j$$

The process repeats until no data points are reassigned to a new cluster in the first step.

These methods are generally quite fast, however they require a priori knowledge of the number of classes, and have been shown to be insufficient as they can produce false neurons, or classes, when the noise is non-Gaussian.

This can lead to cells whose 'centers', under the desire that they represent the most prototypical spike for the class, should technically lie outside the cell. Intuitively, k-means expects clusters to have a spherical shape since the distance metric used to calculate cluster centers is Euclidean. If there is electrode drift, non-Gaussian randomness in the local field potential, or non-white noise in the data, these methods will struggle (S. Gibson et al., 2012). Furthermore, any method reliant on a k-means step will necessarily require the whole dataset and cannot be performed online.

3.1.2 Bayesian methods

The first Bayesian approaches to spike sorting also appear in the 1970s, as cited in Rey et al., 2015. They start by assuming cluster are Gaussian and that they only vary from one another by additive and background noise, both themselves Gaussian (ibid.). From these assumptions the likelihood of some spike event $\vec{x} = [x_{t1}, \ldots, x_{tn}]$ given a particular neuron, or class c_k is:

$$p(\vec{x}|c_k, \mu_k, \Sigma_k)$$

with μ_k, and Σ_k representing the mean and covariance for the class c_k respectively. Then, if $\theta_{1:K}$ is the set of all (μ_k, Σ_k) class parameters, using Bayes Rule the probability of the data points in $\vec{x}$ belonging to any class is given by (Takekawa et al., 2010):

$$p(c_k|x, \theta_{1:K}) = \frac{p(\vec{x}|c_u, \theta_u)p(c_u)}{\sum_{j=1}^{K} p(x|c_j, \theta_j)p(c_j)}$$

.

The two initial challenge with this method is that the value of K, the number of clusters or neurons present in the recording is not known a priori. Schwarz attempted solving this issue by introducing a penalty function on the creation of new classes (ibid.), and software packages like AutoClass, and others, implemented this sort of approach in the late 1990s.

3.2 2000-2015

Motivated by a need for faster and more accurate methods, as well has large multielectrode arrays (MEAs) with 16-256 electrodes coming into more common use, a new generation of algorithms started gaining prominence in the early 2000s. As a point of reference, if a 256 electrode array generates 30,000 values for amplitude per second, a 1 minute recording would comprise

14MB. Today arrays exist with over 4,000 electrodes, and a typical recording is 30-90 minutes.

3.2.1 Density Based Methods

Multiple novel approaches were considered to account for the problems in early density based methods. Super-paramagnetic clustering (SPC) was introduced by (Quiroga et al., 2004), which enhanced the nearest neighbors approach by computing the 'interaction strength', or the tendency of features to change states together when recalculating cluster scores. For a spike i, represented as the feature vector $\mathbf{x}_i$, it's interaction strength to a nearest neighbor $\mathbf{x}_j$ is given by:

$$J_{ij} = \begin{cases} \frac{1}{N}\exp\Big(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2a^2}\Big) & \text{if } \mathbf{x}_i \text{ is a nearest neighbor of } \mathbf{x}_j \\ 0 & \text{o.w.} \end{cases} \tag{1}$$

Here $\mathbf{x}_i$ is a feature vector representing the ith spike in m -dimensions, a is the average nearest-neighbor distance, and N is the number of nearest neighbors.

In a second step, each point in $\vec{x}_i$ is give a random 'state' between 1 and some predefined q (chosen by Quiroga to be 1-20 by experiment), Monte Carlo simulations are then run over a range of temperatures T, where the state of a randomly chosen $\vec{x}_i$ is changed. In this configuration the probability that the neighbors of $\vec{x}_i$ will also change states is given by:

$$p_{ij} = 1 - exp(-\frac{J_{ij}}{T}\delta_{s_i,s_j})$$

Points that change values in an iteration are called the "frontier" and at each cycle the equation above is applied to the frontier to update it and this repeats as in standard k-means until no points change. To form a representative statistic, the researchers then repeated this process from multiple different points. Spikes are then assigned to clusters when their point to point interaction, measured by $\langle \delta_{s_i,s_j} \geq \theta \rangle$, is greater than some predefined θ.

