Pavel Larionov*, Tom Juergens and Thomas Schanze

Correlation-based spike sorting of multivariate data

Abstract: Automated classification of waveforms is an important method of data processing used in various fields of science, such as neuroscience, biomedical engineering, etc. This work shows the possibility of sorting special waveforms i.e. spikes recorded with multichannel electrode arrays by using principles of correlation and data-driven reference. A new method to estimate the number of k-means clusters by using a Monte Carlo method is introduced. To demonstrate the performance of the algorithm, generated signals were used, which are created to mimic multichannel recording of the extra-cellular neuronal signals.

Keywords: spike sorting, clustering, biosignals, multivariate signal processing, *k*-means, scree plot

https://doi.org/10.1515/cdbme-2019-XXXX

1 Introduction

Spike sorting is the process of distinguishing the characteristic action potential activity of one or more neurons present in an extracellular recording. Spike sorting can be performed in different ways, from simple but often inaccurate methods, e.g. spike sorting via thresholding, to more sophisticated ones, like principal component analysis or wavelet analysis [1][2].

It was shown, that it is possible to characterize spikes by correlating them with a data driven reference spike. The use of minimal and maximal correlation values is an effective and efficient dimension reduction approach, since only two parameters are extracted and used for subsequent cluster analysis in order to discriminate neuronal activities [3].

When recording extracellular signals with multiple electrode arrays, each electrode contact senses a signal which is related to its distance to the active neuron. This leads to a neuron specific potential distribution across the electrodes, i.e. a fingerprint of a neuron's activity. This effect is known as the stereotrode effect [4].

This work is a proposal of the possibilities of enhancing the spike sorting capabilities of our previous developed single channel algorithm by introducing multichannel algorithm.

Other goals are to show that new approach leads to a more robust clustering of the neuron specific minimal and maximal correlation values and to introduce a Monte Carlo method to estimate the number of classes required for *k*-means clustering in order to recognize the difference in the spike activities present in the data.

2 Methods

2.1 Multivariate dataset feeding

In order to develop and test our algorithm, a spike-generator was written, which creates a simulation of spike recording with *n* electrodes, given that every electrode writes raw data separately [5]. The simulation also implements quasi-randomized amplitude differences for each electrode, as it would be the case in real world scenario, i.e. mimicking the stereotrode effect. In such a manner a tetrode recording is simulated, i.e. four sources of signal, including different spike waveforms, enriched with Gaussian white noise (Figure 1).

2.2 Waveform detection & sampling

The peaks were detected by finding local minimums and maximums with predefined prominence and optional threshold in each of four signals separately [3]. It is necessary to eliminate duplicated peaks, as peaks may have prominent both negative and positive segments, i.e., side-lobes.

Now it is possible to cut samples, i.e. spike waveforms. That it is generally more profitable to make sample window smaller [6], but with multivariate approach it is not that important, as the spikes will be concatenated. However, it is still profitable to realign the samples around the so-called

***Corresponding author: Pavel Larionov:** IBMT, FB Life Science Engineering (LSE), Technische Hochschule Mittelhessen (THM), Wiesenstr. 14, 35390 Giessen, Germany, e-mail: pavel.larionov@lse.thm.de
Tom Juergens: IBMT, FB Life Science Engineering (LSE), Technische Hochschule Mittelhessen (THM), Giessen, Germany
Thomas Schanze: IBMT, FB Life Science Engineering (LSE), Technische Hochschule Mittelhessen (THM), Giessen, Germany

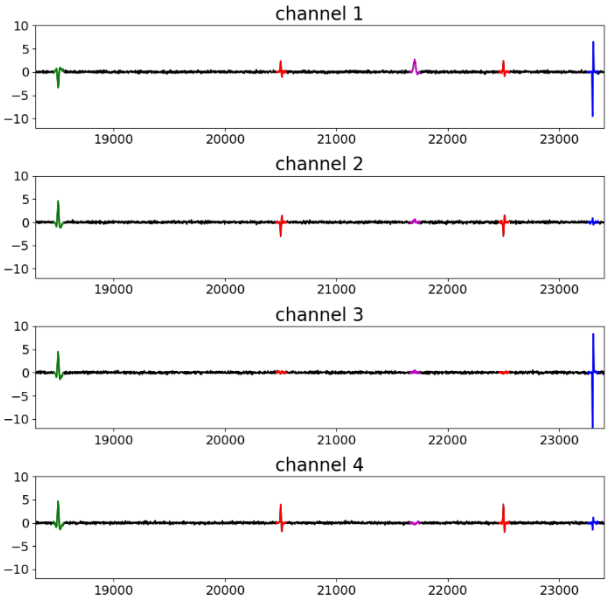

Figure 1: Multichannel data feeding: stereotrode effect can be observed here, as simultaneous spikes occur on all four electrodes, but with different amplitudes. Standard deviation of the noise here is 0.15.

centres of mass of data points, as it contributes to more prominent separation between correlation clouds.

