ELSEVIER

Journal of Neuroscience Methods 163 (2007) 267–282

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Quantifying the isolation quality of extracellularly recorded action potentials

Mati Joshua [a,*], Shlomo Elias [b], Odeya Levine [b], Hagai Bergman [a,b,c]

[a] *The Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem 91904, Israel*
[b] *Department of Physiology, The Hebrew University-Hadassah Medical School, Jerusalem 91120, Israel*
[c] *Eric Roland Center for Neurodegenerative Diseases, The Hebrew University, Jerusalem 91904, Israel*

Received 11 April 2006; received in revised form 18 March 2007; accepted 18 March 2007

Abstract

There have been many approaches to the problem of detection and sorting of extra-cellularly recorded action potentials, but only a few methods actually quantify the quality of this fundamental process. In most cases, the quality assessment is based on the subjective judgment of human observers and the recorded units are divided into "well isolated" or "multi-unit" groups. This subjective evaluation precludes comprehensive assessment of single-unit studies since the most basic parameter, i.e. their data quality, is not explicitly defined. Here we propose objective measures to evaluate the quality of spike data, based on the time-stamps of the detected spikes and the high-frequency sampling of the analog signal of cortical and basal-ganglia data. We show that quantification of recording quality by the signal-to-noise ratio (SNR) may be misleading. The recording quality is better assessed by an isolation score that measures the overlap between the noise (non-spike) and the spike clusters. Furthermore, we use a nearest-neighbors algorithm to estimate the proportion of false positive and false negative classification errors. To validate these quality measures, we simulate spike detection and sorting errors and show that the scores are good predictors of the frequency of errors. The reliability of the isolation score is further verified by errors implanted in real basal ganglia data and by using different sorting algorithms. We conclude that quantitative measures of spike isolation can be obtained independently of the method used for spike detection and sorting, and recommend their reports in any study based on the activity of single neurons.

Keywords: Spike sorting; Multi-unit recording; Signal detection

1. Introduction

The problem of extracting single neuron activity from extracellular recordings has been investigated extensively and comprehensively reviewed (e.g. Lewicki, 1998). The process of detecting action potentials from the extracellular waveforms (spikes) and clustering them into different neuronal sources is known as *spike detection and sorting*. Spike detection and sorting algorithms are not perfect and classification errors can occur for a number of reasons. First, most algorithms are not fully automatic (e.g. Abeles and Goldstein, 1977; Worgotter et al., 1986; Bergman and DeLong, 1992) and their real-time use can lead to human errors (Wood et al., 2004). Second, inaccurate assumptions about the data can also lead to errors. Some algorithms presuppose a parametric statistical model (Lewicki, 1994; Pouzat et al., 2002, 2004; Shoham et al., 2003), whereas other algorithms are based on non-parametric assumptions (Fee et al., 1996a). In both cases these assumptions, whether explicit or implicit, may be violated. For example, the analog trace in Fig. 1 shows significant modulation of the spike waveforms and illustrates how the stationarity (waveform stability) assumption may be violated and thus lead to classification errors.

Although many approaches to the problem of sorting spikes have been put forward, only a few methods have been developed to quantify the quality of the spike sorting (Harris et al., 2001; Pouzat et al., 2002; Schmitzer-Torbert et al., 2005). In most cases, the quality assessment is done subjectively by a human observer, and units with high scores are then reported as having a "high signal-to-noise ratio" and being "well isolated". These subjective reports do not permit comparison of data quality across different studies and unfortunately are predisposed to personal bias.

* Corresponding author. Tel.: +97226757168.
E-mail address: mati@alice.nc.huji.ac.il (M. Joshua).

