COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 91 (2008) 232–244

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Performance evaluation of PCA-based spike sorting algorithms

*Dimitrios A. Adamos, Efstratios K. Kosmidis**, *George Theophilidis*

Laboratory of Animal Physiology, School of Biology, Aristotle University of Thessaloniki, 54 124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 12 August 2007
Received in revised form
8 April 2008
Accepted 28 April 2008

Keywords:
Spike sorting
PCA
Clustering
Noise
Single-electrode extracellular recordings

ABSTRACT

Deciphering the electrical activity of individual neurons from multi-unit noisy recordings is critical for understanding complex neural systems. A widely used spike sorting algorithm is being evaluated for single-electrode nerve trunk recordings. The algorithm is based on principal component analysis (PCA) for spike feature extraction. In the neuroscience literature it is generally assumed that the use of the first two or most commonly three principal components is sufficient. We estimate the optimum PCA-based feature space by evaluating the algorithm's performance on simulated series of action potentials. A number of modifications are made to the open source nev2lkit software to enable systematic investigation of the parameter space. We introduce a new metric to define clustering error considering over-clustering more favorable than under-clustering as proposed by experimentalists for our data. Both the program patch and the metric are available online. Correlated and white Gaussian noise processes are superimposed to account for biological and artificial jitter in the recordings. We report that the employment of more than three principal components is in general beneficial for all noise cases considered. Finally, we apply our results to experimental data and verify that the sorting process with four principal components is in agreement with a panel of electrophysiology experts.

1. Introduction

Extracellular recordings of spontaneous nerve activity using either hook or suction electrodes is a common practice for a number of electrophysiological experiments providing valuable information concerning peripheral and central nervous system physiology of vertebrates and invertebrates [1–6]. Extracellular electrodes record voltage potentials representing the activity of an unknown number of activated axons which may serve different functions. It is generally assumed that neurons encode information into series of action potentials (AP), their spike trains, so there is a special interest in reconstructing the waveform of individual neurons from the recorded trace. The procedure of proper assignment of spikes to neurons, in order to draw inferences from neural recordings, is referred to as neural spike sorting. Spike sorting from nerve activity is based on the assumption that the APs of a neuron have the same size and shape as they depend mainly on the axon's diameter and its distance from the electrode. The experimentalist has to identify the number of neurons from the recorded trace and classify each action potential into separate spike trains in a time consuming procedure that grows with the number of axons. The quality of spike sorting depends on the researcher's experience and his objective judgment. Consequently, a significant variability in human spike sorting performance has been noted [7]. Over the last decade, a considerable amount of research has been devoted to computer aided spike sorting

---

* *Corresponding author*. Tel.: +30 2310 998261; fax: +30 2310 998269.
E-mail addresses: dadam@bio.auth.gr (D.A. Adamos), kosmidef@bio.auth.gr (E.K. Kosmidis), theophil@bio.auth.gr (G. Theophilidis).

doi:10.1016/j.cmpb.2008.04.011

which is today an indispensable tool in neuroscience research [7–9,11–21].

A generic algorithm for spike sorting, implemented in popular commercial and open source titles, is as follows: spikes are extracted initially from the continuously extracellular recorded signal and a spike vector is created. The classification of detected spikes into multiple groups of neurons is based on spike shape characterization. In order to reduce the dimensionality of the data, features of the shape are selected to be used to represent the most prominent dynamics of spike waveforms. In the final step, clustering techniques are used to achieve the best cluster separation decision making.

Principal component analysis (PCA) is a powerful method employed to automatically select features and use them to create feature vectors [22]. PCA seeks an ordered set of orthogonal basis vectors, the principal components, which capture the directions in the data of largest variation [18]. A smaller subspace created by some of the initial principal vectors is then used to make an approximate projection of the data. In this projection, clusters of different units in the data, corresponding to separate neurons, are revealed. It has been argued that the use of the first two or more commonly three principal components is sufficient to accurately describe the spike [8,10,14–19,23–30]. A very popular choice for the final step is the expectation–maximization (EM) clustering algorithm [31]. The EM algorithm is an ideal candidate for solving parameter estimation problems. It computes probabilities of cluster memberships based on one or more probability distributions while the goal is to maximize the overall probability or likelihood of the final data.