The key idea here is to determine the optimum number of clusters by progressively adjusting the the temperature parameter, thus controlling the scale of the underlying probability distribution representing the likelihood of points to change state together. This class of algorithm is known as simulated annealing, and is modeled after the process of heating and cooling

metal to control the energy in the material and reduce defects. The 'energy', or interaction strength here, is scaled by T such that at higher values the algorithm will have a higher probability of moving uphill, and likewise, as the temperature is decreased over the iterations of the Monte Carlo simulation, the probability of uphill moves decreases. This allows Quiroga to avoid problems other optimization algorithms encounter where they get stuck in local maxima. It also allows them to automatically determine the number of needed classes as at some continuous set of temperatures, the algorithm stabilizes and produces a consistent number of classes. This approach removed the need for supervision, or a prior knowledge of the number of neurons present in the recording. It also performed quite well against data with non-Gaussian noise (e.g. where a cluster's centroid falls outside it's boundary).

3.2.2 Bayesian

While early Bayesian methods solved their supervision problem by setting a penalty on the creation of classes, Wood et al., 2008 proposed one of the most robust models for single channel data by modeling the problem non-parametrically. Wood suggests a recording $R = [\vec{t_1}, \ldots, \vec{t_N}]$ with N waveforms detected on a single channel, and where $\vec{t_i} = [t_i^1, \ldots, t_i^n]^T$ are $n = 40$ voltage samples for the ith waveform. PCA is applied, yielding the representation:

$$\vec{t_i} \approx \mu + \sum_{d=1}^{D} y_i^d \vec{u}_d$$

where μ is the mean waveform in $\mathbb{R}^n$, y_i^d the linear coefficients, and $\vec{u}_d$ the dth PCA basis. Wood then deals with the low dimensional representation $\mathcal{Y} = [\vec{y}_1, \ldots, \vec{y}_N]$ of the data, with $\vec{y}_i = [y_i^1, \ldots, y_i^D]$. Representing the spike as it's mean waveform and the statistics about how it varies between spikes does better to capture the effects of variances like electrode drift as averaging the results of this model over successive runs provides evidence in support of a class partition. (ibid.).

Starting from a simplified version of the model, Wood defines

$$P(\vec{y}_i) = \sum_{k=1}^{K} P(c_i = k) P(\vec{y}_i | \theta_k)$$

as the discrete probability of a waveform belonging to a particular class, i.e. that the wave was generated by particular neuron $k \in K$, $P(c_i = k)$,

times a multivariate normal $P(\vec{y}_i|\theta_k)$. A fixed K version of the joint probability for this model, that is, assuming they number of neurons K in the recording is known, is given by:

$$P(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha, \mathcal{H} = \Big(\prod^{K} P(\theta_j, \mathcal{H})\Big)\Big(\prod^{N} P(\vec{y}_i|c_i, \theta_{c_i})(P(c_i|\vec{\pi})\Big)P(\vec{\pi}|\alpha)P(\alpha)$$

With $\mathcal{C}$ as the class indicators for each neuron in class K, $\Theta = \vec{\mu}_k, \mathbf{\Sigma}_k$ as the class parameters for each k, $\vec{\pi}$ as the prior probabilities that a a waveform was generated by a neuron k, and α and $\mathcal{H}$ are hyperparameters.

For the hyperparameters, Wood and Black choose $\vec{\pi}|\alpha \sim Dirichlet(\cdot|\frac{\alpha}{K}, \dots \frac{\alpha}{K})$ and $\Theta \sim$ Inverse Wishart $\cdot$ Gaussian.