0165-0270/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2007.03.012

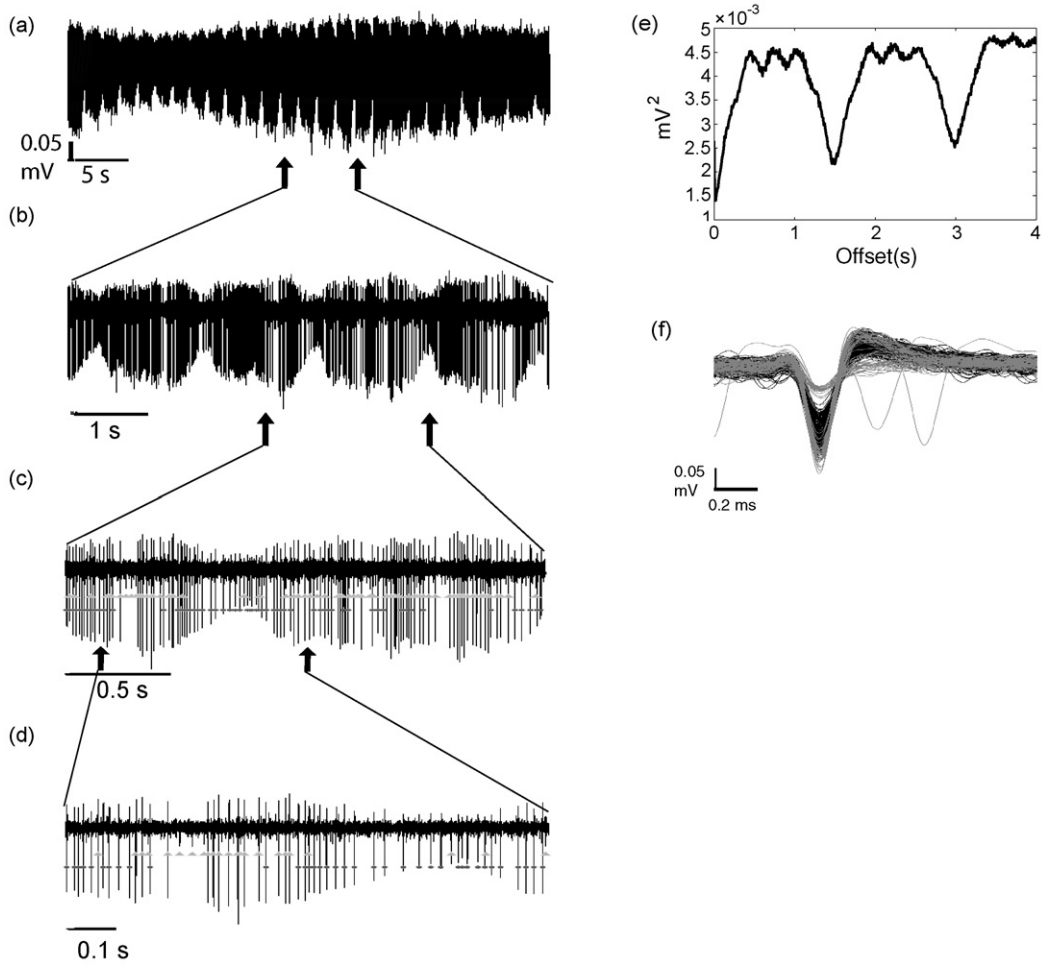

Fig. 1. An extreme example of non-stationarity of extracellular recording. Instability in the extracellular recording can lead to misclassifications by the spike sorting algorithm. (a–d) A single trace of the extracellular recording, at different time scales (b depicts the arrow-marked interval in (a), etc.). In (c and d) spikes detected by the real-time spike sorter are marked by black dots; high noise events (events that crossed threshold but not classified as spikes) are marked by gray triangles. (e) Average of squared peak-to-peak differences of spike waveform, over all time intervals as a function of time starting from the real-time detection. This is a gross measure of the change in spike waveform shape in time, similar to the autocorrelation function. Note the large periodic changes at 0.66 Hz and small periodic changes at ∼3 Hz. These are probably due to periodic changes in electrode position, caused by respiratory (∼40 min^{-1}) and cardiac (∼180 min^{-1}) waves, respectively. (f) Events classified in real-time as a single unit are in black; the noise events are in gray. Note that the noise cluster contains two different classes of events. One class forms a smooth continuum with the boundaries of the spike cluster (probably missed spikes). The other class can be dissociated from the spike cluster due to its smaller waveforms (probably spikes from other units). The scores of this unit are: isolation score, 0.93; false negative score, 0.12; false positive score, 0.002; $\text{SNR}_{\text{No Spk}}$, 5.35; SNR_{Spk}, 5.26. The false negative score is suggestive of the instability shown above.

In this article we propose objective measures to assess the quality of spike detection and sorting. Our measures quantify two different aspects of the data:

1. Quality of the recording, by calculating SNR (Section 3.1). We present and discuss two calculations of the SNR that differ in their noise estimation. The first is based on the noise when an action potential occurs and the second is based on the noise between action potentials.
2. Clustering quality. We introduce an *isolation score* for quantifying the overlap between the spike and the noise (non-spike) clusters (Section 3.2). We then present classification error scores that estimate the fraction of events that were misclassified as spikes (false positive errors) or misclassified as noise-events (false negative errors) (Section 3.3).

To validate these measures, we simulate spike-sorting errors (Section 3.4.1) and test the isolation score and classification error scores as a function of the fraction of simulated errors for different units. We check the scores under different conditions by applying several clustering algorithms (Section 3.4.2). We use real data from the basal ganglia and simulated errors to investigate the score parameter space (Section 3.4.4). Finally, we compare the results of the different scores (Section 3.4.5).

2. Methods

2.1. Neuronal recording procedures

The data were collected from experiments performed on two vervet monkeys (Monkey Cu: *Cercopithecus aethiops*, female, weighing 3.5–4 kg and monkey T, female, weighing 3 kg) and

two *Macaque fascicularis* (monkey Y, male, weighing 5 kg and monkey P, female, weighing 3 kg). Details of the behavior of the monkeys and animal care are described elsewhere (Heimer et al., 2002; Morris et al., 2004; Elias et al., 2007). Recordings were made in the external segment of the globus pallidus (GPe), a central nucleus of the basal ganglia and in the primary motor cortex (M1, monkey T only). Animal care and surgical procedures were in accordance with the *NIH Guide for the Care and Use of Laboratory Animals* (1996) and the Hebrew University guidelines for the use and care of laboratory animals in research, supervised by the institutional committee for animal care and use.

During the recording sessions, glass-coated tungsten microelectrodes (impedance at 1 kHz equals 0.2–0.6 MΩ) were advanced to the target. Neuronal activity from each electrode was amplified (monkey P Y and T × 5000, monkey Cu × 10,000), filtered (monkey P and Y: 1–6000 Hz, monkey Cu and T: 300–6000 Hz), and continuously sampled at 24 kHz/electrode (AlphaMap, Alpha-Omega Engineering, Nazareth, Israel).

2.1.1. Real-time spike detection and sorting algorithm

The electrode output was processed and classified in real time (MSD, ASD, Alpha-Omega Engineering) by a template-matching algorithm (Worgotter et al., 1986). The electrode signal was continuously sampled at 36–50 kHz, placed in a buffer containing the last 100 samples (2–2.8 ms), and compared continuously with one to three templates. Each template was constructed of eight equally spaced points separated by five sampling points (e.g. 0.1 ms for the 50 kHz sampling), and was defined by the experimenter following a learning process of threshold crossing signals. The sum of squares of the differences between eight points in the buffer and the templates was calculated. When this sum reached a minimum that was below a user-defined threshold, detection was hardware reported. In cases where a buffer was double matched (e.g. a signal passed the criterion of more than one template), an error signal was given to the user, but no hardware report was created. A dead time of 0.06 ms followed detection. The timing of the hardware detections (100 μs active-low TTL pulses) was edge sampled at 12 kHz (33 kHz in monkey T) in parallel with the analog signals of the electrode output. During recording sessions, the experimenter closely followed the spike shape and discharge rate. The experimenter graded the isolation quality approximately every 2–4 min (see below) and when necessary adjusted the template, detection threshold, or rarely the electrode position.