Typical problems in spike sorting are the presence of noise and spike overlaps. Biological noise and noise from the recording devices may introduce problems in spike detection and in spike classification. With noise, similar APs belonging to different neurons may appear the same or APs belonging to the same neuron may appear different. Spike overlaps occur when two or more neurons fire simultaneously or almost simultaneously and produce APs of significant size. Depending on when the peaks and dips of the APs occur the size and shape of the resulting trace will vary. Another inherent weakness in spike sorting is the *a priori* ignorance of the number of active neurons present in the recorded trace. The *a posteriori* supervision of the classification results by the researchers and their decision making on whether more or less clusters should be considered introduce the human error factor in the process.

In this paper we empirically estimate the size of PCA-based feature space in a typical semi-automatic spike sorting approach. Statistical approaches to estimate this subspace dimension in a blind fashion also exist in the literature (for a comparative study see in [32]). We apply the PCA–EM combination to simulated spike trains representing single electrode records from the neural cord of the beetle *Tenebrio molitor*. These neurons display spontaneous activity and multiple Single Fiber Action Potentials (SFAP) can be seen on the trace. We evaluate the optimum volume of principal components that participate in the spike sorting approach under different types and levels of background noise. This is possible based on our *a priori* knowledge of the number of simulated neurons, the exact occurrence timings and overlaps of the APs generated by each one of them. Finally, we apply the method to our experimental data and compare the results with those obtained by a group of neurophysiology experts.

The remainder of this paper is organized as follows: in Section 2, we give the background mathematics on the proposed simulation method and the part of the spike sorting methodology that we exploited during the evaluation procedure; in Section 3, we describe the detailed implementation of our work; and in Section 4, we present the results derived from the evaluation of the spike sorting method on the simulated and experimental data.

2. Background mathematics

2.1. Single fiber action potential model

The model for SFAP is a damped sinusoid as suggested in [33]:

$$f(t) = A \sin\left(\frac{t}{\tau_1}\right) e^{-t/\tau^2} \qquad (1)$$

where A, τ_1 and τ_2 are parameters that determine the amplitude, rising phase rate and the total duration of each SFAP, respectively. Spike, AP and SFAP will be used interchangeably throughout this text. A small amount of bound-limited, uniformly distributed jitter was added to all three parameters to account for naturally occurring variability in spike shape of the same axon.

A dead time Poisson process (DTPP) with rate λ, defined for each unit separately, and absolute refractory period $\Delta = 2.5$ ms has been used to generate spike interval times for each unit [34–36].

2.2. Principal component analysis

In PCA a spike vector is constructed in an m-dimensional space, where m is the number of measurement types. Forming an $n \times m$ matrix of values, each of the n rows represents an object that can be regarded as an m-dimensional vector, a row vector in R^m. In R^m, PCA searches for the best-fitting linear combined set of orthogonal axes to replace the initial set of m axes in this space. The idea behind this step is to find a set of $m' < m$ principal axes allowing the objects to be adequately characterized on a smaller (m'-dimensional) space, while the m–m' dimensions may be ignored as describing noise. In PCA, the projections of points on the axis sought for, need to be as elongated as possible, i.e. the variance of the projections needs to be as great as possible. The eigenvectors associated with the m' largest eigenvalues yield the best-fitting m'-dimensional subspace of R^m. For an excellent review on PCA and its applications see in [37].

3. Methodology

3.1. Spike train generation

In order to evaluate the algorithm's performance in more realistic situations we have modeled series of randomly distributed action potentials with different characteristics (A, τ_1