While these updates do more effectively solve the supervision problem, this method doesn't account for multiple channels, and there is a stopping problem as the number of classes K is not known a priori. To solve for this Wood and Black extend the model above to an infinite Gaussian mixture model by integrating $\vec{\pi}$ out of the mixture model as $K \to \infty$. The limiting expression for the total probability of an arbitrary partition is given as:

$$P(\mathcal{C}|\alpha) = \alpha^{K_+}\Big(\prod_{k=1}^{K_+}(m_k - 1)!\Big)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

Where K_+ is the number of classes with at least one observation, and m_k the number of observations in class k. This together with the likelihood obtainable from the joint given above, provides the necessary information to formulate a Gibbs sampler.

The intuition is captured by:

$$P(c_i = k|\mathcal{C}_i) = \begin{cases} \frac{m_k}{i-1+\alpha} & k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & k > K_+ \end{cases} \tag{2}$$

Which explains that if the number of neurons in a given class m_k is large, then the likelihood for a new neuron to be added to that class is also high.

3.2.3 Templates

Early template-based approaches relied on the similarity of spikes to those in a dictionary, and were prone to false positives when collided spikes produced large amplitudes or abnormal shapes (Wouters et al., 2018). More

recent efforts like Kilosort (Pachitariu et al., 2016), SpikeDetekt (Kadir et al., 2013), and Spyking circus (Yger et al., 2018) have used some combination of methods to attempt to surmount these problems.

3.3 2016-On

MEAs are challenging to tackle for the early algorithms, and even more so recently, since electrode arrays with $20\mu m$ pitch and 1000s of electrodes are now coming into use. These further exacerbated the problems with the curse of dimensionality that the density based methods have always struggled with. i.e. that the methods that work on low dimensional spaces tend to fail in high dimensions as the data points become sparse as the space grows.

The most modern incarnations of algorithms take hybrid approaches to this. They apply ideas from all past concepts like density based, Bayesian, and template approaches, as well as new methods like neural nets and dictionary learning. Most of these most recent algorithms are real-time, in the sense that they take about as long as the recording takes to process rely on having the entire dataset present at processing time, thereby rendering them incapable of being used online.

Yger's Spyking Circus algorithm (ibid.) from 2018 claims to be able to process data generated on up to 4225 electrodes, and is what we will test with ground truth data in the next section. Yger's updates to the 2000's on methods for spike sorting, namely their approach combining multiple of techniques explored above in their algorithm, is prototypical of this upcoming wave of research. Their key addition is a "Whitening" step where the dimensionality of the problem is reduced by taking the waveform as heard only on the strongest channel of the MEA.

The Spyking Circus algorithm comprises three main steps. First the spikes are detected as threshold crossings, clusters are built, and finally templates are learned from those clusters. Since the algorithm is designed with MEAs in mind, for the initial 'pre-clustering' step, Yger takes only a sample of spikes, and only the waveform on the electrode with the strongest amplitude. This initial set of peaks are clustered according to the procedure defined in Rodriguez and Laio (Rodriguez et al., 2014). Once similar clusters are merged, templates are estimated and represented by two components $\mathbf{w}_m$ and $\mathbf{v}_m$, where $\mathbf{w}_m$ is the point-wise median of waveforms in cluster m:

$$\mathbf{w}_m = \text{med}_l \mathbf{s}(t_l^m)$$

And the second component $\mathbf{v}_m$ is calculated by first projecting the waveforms $\mathbf{s}(t)$ into a space orthogonal to $\mathbf{w}_m$, i.e.

$$\mathbf{q}_l = \mathbf{s}(t_l^m) - \frac{\mathbf{s}(t_l^m c) \cdot \mathbf{w}_m}{||\mathbf{w}_m||}\mathbf{w}_m;\ \forall l$$

and then taking the first principal component of $\mathbf{q}_l$.

From this, Yger's fundamental claim is that a signal is decomposable into a spatio-temporal sum of kernels, called "templates", such that:

$$\mathbf{s}(t) = \sum_{ij} a_{ij}\mathbf{w}_j(t - t_i) + b_{ij}\mathbf{v}_j(t - t_i) + e(t)$$

where $s(t)$ is the signal voltage at time t, $\mathbf{w}$ and $\mathbf{v}$ are the two components of the template and e is noise from the LFP. Template matching then is the computation of (a_{ij}, b_{ij}) for each time step t such that they reconstruct the original signal $\mathbf{s}(t)$.