2.2. Real-time grading of isolation quality

In most cases one experimenter controlled the position and spike sorting of four electrodes. The quality of the detection and spike sorting was estimated on-line experimenter. This quality estimation was based on the superimposed analog traces of the recently (20–100) sorted spikes as well as the waveforms of events that passed an amplitude threshold that was set by the experimenter but were not classified as spikes. The grade scale ranged from 1 (highest score) to 8 (lowest score). Generally, a score of 1 meant perfect isolation, where the experimenter judged that close to 100% of the spikes emitted by a single neuron were detected with no false detections (zero false positive and negative errors). Scores between 3 and 4 meant that most (but not all) spikes generated by a neuron were detected (small fraction of false negative errors), but still with a negligible fraction of false detections ("no false positive errors"). Scores of 5–6 meant a mixture of two to three units. Finally, a grade of 7–8 meant a recording of multi-unit activity (significant fraction of false negative and positive errors).

2.3. Algorithm development

All the functions used for both the analysis and algorithm implementations are Matlab 7.1 (Mathworks, Natick, MA, USA) compatible, and are available at: http://alice.nc.huji.ac.il/~mati/sorting_quality_programs.

2.4. Data preprocessing and event representation

The initial input for the estimation of the spike isolation quality consisted of the time stamps of the detected spikes (spike trains) and the entire analog signal. We defined two clusters of events: (i) *spike cluster* – a cluster classified as a single unit and (ii) *noise cluster* – a cluster of events not classified with that unit. The spike cluster was simply constructed from segments of the analog signal according to the spike trains. The noise cluster was extracted from the same analog trace, and contained events that were not detected as spikes of the given unit. However, the noise events had some similarity to the events in the spike clusters (e.g. similar amplitude, see details below). Each event, in both spike and noise clusters, was represented as a point in a high-dimensional space. Fig. 2 and Sections 2.4.1–2.4.3 depict step by step the extracting of the spike and noise clusters.

2.4.1. Up-sampling using cubic spline

Discrete sampling of analog traces leads to a time jitter between superimposed frames of events (Fee et al., 1996a; Pouzat et al., 2002). This jitter contributes to the variability in extracellular waveforms from the same cell (Fig. 2c). To reduce this variability, we up-sampled the data using cubic spline interpolation (Fig. 2d). The factor by which we up-sampled the data was fixed at 4; using this value for the up-sampling factor had a maximal effect on the reduction of variability between our data waveforms (data not shown).

2.4.2. The spike cluster

To represent an event we used 1.5 ms (144 sampling points after the cubic spline interpolation) of the corresponding up-sampled analog trace. The resulting vector, whose ith value is the voltage measured after i time steps from the beginning of the event, can be viewed as a point in a high-dimensional space:

$$\vec{X} = (V_1, V_2, \ldots, V_{144})' \quad (1)$$

All events were aligned by the largest negative peak. The offset of this peak from the beginning of the event was set to 0.5 ms

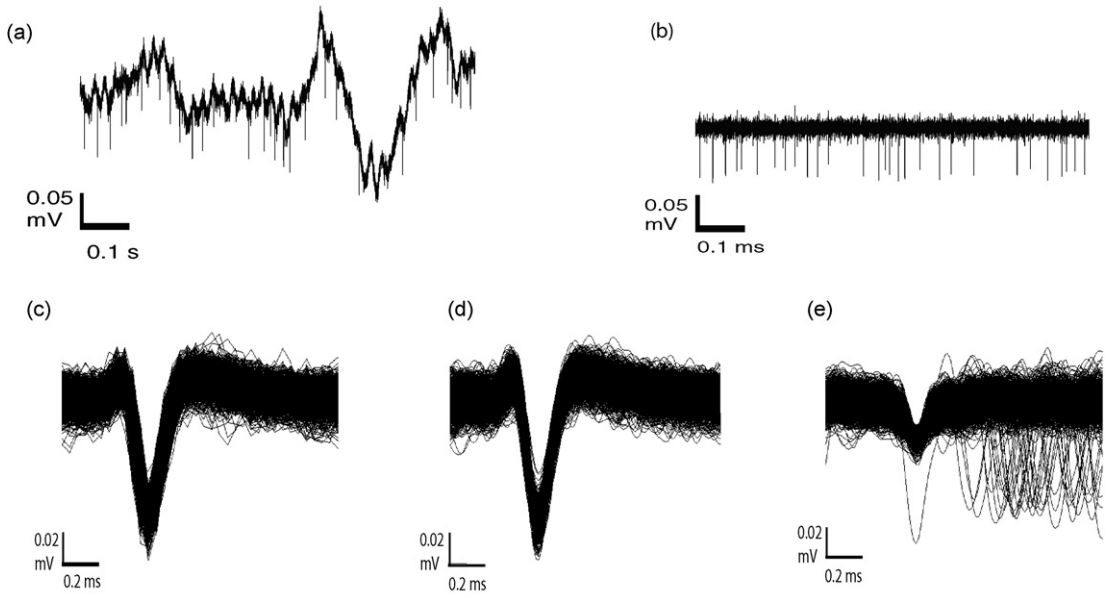

Fig. 2. Preprocessing. The process of extracting the spike and noise clusters. (a) The raw analog trace (1–6000 Hz band-pass hardware filtered and digitally sampled at 24 kHz). (b) Analog trace after filtering with a digital high-pass filter (>300 Hz, two pole Butterworth filter, a zero-phase forward and reverse digital filter, Matlab filtfilt function). (c) Superimposed waveforms of spikes extracted according to classification by the spike sorter (*spike cluster*), and aligned to the largest negative peak (before the spline upsampling). Note the large variability during the fast phase of the action potential that results from the limited sampling rate. (d) Spike cluster after upsampling the events and realignment to the negative peak revealing reduction in variability compared to (c). (e) The noise cluster, detected by threshold crossing. The same upsampling and alignment process was used. The scores of this unit are: isolation score, 0.98; false negative score, 0.01; false positive score, 0; $SNR_{No\ Spk}$, 2.57; SNR_{Spk}, 2.52.

(i.e. V_{48}). Our extracellular recordings are from neurons in the GP and the primary motor cortex with a large negative phase. Hence, we used the largest negative peak for alignment of the spike vector (however one can easily generalize this algorithm to positive peaks). Finally, the aligned vector was normalized to have a zero mean. The cluster of up-sampled, aligned and normalized spike events is denoted as S_{cluster}.

2.4.3. The noise cluster

Detection of the events comprising the noise cluster was based on threshold amplitude crossing. Because of the typical spike shape in our extra-cellular recordings, we only used a negative (lower) threshold to detect the noise cluster (however one can easily generalize this algorithm to upper or dual, i.e. upper and lower, threshold crossing events).