2.3 Concatenation & reference spike

To collect all relevant information about a spike from our four signals and to be able to somehow compare them, the four spike waveforms are concatenated in virtual composite samples [5]. It is assumed, that travel time of a neuron's signal to all electrode contact of an electrode array is neglectable short and therefore we can suppose, that at every data point, where a spike is detected in one signal channel, there should be the same spike on all four electrodes, even if it is too small to detect or corrupted by any means, e.g. noise (homogenous electrical tissue properties). Duplicates' elimination was performed again here, in order to pick only one virtual composite sample per spike.

The obtained concatenated data are then used to build a data-driven reference spike. Reference spike lets the algorithm "know", what trend do given data represent and allows to differentiate between similar waveforms much better, than predefined reference spike or, for example, a straight line [3].

2.4 Cross correlation

Now it possible to calculate cross-correlations between the virtual composite samples and the composite reference. The lowest correlation value and the highest one for each sample were taken and projected them on a plane. This method is fast and allows to clearly see the so-called correlation clouds, which supposable represent the waveforms, that are contained in the signals [3].

The shape and separation of the clouds depend on several factors, such as waveform similarity, correct alignment, and sample window size. A strong discretization of the correlation clouds may occur due to inadequate signal sampling. This can be avoided by increasing the sampling rate of recording or fixed by interpolation of the data.

2.5 *K*-Means with estimation the number of classes

In order to automate the process of analysis it is necessary to apply a clustering algorithm on the correlation clouds [3].

First, *k*-means calculates the squared Euclidian distances between some random start points (centroids) and each observation of the clustering massive, initially assigning them to the nearest centroid. Then it calculates new centroids, based on previous assignment. The algorithm converges when the assignment is stable (no reassignment needed) [7].

One of the prerequisites of *k*-means is knowing the number of clusters *k*, which is, of cause, scarcely the case. We took the basic scree-elbow method of estimation the number of classes [8] and customized it to the point of full automatization, as workaround for this problem.

In order to get the scree plot, it is necessary to run *k*-means algorithm with incrementing argument for number of clusters,

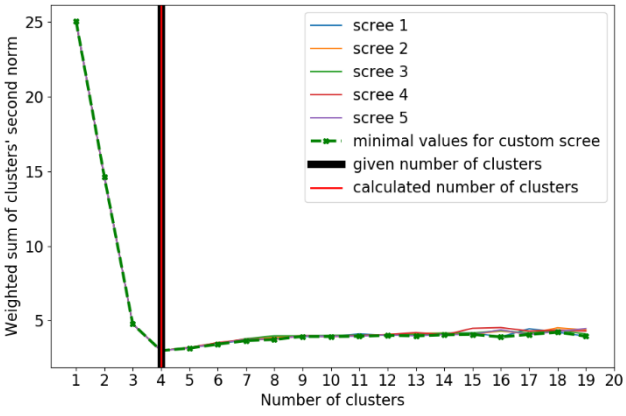

Figure 2: Scree plot with estimated number of clusters. Colored lines without marks are the scree values from different *k*-means run, gain though setting different random states. The green dash line with x-markers is the least values across all scree values, which are chosen for final estimation. The red vertical line is set at the lowest point of the plot, which represents the most probable number of clusters. The black vertical line is set on the correct number of clusters.

starting with one and ending with any number, which must be much larger than the maximal expected number of clusters [8].

After each *k*-means run, the calculation of mean squared differences (d_j) between coordinates of cluster's centre (x_c) and coordinates of each observation for every found cluster (x_i) occurs. The random seed must be reset after each run.

The squared Euclidian norm to each pair of the differences was calculated and summed up for each found cluster, and then divided by the number of observations in the cluster (n_j) (see eq. 1).

$$d_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\left\|\overrightarrow{x_{j,i}} - \overrightarrow{x_{j,c}}\right\|_2^2 \quad (1)$$

These values are then summed up again to get cumulative spread (D_k) for all clusters k in a *k*-means run (see eq. 2).

$$D_k = \sum_{j=1}^{k} d_j \quad (2)$$

The whole procedure should be repeated several times, say 100 totally, for different random initial values, i.e. random initial centroid placement, in order to gain multiple scree values to calculate the number of clusters. This is needed because of potential issues with random assignment of the initial coordinates for *k*-means' centroids, which sometimes delivers incorrect results, such influences must be minimized.