Yger's methodology enables the online processing of massive amounts of data, as once the templates are calculated (from a small subset of data) the only information necessary for classification are the estimated templates $(\mathbf{w}, \mathbf{v})$ and the sparse coefficients (a_{ij}, b_{ij}). Their methods however do not necessarily handle electrode drift and spike collisions well, as they share problems with the rest of the PCA based methods.

4 Validation

4.1 General Practice

A consistent method for validation of spike sorting algorithms has not been established. The majority of authors compare results on an algorithm to algorithm basis. In effect, with little ground truth data for use, studies compare the success of the clustering stage of their new algorithm to a 'hybrid ground truth' (C. Rossant et al., 2015) comprised of the results of a known good algorithm, and human curation. In (Pachitariu et al., 2016), pseudo-ground truth data was generated via the software KlustaKwik (Kadir et al., 2013), and curated with a human expert's review. In a best case scenario the authors claim to identify 69% of matches, compared to KlustaKwik's 60%.

Simulated data is another common practice. Here spikes are generally copied from known templates, or online template banks, and overlayed with noise that attempts to mimic the LFP signal, again usually by simulation of data or background noise. This is by far the most common methodology we encountered, and can be found in: Quiroga et al., 2004, Chung et al., 2017, Lee et al., 2017, and Yger et al., 2018.

The ideal validation would be done on 'Ground truth data', and such a recording would comprise: 1) a known number of neurons, 2) the identity, ideally by position, of each of those neurons known and recognizable by some prototypical wave, and 3) the times at which least one of those neurons fired. Unfortunately such a dataset does not currently exist, and even close approximations still rare.

4.2 Attempts at Ground Truth Validation

The study by Yger, responsible for the Spyking circus library, attempted a ground truth validation of their data using the hybrid data approach. First, a known "good" data set was processed using the Kilosort package for Matlab, and then curated by an expert. Against their testing, Yger found a 10% error when comparing Spyking Circus to an optimal classifier (which accounts for amplitude dropoff for neurons farther than 60µm.

In 2016 Kampff-lab released several datasets where 32 and 128 channel MEA extracellular recordings paired with time aligned juxtacellular recordings (Neto et al., 2016). To create these, MEAs were surgically implanted into the motor, sensory, or parietal cortex of anesthetized rodents (ibid.). Adjacently, via a microscope, a second electrode was aligned until in contact with a cell membrane. This was detected by resistance on the electrode, and spikes, at which point suction was applied to affix the probe to the site. The juxtacellular recording then gives a time aligned data set to the MEA recording, where the juxtacellular probe's position is known to be within 30-150µm of a specific electrode on the polytrode array.

Kampff details their internal comparison, using the algorithm SpikeDetekt (ibid.), and found widely varying results. In one recording SpikeDetekt found 386 spikes compared to 348 on the juxtacellular channel; and in another recording 35 were detected to the 150 juxtacellular. In general, Kampff also questions the reliability of the MEAs at distances. In their experiments a 50µV voltage differential was only perceptible for some probes inside of approximately 90µm diameter ball around the recording site.

4.3 Replicating that effort

Given the limited amount of ground truth data available, and thus the limited number of studies using such data to verify modern spike sorting algorithms, we performed a similar analysis as to the one explained in Kampff (ibid.). We apply similar techniques to the same data set Kampff used to validate KlustaKwik, but against a different sorting algorithm, allowing us to

compare Kampff's findings across two spike sorting techniques. With these results we are able to substantiate Kampff's concerns about the true sensitivity of MEAs to voltage differentials, and give explicit examples of where the challenges explained in the previous sections appear when confronted with real world data.

the First the data was sorted via the Spyking Circus algorithm, a competitor to KlustaKwik–used at this step by Kampff. Then the juxtacellular data was sorted and compared.