The noise cluster was constructed in the following manner. First we selected from S_{cluster} the 2% of the spikes with the smallest negative peak (closest to zero).We then took the average of these negative peaks and defined the threshold as half of this value:

$$\text{threshold} = \frac{\text{average negative peak}_{0.02}}{2} \quad (2)$$

where 0.02 is the fraction of spikes used for calculation of the average negative peak.

Next, we identified all events that crossed this threshold, but removed the events already marked as spikes. Finally, we up-sampled and aligned the noise events (0.5 ms offset similar to the spike waveforms) by the local minimum between the first two (down and up) threshold crossings (Fig. 2e).

The noise cluster models all high amplitude events that are not classified as spikes from the given unit. The noise-cluster should contain all unclassified putative spikes; i.e. events that are close, but not in, the spike cluster. This is achieved by using only the S_{clsuter} events with the smallest negative peaks to determine the threshold. The spike sorting quality measures are insensitive to inclusion of more noise event crossings; i.e. with a more conservative threshold. To verify this insensitivity we modified the threshold parameters and found that the quality measures were stable (see below). Therefore, we recommend the use of a conservative threshold that ensures that the noise cluster contains putative spikes even when the noise cluster is overly large.

3. Results

Spike detection and sorting quality depends first on the recording quality and then on the quality of the clustering algorithm. To evaluate recording quality we used the signal-to-noise ratio (SNR) (Section 3.1). Although the SNR can be used for initial estimation of recording quality and a high SNR is usually a necessary condition for good unit isolation, the SNR is not a direct measure of the isolation of a single unit. Sorting of recordings with a high SNR may nonetheless result in a spike cluster that excludes spikes (false negative errors, e.g. Fig. 1) or a cluster composed of two large units (false positive errors). We therefore applied more direct measures of cluster quality by measuring the isolation of the spike cluster from the noise cluster: the isolation score (Section 3.2), and false positive and false negative measures (Section 3.3). In Section

3.4 we compare and validate the scores under different sorting methods and simulated error frequencies.

3.1. Signal-to-noise ratio measures

Several previous studies have taken a initial step towards assessing the quality of spike data by reporting some variants of the spikes SNR (Pare and Gaudreau, 1996; Likhtik et al., 2005). However, there is no explicit definition of the SNR in these reports, making them very difficult to compare. The spike signal-to-noise ratio can be computed in two ways. Both methods compute the signal in the same fashion but differ in their noise calculation.

3.1.1. Signal calculation

The average of S_{cluster} (up-sampled and aligned by the negative peak) is defined as:

$$S_{\text{avg}} \equiv \frac{1}{|S_{\text{cluster}}|} \sum_{\vec{x} \in S_{\text{cluster}}} \vec{X} \tag{3}$$

We quantify the signal as the difference between the minimum and the maximum of the average spike waveform (Fig. 3a):

$$\text{peak-to-peak} \equiv \text{Max}(S_{\text{avg}}) - \text{Min}(S_{\text{avg}}) \tag{4}$$

We prefer using the peak-to-peak to quantify the signal rather than other methods that integrate the area enclosed by the spike waveform (spike energy). This is because the peak-to-peak signal value does not depend on the duration of the spike waveform, which is conditioned by the filter and the edge detection parameters.

3.1.2. Noise calculation

We quantified noise in two ways:

1. The noise underlying the spike events, which corresponds to the intra-cluster variability (Fig. 3b). For each spike waveform, X_k, we subtracted the mean waveform, S_{avg}, to produce Resid_k. We then concatenated all resulting residuals to produce one long vector Resid. The noise is then defined as the standard deviation of this vector:

$$\text{Noise}_{\text{Spk}} \equiv \text{S.D.(Resid)} \tag{5}$$

   Since our filters exclude the low frequencies and we use segments larger than most spikes, we can disregard the increase in variability that may be created by the concatenation process.
2. The noise from the inter-spike-intervals (Fig. 3c), where spikes are defined as events in the signal cluster. For each spike in the signal cluster, we extract the 1.5–3 ms period before the spike event negative peak, Prev_k (Fig. 3c), unless another spike from the cluster occurred in that interval. We then concatenate all such intervals to produce one long vector