As result, a classic "elbow"-spot of the scree is transformed into "v"-spot (lowest point of the graph), which indicates the most probable number of clusters (see Figure 2). The scree plot is now actually become a valley plot. Note that this number of classes is assumed to be identical with the number of different spikes present in the data.

3 Results

Current version of the algorithm yields robust spike detection, which allows to collect information from all data channels (electrodes) because of multichannel approach (see the first two rows of the Figure 3) and concatenated spike samples (see the first plot in the third row of the Figure 3).

Novel method of customized scree-elbow plot (which could also be called "v-shaped-valley plot") allows to bypass problem of the predefined number of clusters – this estimation is now calculated and delivers stable results, as long as the correlation clouds stay separated (see Figure 2).

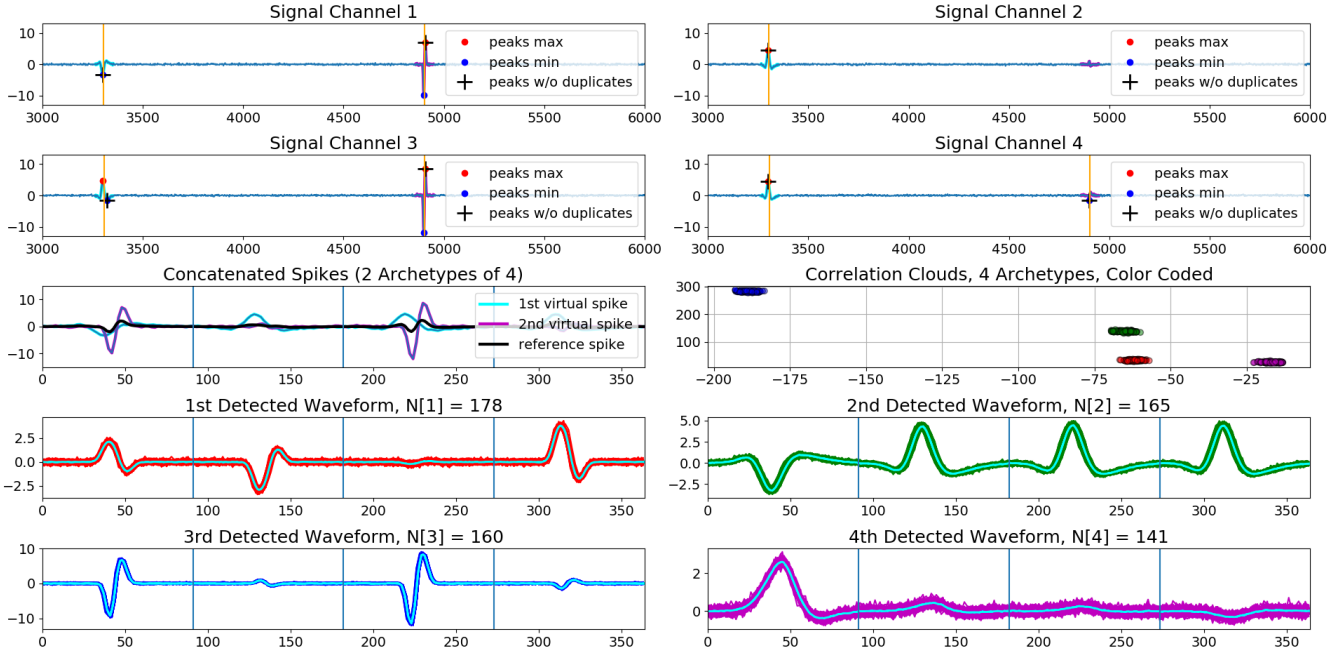

Figure 3: Final algorithm output. Upper two rows of plots represent peak detection and duplicate elimination. Detected waveform is highlighted with color, it's maximal and minimal point is marked with red or blue dot and thecenter of mass of it is marked with orange vertical line. On the top right plot there is no peak detected for one waveform, as the amplitude of the waveform is too low, however, robust peak detection algorithm assumes that there must be waveform there. On the first plot in the third row are two examples of spikes' concatenation for two different waveforms. Vertical blue lines mark the sample size, so one virtual composite sample consists of four segments. The second plot in the third row is the scatter of correlation clouds, which are color coded as they are clustered. The last two rows are the color-coded waveforms after clustering – cyan line represents mean of the waveform and overlapped red, green, blue and magenta lines represent cluster's observations as virtual composite samples.