Since the juxtacellular channel has been confirmed to be generating spikes at the recording, and since we can isolate the juxtacellular channel from the MEA data, and we know it represents the electrode suctioned to the cell membrane nearest to one specific channel on the extracellular array, we will treat it as a ground truth of spike times. From that we are able to compare the spikes detected on those two channels as an evaluation of the accuracy of a 3rd party algorithm for sorting. In an effort to validate the algorithm, we would ideally find a coincidence of all occurrences of spikes on the juxtacellular channel and their corresponding event on the closest extracellular channel (by proximity).

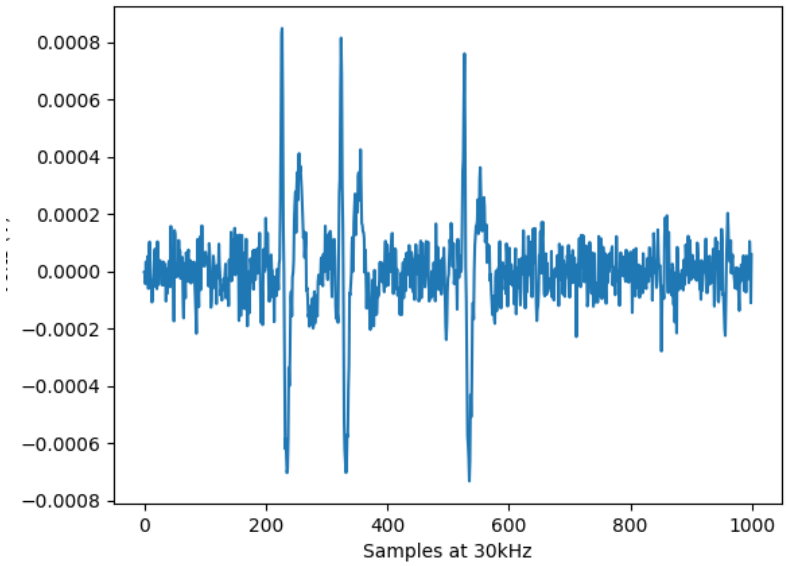

Figure 1: 1000 samples (30ms) of the filtered extracellular signal, containing a bust of spikes.

To test this we used the recording 2014_11_25_Pair3.0 from the Kampff dataset, and first read the data for the juxtacellular channel into a python numpy array, as per the procedure laid out by Kampff. The signal was band-pass filtered using a butterworth filter between 500Hz-15kHz, as suggested in (Neto et al., 2016), and then converted from its digitized form to a timeseries in volts. From this timeseries, spikes were detected as threshold crossings above a noise floor (LFP).

The polytrode, or extracellular, data was processed via Spyking Circus's python library, which utilizes a hybrid clustered-template match approach to sorting, as described above. After sorting the software outputs a HDF5 file which was used to extract the spike times as see by Spiking Circus. These data were plotted as time series using phy (Rossant, 2019), a small visualization library for manual sorting of electrophysiological data (Fig. 2).

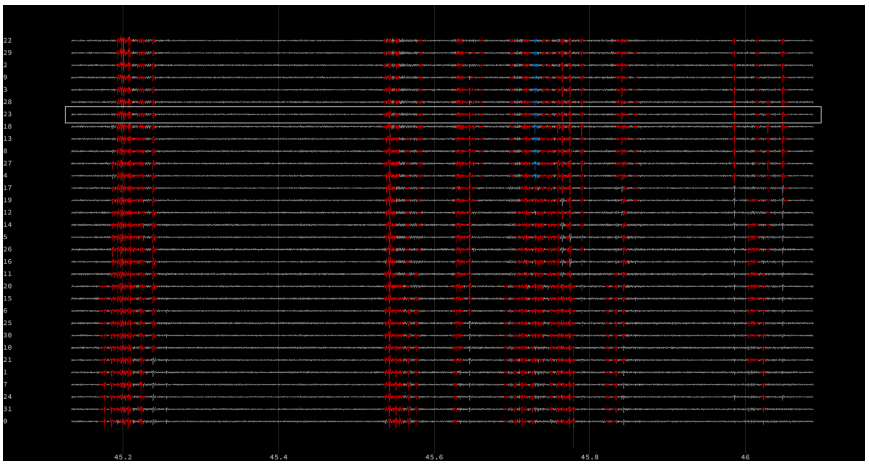

Figure 2: An approximately one second sample of the full 32-channel MEA data. Spikes detected by Spyking Circus are highlighted. The red colored segments demark any segment of signal Spyking circus classifies as a spike, whereas the blue indicates a spike found at the juxtacellular pipette. The channel nearest the juxtacellular probe, number 23, is boxed in a white boundary.