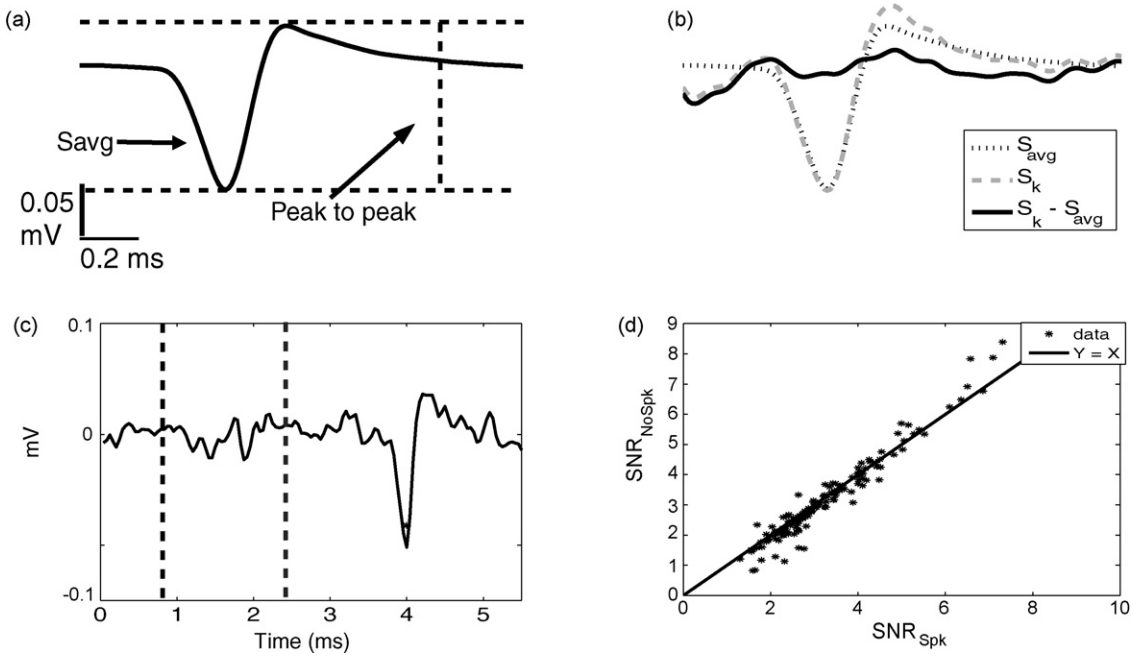

Fig. 3. Signal-to-noise ratio calculation. (a) Average spike waveform (solid line) and peak-to-peak (dotted lines). (b) $\text{Noise}_{\text{Spk}}$. For a given spike event (dashed gray line), we subtracted the average spike waveform (dotted black line) which results in the noise during the spike event (the solid black line). We then concatenated segments from all spikes and calculated the standard deviation of this vector. (c) $\text{Noise}_{\text{No Spk}}$. For a given spike we extracted the analog trace 1.5–3 ms before the negative peak of the spike event (between the dashed lines). We then concatenated all these traces from all spikes and calculated the standard deviation of this vector. (d) $\text{SNR}_{\text{No Spk}}$ vs. SNR_{Spk}. The SNR scores are the ratios between the peak-to-peak and the noise estimations (S.D. × 5). The SNR scores for 155 GP units. Scores are highly correlated ($R^2 = 0.94$). Generally, for high values—$\text{SNR}_{\text{No Spk}}$ is larger than SNR_{Spk} (large values tend to be above the line $Y = X$), as changes in the waveform that contribute to $\text{Noise}_{\text{Spk}}$ but not to $\text{Noise}_{\text{No Spk}}$ are more likely when the electrode is close to the cell. On the other hand, when the scores are small, SNR_{Spk} is larger (small values are below the $Y = X$ line), probably due to failure in detecting overlapping spikes.

Prev. The noise is then defined equivalently:

$$\text{Noise}_{\text{No Spk}} \equiv \text{S.D.(Prev)} \tag{6}$$

The two SNR scores are highly correlated in our data (Fig. 3d). There are some cases where these two measures are not equal. When the waveform of a single unit changes (e.g. due to electrode drift, or intrinsic firing properties), or when the spike cluster actually reflects multi-unit instead of single-unit activity, $\text{Noise}_{\text{Spk}}$ will be larger than $\text{Noise}_{\text{No Spk}}$. Surprisingly, the opposite can also occur. For example, when the spike of a second unit temporally overlaps with the spike of the given unit, the sorting algorithm may drop these spikes (Bar-Gad et al., 2001). As a result, the second unit will contribute only to $\text{Noise}_{\text{No Spk}}$ as its coincidence with the first spike is ignored.

3.1.3. Signal-to-noise ratio

We define the two signal-to-noise ratios as simply:

$$\text{SNR} \equiv \frac{\text{peak-to-peak}}{\text{Noise} \times C} \tag{7}$$

where Noise is calculated by one of the two methods and C is a scaling factor (commonly set by us to 5) which scales the noise measures to peak-to-peak equivalent units. Examples of spikes and their SNR measures are given in Figs. 1, 2, 6 and 8.

3.2. Isolation score

As stated above, SNR might be problematic, especially in cases where the spike cluster actually reflects high amplitude multi-unit activity. We therefore then assessed the quality of the spike isolation directly. The isolation score quantifies the distance between the spike cluster and the noise cluster. We computed this distance on the raw events directly, without mapping the spikes to some feature space, e.g. PCA (Abeles and Goldstein, 1977) or wavelet transform (Quiroga et al., 2004; Nenadic and Burdick, 2005).

3.2.1. Mandatory features of the isolation score

The isolation score needs to exhibit several critical properties:

1. The score should decrease with the number of real spikes missed by the sorting algorithm (false negatives).
2. The score should decrease with the number of noise events that were classified as spikes (false positives).
3. The score should be insensitive to the size of the extracted noise cluster.
4. The score should span an intuitive range, e.g. 0–1.

3.2.2. Isolation score: definition

The isolation score quantifies the distance between events in the spike cluster to the noise cluster. Nevertheless, since we are only interested in the spike cluster events, this measure is not symmetric. First, we compute the normalized similarity between each event in the spike cluster, X, to all other events (spikes and noise), Y:

$$\text{Similarity}(X, Y) \equiv \exp\left(\frac{-\text{d}(X, Y)\lambda}{d_0}\right) \tag{8}$$

where d(X,Y) is the Euclidean distance between vectors X,Y. Note that Similarity(X,Y) between close events is close to one (exp(0)), whereas between distant events it is closer to zero (exp($-\infty$)). d_0 is the average Euclidian distance in the spike cluster; this parameter normalizes the Euclidian distance to avoid dependence on the units of a particular data set. The exponent function stretches the Euclidean distance nonlinearly; thus Similarity(X,Y) of remote events become infinitesimally small. λ is a gain constant ($0 < \lambda \ll \infty$) that sets the gain of this stretch. With $\lambda \ll 1$, all events are similar and Similarity(X,Y) is close to one, whereas with $\lambda \gg 1$, all events are dissimilar and Similarity(X,Y) become infinitesimally small.

In order to turn the above similarity index into a probability-like quantity (positive values that sum to 1), we normalize it by the sum of similarities between a given event, X, from the spike cluster, to all other events (spikes and noise):

$$P_X(Y) \equiv \frac{\exp(-\text{d}(X, Y)(\lambda/d_0))}{\sum_{Z \neq X} \exp(-\text{d}(X, Z)(\lambda/d_0))} \tag{9}$$

For each event X we get a function P_X that takes the form of the Boltzmann–Gibbs distribution, also known as "softmax" (Goldberger et al., 2004). Note that the parameter λ controls the "softness" of the max operation; i.e. λ behaves like 1/temperature in some notations of the softmax equation. For a given event X when λ approaches infinity (zero temperature) $P_X(Y)$ is the deterministic probability function; i.e. $P_X(Y) = 1$ for the event nearest X and zero for all other events. On the other hand when λ approaches zero (maximal temperature) $P_X(Y)$ is the uniform distribution; i.e. $P_X(Y)$ is equal for all events. In this manuscript we used $\lambda = 10$; i.e. we stretched the distances between near and remote events.