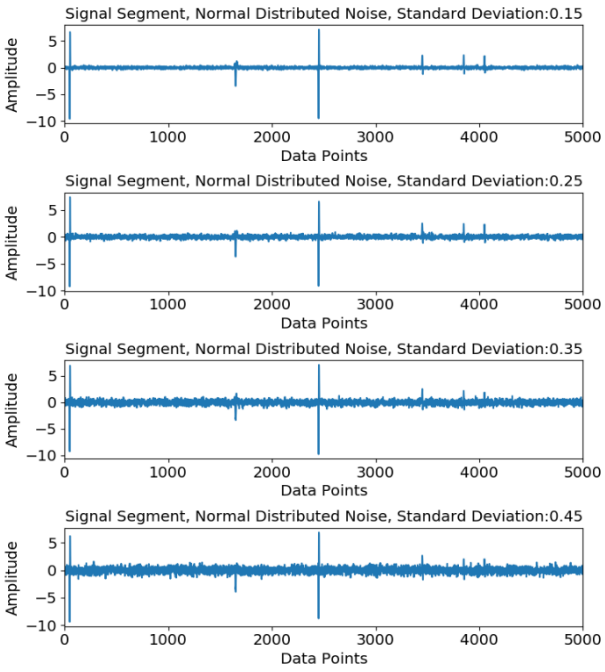

Figure 4: Increasing noise standard deviation levels. Spikes with small amplitudes on the 4th plot are almost indistinguishable from the noise.

Due to multivariate nature of the algorithm, definite separation of the correlation clouds was achieved (see the scatter in the third row of the Figure 3), which allows confident automated 100% correct clustering (see extracted waveforms on the last two rows of the Figure 3) within the standard deviation level for Gaussian white noise of 0.35 (see Figure 4). Mean maximal amplitude of the spikes in the simulated dataset across all four channels is thereby 5.40.

4 Discussion

The correlation based single channel spike sorting algorithm [3] was extended to a multiple channel one. In addition, a new method of calculation of the optimal number of classes for *k*-means clustering was developed, which reliably estimates number of different waveforms (spikes) present in the data.

The concatenating approach of related single channel spikes allows to analyse multivariate signals as united one without losing information from every channel. However, this can be viewed as dimension increase in order to obtain a "better" separation of the correlation clouds.

The use of a data driven reference spike to compute the correlation clouds is an easy understandable and efficient method to parameterize spike waveforms.

The algorithm needs to be improved and automated further, e.g. the part of peak detection needs to be enriched with optimal prominence estimation. The waveform sampling problem must be analysed more profound as a part of the statistical testing of the algorithm, as well as the introduced Monte Carlo method to estimate the number of *k*-means classes.

5 Conclusion & outlook

In this work, further possibilities of correlation-based spike sorting were shown. The usability and robustness of the previous approach were increased. The algorithm is simple and fast, it can be applied to any signals containing spike-like waveforms, e.g. ECG [6], and may also be used to discover possible irregularities of spike-like components in a wide range of biomedical signals.

Author Statement

Research funding: The authors state no funding involved. Conflict of interest: Authors state no conflict of interest.

References

[1] Rey, H. G., Pedreira, C., & Quian Quiroga, R. (2015). Past, present and future of spike sorting techniques. Brain Research Bulletin, vol. 119, pp. 106–117.

[2] Lewicki, M.S. (1998). A review of methods for spike sorting: the detection and classification of neural action potentials. Network: Computation in Neural Systems, vol. 9, pp. R53–R78.

[3] Larionov, P., & Schanze, T. (2018). Correlation Based Spike Sorting. Automed 2018 Tagungsband, pp. 71–73.

[4] McNaughton, B.L., O'Keefe, J.; Barnes, C.A. (1983). The stereotrode: A new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. Journal of Neuroscience Methods; vol. 8(4), pp. 391–397.

[5] Doerr, C., & Schanze, T. (2015). Are Heptodes Better than Tetrodes for Spike Sorting? IFAC-PapersOnLine, vol. 48(20), pp. 94–99.

[6] Larionov, P., Janßen, J.-D., & Schanze, T. (2018). Adaption of a spike sorting algorithm to ECG signals. Biomedical Engineering / Biomedizinische Technik, vol. 63(s1), p. 394.

[7] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.

[8] Lewith, G.T., Jonas, W. B.; Walach, Harald (2010). Clinical Research in Complementary Therapies: Principles, Problems and Solutions. Elsevier Health Sciences. p. 354.