The timings of the spikes detected on the MEA electrode nearest the juxtacellular pipette were extracted from Spyking Circus, and ignoring the cluster, there were a total of 1,576 spikes found on that channel. Based on our threshold crossings analysis, we detected 395 spikes on the juxtacellular channel, compared to the 348 Kampff noted as having detected. Irrespective of our slight over counting of spikes on the channel, there still seems to

be a bias in Spyking Circus towards the over detection of spikes. This is graphically depicted in Fig 3 which shows a 1 minute section of the raw data from the recording, with spikes detected both on the juxtacellular channel and those recognized by Spyking Circus overlayed.

There are several possibilities to evaluate in understanding this discrepancy of spike counts. Before assuming fault on the part of the sorting software, the potential for overlapping spikes and poor readings from the physical MEA device should be considered as potential sources of inaccuracy. Multiple authors, including Kampff and Yger have also noted concerns about the salient detection distance of MEAs. Unfortunately these theories would, however, primarily account for an under count of spikes on the channel, and not an over count, whereas we have both.

A more robust examination of this current generation of models would compare the results of multiple modern approaches when tested against multiple of these ground truth datasets. Henning's 2018 review provides an overview of those systems, and future work would consider a more fully automated process by which this analysis could be repeated over large numbers of algorithms and datasets. Furthermore, controlling for electrode distances in the final analysis might also shed more light on the problem of how sensitive the signal reported by the MEAs can be believed to be.

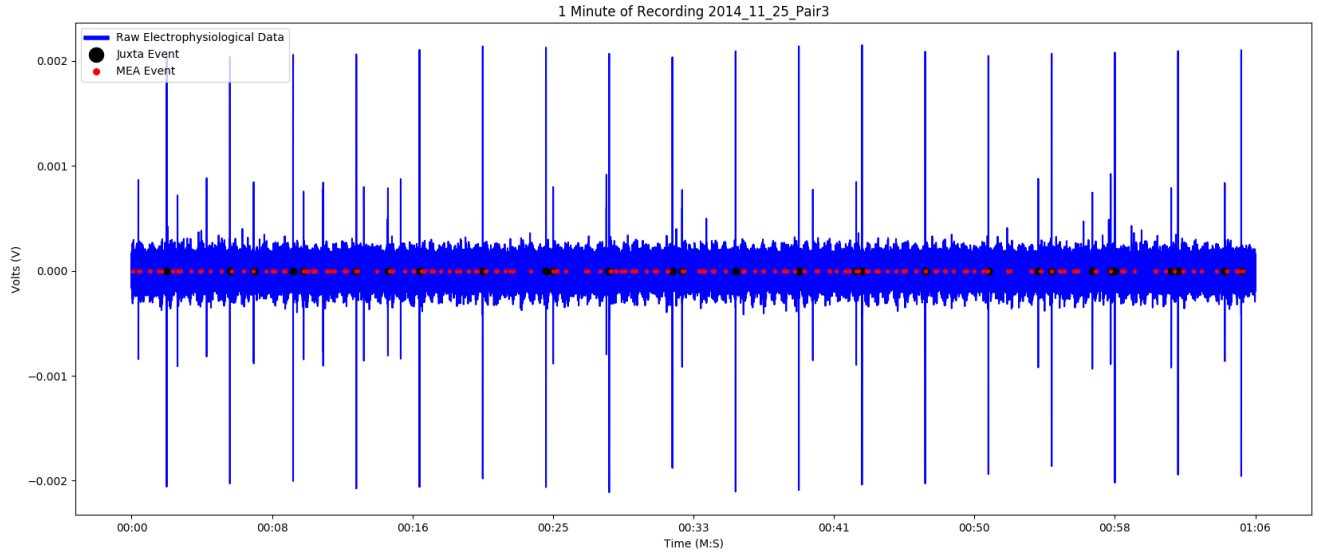

Figure 3: A 1 minute sample of the raw electrophysiological recording analyzed here. Spikes detected on the juxtacellular channel are overlayed onto the data with black circles (48), and the spikes which registered on the nearest MEA channel, as detected by Spyking Circus, are shown in red circles (198).