In the next step, for each event in the spike cluster, X, we sum over all the normalized similarity values $P_X(Y)$ for all the Y's in the spike cluster:

$$P(X) \equiv \sum_{Y \in S_{\text{cluster}}} P_X(Y) \tag{10}$$

$P(X)$ is therefore a measure of how close event X is to the spike cluster compared to the noise cluster. Intuitively, $P(X)$ is the probability that event X belongs to the spike cluster. The calculation of $P(X)$ is illustrated in Fig. 4 (note that $P(X)$ and $P_X(Y)$, Eqs. (10) and (9), respectively, are not the same).

The isolation score is defined as:

$$\text{isolation score} \equiv \frac{1}{|S_{\text{cluster}}|} \sum_{X \in S_{\text{cluster}}} P(X) \tag{11}$$

and can be intuitively considered as the average probability that an event classified as a spike belongs to the spike cluster. Thus, our isolation score is a combination of two approaches:

1. Quantifying the connectivity between two clusters using the energy at the interface of the two groups (Fee et al., 1996a).

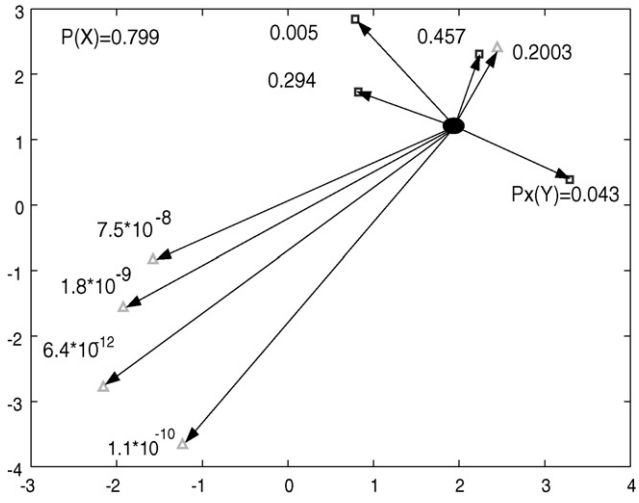

Fig. 4. Calculation of the isolation score. Calculating the proximity of a spike to the spike cluster, relative to the noise cluster. This figure is a schematic representation of the isolation score calculation. The *x*–*y* coordinates represent the 144 dimensions of a waveform from the spike and noise cluster. The gray triangles represent points in the *noise cluster*, whereas the black squares represent spike events in the *spike cluster*. For a given point *X* in the spike cluster (black oval), the numbers next to each of the other points, *Y*, are $P_X(Y)$. The arrows denote the Euclidian distance. Finally, $P(X)$ for the given point (black oval), is the sum of all $P_X(Y)$ values for all other spike events (black squares). Note that for events far from *X*, $P_X(Y)$ is infinitesimal, and hence they have only a small influence on the $P(X)$. On the other hand, noise events that are close to the spike cluster significantly decrease the $P(X)$ values (e.g. gray triangle in the upper right-hand corner).

2. Grading the distance of two events using the "softmax over Euclidean distances" function (Goldberger et al., 2004).

The range of the isolation score is from 0 to 1, where a score of 1 means ideal isolation, with minimal distances between the elements of the spike cluster, and a large distance between them and all the elements of the noise cluster. A score close to zero means very poor isolation, where the Euclidian distance among elements in the spike cluster is larger than the Euclidian distances between them and the noise cluster; i.e. elements from the spike cluster are surrounded by elements from the noise cluster.

The isolation score satisfies the requirements defined in Section 3.2.1:

1. Spikes that were missed by the spike-detection or sorting algorithm (false negatives) are nonetheless close to the spike cluster. As a result, events in the cluster that are close to such misses will have a reduced $P(X)$ (Fig. 4), which in turn will reduce the overall isolation score.
2. Likewise, noise events that were classified as spike events (false positives) are close to the noise cluster. For these false positives the $P(X)$ value is reduced, due to their proximity to the other noise events, thus again reducing the overall isolation score.
3. The isolation score is insensitive to the size of the noise cluster. This is a result of the exponential decay of the similarity value, $P_X(Y)$, between a spike event *X* and a distant event *Y*. Therefore, adding more noise events, which are mostly distant from the spike cluster, contributes only small additional values to $P_X(Y)$.
4. The isolation score is the average of probability-like values and hence is bounded between 0 and 1.

It is crucial to note that the isolation score does not measure the distance between the noise and spike distributions directly. Nor does it directly measure the performance of the clustering procedure. Rather, it measures how far away the noise and the spike clusters are. It is similar to the gap measure common in classification discussions, except that it recognizes that there is no real gap between the two clusters.

3.3. Scores of classification errors

As described in the previous Section 3.2, the isolation score quantifies the separation of the spike cluster from other events, but it does not estimate the number of spikes missed by the spike detection and sorting process or the number of noise events that were erroneously classified as spikes (false negative and positive errors, respectively). Moreover, the isolation quality measure cannot separate these errors. However, some physiological studies are more sensitive to one of the two errors and therefore their separate estimates may provide a better database for these experiments. In this section we describe a method for estimating these errors. For each event we find its *K* nearest neighbors (KNN) (Vapnik, 1998) and compare the classification of the majority of these neighbors to the event classification (produced by the sorting algorithm). This method is illustrated in Fig. 5.

3.3.1. False negatives score

False negatives are spikes that were missed by the spike detection or sorting algorithm. We estimate these by the number of noise events having most of their *K* nearest neighbors (see below

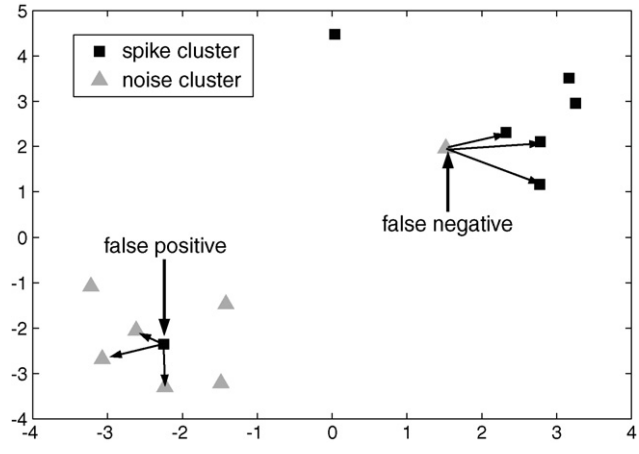

Fig. 5. Illustration of the KNN algorithm for estimating classification error scores. The *x*–*y* coordinates represent the 144 dimensions of a waveform from the spike and noise cluster. Black squares represent spike events, gray triangles represent noise events. The notations are similar to those used in Fig. 4. For each event we calculated the *K* nearest neighbors (KNN, here $K = 3$). Spike events having most of their KNN from the noise cluster are considered *false positives*; similarly noise events with a majority of their KNN from the spike cluster are considered *false negatives*.

for the choices of K) from the spike cluster, denoted N_{fn}. The false negatives score is thus defined as:

$$\text{false negatives score} \equiv \frac{N_{fn}}{N_{fn} + |S_{cluster}|} \quad (12)$$

When the number of false negatives is small the score is close to zero. The score then increases with the number of false negatives. When all the real spike events are missed ($N_{fn} \gg |S_{cluster}|$) the score reaches the maximum of 1. However, in practical terms the score cannot reach this limit (see below Section 3.3.4).

3.3.2. False positives scores

Similarly, we estimate the number of false positive events (events that were classified as spike events but are noise events):

$$\text{false positives score} \equiv \frac{N_{fp}}{|S_{cluster}|} \quad (13)$$

where N_{fp} is the number of spike events having most of their K nearest neighbors as noise events. When the number of false positives is small the score is close to zero. As this number increases the score increases. When all spike cluster events are surrounded by noise events ($N_{fp} = |S_{cluster}|$) the score reaches the maximum of 1.

3.3.3. Choosing K

Choosing inappropriate values of K leads to biases in the classification error scores. For example, if too small a value is chosen for K, false negative events may erroneously lead their correctly classified spike-event-neighbors to be considered as false positives. Likewise, to take an extreme example, when the spike cluster is larger than the noise cluster, using a K value that is larger than twice the size of the noise cluster will cause all noise events to be considered false negatives. Generally large values of K may lead to biased estimations of error rates of events that are close to the boundaries of the clusters.

In this study, we selected an intermediate value for K, such that a small number of clustering errors did not cause a large bias, and the K value was far smaller than the size of both clusters. Typically, our validation tests were performed on clusters that contained 1500 events, using $K = 31$. In our experience, a good rule of thumb is that K should equal 1–5% of the number of events in the signal cluster.

3.3.4. Using the classification error scores

The classification error scores are a refinement of the isolation score. These false positive and negative estimates may help constrain neuronal data analysis. For example, the existence of false positives is one reason one should not expect to find a perfectly oscillatory cell, or should not be surprised by multi-parameter encoding of a single neuron. Naïve use of these error scores, however, may be misleading. When the spike and noise clusters overlap highly these scores are biased (Figs. 6b and c, 7c and d and 9c and d). Thus, when a large ratio of the spike events are missed, real spike events from the spike cluster may have most of their KNN from the noise cluster and hence be considered false positives; for the same reason the estimation of false negatives will be low. This is the reason we argued (Section 3.3.1) that the false negative score will not reach its theoretical upper limit of 1. Similarly, real noise events may have most of their KNN from the spike cluster and hence be considered false negatives. Nonetheless, the isolation score measures the overlap between the spike and the noise clusters. When the isolation score is high the classification error scores are good estimates of the frequencies of false positive and negative errors and can be used. When the isolation score is low, the error classification scores are biased; however low isolation scores should dissuade us from using the data and therefore further refinement of the errors is unnecessary.

3.4. Validation of the isolation scores by simulation and real data

3.4.1. Random simulation of false negative and false positive errors

To test the efficiency of the various scores we simulated spike-sorting errors and calculated the isolation and the classification error scores (Fig. 6). The error simulation was carried out by modifying well-sorted data (original isolation quality >0.99, less than 1% false errors) of four real-time detected GPe neurons with different signal-to-noise ratios (Fig. 6a). To validate the quality of the real-time sorting of the selected units we further examined the data using the off-line PCA method and also checked for inconsistency by screening of the analog signal.

To simulate false negative errors we eliminated spike events from the spike cluster and marked them as noise events (Fig. 6b). The independent variable was the ratio between the number of eliminated (i.e. missed) spikes and the real number of spikes (false negative ratio). Zero means no false negatives were generated and 1 means all spikes are classified as noise events. As expected the isolation score was close to 1 when the ratio of missed events was 0, and dropped to 0.5 when the missed ratio was 0.5 (Fig. 6b1). We conclude that the isolation score decreases linearly with the ratio of missed spikes when they are equally distributed. Moreover, the scores of the four different units were highly correlated ($R^2 > 0.99$). This demonstrates the consistency of the isolation score; i.e. the same ratio of errors yields the same isolation score.

The estimated false negative score was a good estimation of the simulated false negative ratio values between 0 and approximately 0.35 (Fig. 6b2). When the fraction of simulated errors was above 0.35 the estimation of the false negative was noisy, and it fluctuated around 0.35. The false positive score was a valid estimator when the fraction of simulated false negative errors was between 0 and 0.3 (Fig. 6b3); in this range the estimate rate of the false positive errors was close to zero, as expected. However, a ratio of 0.3 or more of simulated false negatives caused the estimate of the false positive error to erroneously increase. By contrast to isolation and classification scores the SNR_{SPK} does not change as a function of the simulated false negative ratio (Fig. 6b4).