5 Conclusion

The problem of spike sorting is the trade-off between accuracy and computational complexity. As the current generation of MEAs grow in the number of simultaneous recording sites, the data sizes continue to grow too. This makes it physically difficult to move and compute on the data, but also further increase the dimensionality of the problem. The most recent generation of solution seem well equipped to tackle these weaknesses, being particularly tuned to account for massive data, online processing, and corrections for electrode drift. Nonetheless the results from Yger, Kampff, etc. are still relatively unevaluated against ground truth data, and the few evaluations which have been completed suggest there still multiple open problems in spike sorting.

Discrepancies between the juxtacellular and MEA recordings seem to suggest one of three basic problems: 1) a signal spike being detected as the results of two waveforms colliding at the electrode (Lee et al., 2017), 2) electrodes (or cells) drifting and distorting the waveform received at the recording site (Yger et al., 2018), and 3) lack of understanding of how the sampling fidelity of MEAs decays as a function of distance and neural tissue type (Neto et al., 2016). In order to get a better understanding of the true number of neurons likely detectable in a given recording, and continue to improve on the gaps between what is detected via loose path juxtacellular recordings and large MEAs, each of these issues should be addressed.

Inaccurate spike detection leads to problems in all downstream process relying on spike sorting. With more high quality ground truth data coming available, effort should be made to develop new, robust evaluation procedures for the comparison of the accuracy between algorithms. These procedures could also be leveraged to aid in the development of newer AI assisted sorting models as some authors have suggested moving towards (Clark, 2013), as ground truth datasets and training will become even more imperative.

References

Adamos, Dimitrios A, Efstratios K Kosmidis, and George Theophilidis (2008). "Performance evaluation of PCA-based spike sorting algorithms". eng. In: *Computer Methods and Programs in Biomedicine* 91.3, pp. 232–244. ISSN: 0169-2607 (cit. on p. 4).

Chung, Jason E. et al. (Sept. 2017). "A Fully Automated Approach to Spike Sorting". en. In: *Neuron* 95.6, 1381–1394.e6. ISSN: 08966273. DOI: 10.1016/j.neuron.2017.08.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627317307456 (visited on 10/01/2018) (cit. on p. 11).

Clark, Andy (June 2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". en. In: *Behavioral and Brain Sciences* 36.03, pp. 181–204. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X12000477. URL: http://www.journals.cambridge.org/abstract_S0140525X12000477 (visited on 09/29/2018) (cit. on p. 17).

Gibson, S., J. W. Judy, and D. Markovic (2012). "Spike Sorting: The First Step in Decoding the Brain: The first step in decoding the brain". eng. In: *Signal Processing Magazine, IEEE* 29.1, pp. 124–143. ISSN: 1053-5888 (cit. on p. 6).

Gibson, Sarah, Jack W. Judy, and Dejan Markovic (Aug. 2008). "Comparison of spike-sorting algorithms for future hardware implementation". In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. DOI: 10.1109/iembs.2008.4650340. URL: http://dx.doi.org/10.1109/iembs.2008.4650340 (cit. on pp. 3, 4).

Hodgkin, A. L. and A. F. Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve". In: *The Journal of Physiology* 117.4, pp. 500–544. DOI: 10.1113/jphysiol.1952.sp004764. eprint: https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1952.sp004764. URL: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1952.sp004764 (cit. on p. 2).

Kadir, Shabnam, Dan Goodman, and Kenneth Harris (Sept. 2013). "High-Dimensional Cluster Analysis with the Masked EM Algorithm". In: *Neural computation* 26. DOI: 10.1162/NECO_a_00661 (cit. on pp. 10, 11).

Lee, JinHyung et al. (June 2017). "YASS: Yet Another Spike Sorter". In: DOI: 10.1101/151928. URL: http://dx.doi.org/10.1101/151928 (cit. on pp. 11, 17).

Lewicki, Michael S (1998). "A review of methods for spike sorting: the detection and classification of neural action potentials". eng. In: *Network:*

Computation in Neural Systems 9.4, R53–R78. ISSN: 0954-898X. URL: http://www.tandfonline.com/doi/abs/10.1088/0954-898X_9_4_001 (cit. on p. 5).

Neto, Joana P. et al. (2016). "Validating silicon polytrodes with paired juxtacellular recordings: method and dataset". In: *Journal of Neurophysiology* 116.2. PMID: 27306671, pp. 892–903. DOI: 10.1152/jn.00103.2016. eprint: https://doi.org/10.1152/jn.00103.2016. URL: https://doi.org/10.1152/jn.00103.2016 (cit. on pp. 12, 14, 17).

Pachitariu, Marius et al. (June 2016). "Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels". In: DOI: 10.1101/061481. URL: http://dx.doi.org/10.1101/061481 (cit. on pp. 10, 11).

Quiroga, R. Quian, Z. Nadasdy, and Y. Ben-Shaul (Aug. 2004). "Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering". en. In: *Neural Computation* 16.8, pp. 1661–1687. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976604774201631. URL: http://www.mitpressjournals.org/doi/10.1162/089976604774201631 (visited on 09/29/2018) (cit. on pp. 7, 11).

Rey, Hernan Gonzalo, Carlos Pedreira, and Rodrigo Quian Quiroga (Oct. 2015). "Past, present and future of spike sorting techniques". en. In: *Brain Research Bulletin* 119, pp. 106–117. ISSN: 03619230. DOI: 10.1016/j.brainresbull.2015.04.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0361923015000684 (visited on 09/29/2018) (cit. on pp. 3, 4, 6).

Rodriguez, Alex and Alessandro Laio (2014). "Machine learning. Clustering by fast search and find of density peaks". eng. In: *Science (New York, N.Y.)* 344.6191, p. 1492. ISSN: 00368075 (cit. on p. 10).

Rossant (2019). *phy: interactive visualization and manual spike sorting of large-scale ephys data*. Version 2.0-dev. URL: https://github.com/cortex-lab/phy (cit. on p. 14).

Rossant, Cyrille et al. (Feb. 2015). "Spike sorting for large, dense electrode arrays". In: *bioRxiv*. DOI: 10.1101/015198. URL: http://dx.doi.org/10.1101/015198 (cit. on p. 11).

Takekawa, Takashi, Yoshikazu Isomura, and Tomoki Fukai (2010). "Accurate spike sorting for multi-unit recordings". In: *European Journal of Neuroscience* 31.2, pp. 263–272. DOI: 10.1111/j.1460-9568.2009.07068.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2009.07068.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2009.07068.x (cit. on p. 6).

Wood, Frank and Michael J. Black (Aug. 2008). "A nonparametric Bayesian alternative to spike sorting". In: *Journal of Neuroscience Methods* 173.1, pp. 1–12. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2008.04.030. URL: http://dx.doi.org/10.1016/j.jneumeth.2008.04.030 (cit. on p. 8).

Wouters, Jasper, Fabian Kloosterman, and Alexander Bertrand (2018). "Towards online spike sorting for high-density neural probes using discriminative template matching with suppression of interfering spikes". In: *Journal of Neural Engineering* 15.5, p. 056005. URL: http://stacks.iop.org/1741-2552/15/i=5/a=056005 (cit. on p. 9).

Yger, Pierre et al. (Mar. 2018). "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo". In: *eLife* 7. ISSN: 2050-084X. DOI: 10.7554/elife.34518. URL: http://dx.doi.org/10.7554/elife.34518 (cit. on pp. 10, 11, 17).