To simulate false positive errors we added events from the noise cluster to the spike cluster (Fig. 6c). The independent variable was the ratio between the number of noise events included

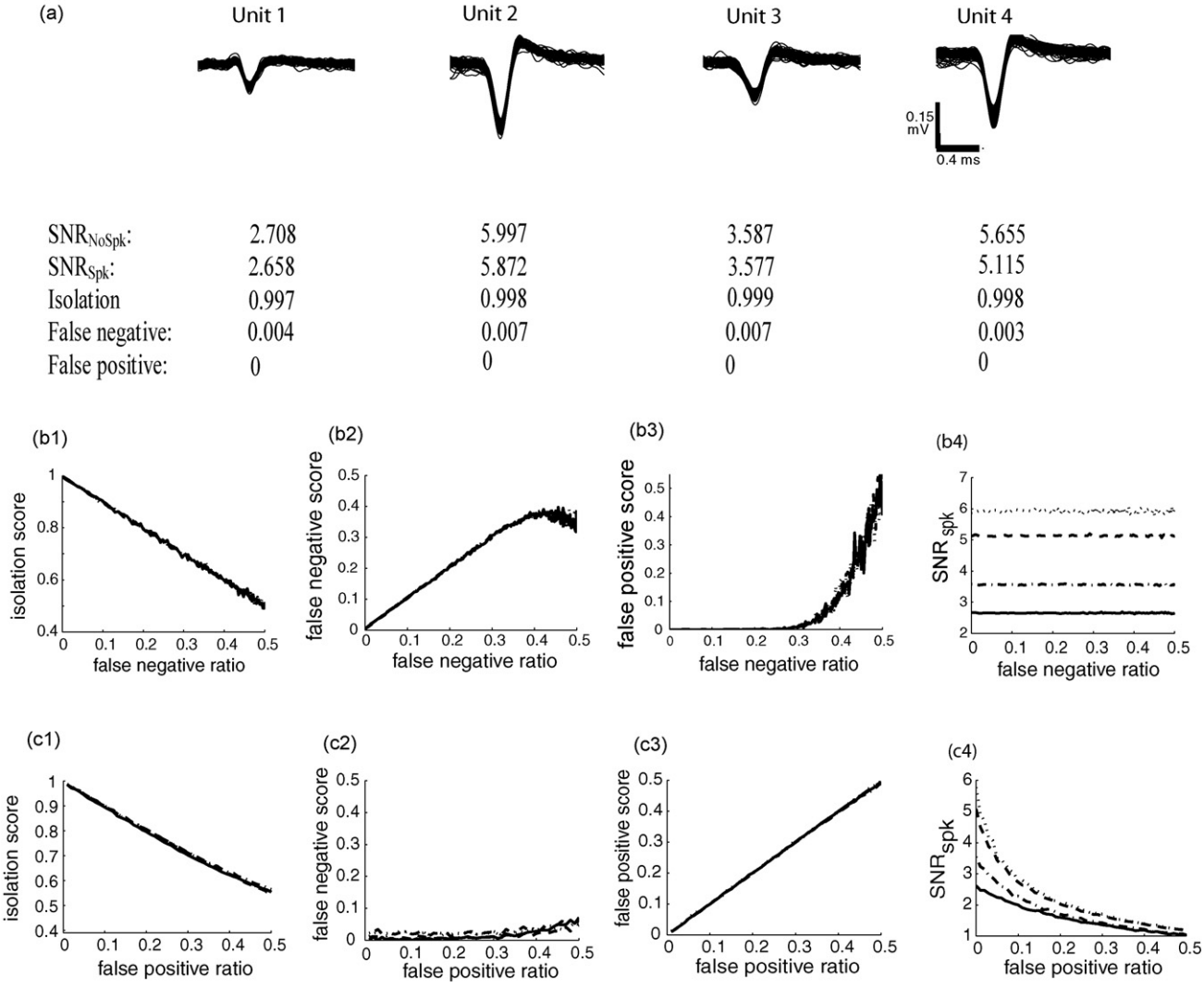

Fig. 6. Simulation of false positive and negative errors. False negatives were simulated by random reclassification of spike events as noise events. The independent variable is the ratio between the number of false negatives introduced and the size of the original spike cluster. The false positive errors were simulated by setting the noise cluster to be 10 times the size of the spike cluster and reclassifying noise events as spike events. The independent variable is the ratio of the number of false positives to the size of the spike cluster after reclassifying. (a) Spike waveforms from four well-sorted units with different signal-to-noise ratios. (b) Simulation of false negative errors. (b1) Isolation score. The score decreases with the number of false negatives; the difference between the units is negligible. (b2) False negative score. The score predicts the ratio of simulated false negatives well when the fraction of misclassified units is below 0.3. In this range, the difference between the score and the error ratio is less than 0.02. For larger simulated error ratios the score is misleading; instead of increasing, the score is bounded by 0.45. (b3) False positive score. The score predicts the false positive errors well when the fraction of misclassified units (simulated false negative) is below 0.3. For large error ratios of simulated false negatives the false positive score is misleading; instead of remaining at zero the score rises to 0.5. (b4) SNR_{SPK}. The SNR_{SPK} does not change as a function of the false negatives. (c) Simulation of false positive errors. (c1) Isolation score. The score decreases with the number of simulated false positive errors; no significant difference was found between the different units. (c2) False negative score. When the ratio of simulated false positive errors is larger than 0.3 the score increases from 0.01 to 0.08 due to biases when the noise and the spike cluster overlap. (c3) False positive score. The score follows the ratio of simulated errors (less the 0.02 difference). (c4) SNR_{SPK}. The SNR_{SPK} decreases with the number of false positives; however the SNR_{spk} of different units is not consistently modified by the fraction of simulated false positives.

in the spike cluster and the size of the spike cluster (false positive ratio). Zero means no false positives were generated and 0.5 means the number of simulated false positives was equal to the number of real spikes in the spike cluster. Unlike the case of simulated false negative errors, the reduction in size of the noise cluster may influence this simulation. To minimize such effects we set the noise cluster to be 10 times the size of the spike cluster before generating the errors. The isolation score was close to 1 when the error ratio was 0 and decreased to 0.55 when the simulated error ratio was 0.5 (Fig. 6c1). The changes in the isolation score as a function of the simulated false positive error were highly correlated ($R^2 > 0.99$) for the four different units depicted in Fig. 6. The false positive score was a good estimation of the ratio of errors; the difference between the score and the ratio of the simulated errors was less than 0.02 (Fig. 6c3). The false negative score changed only slightly when the simulated false positive error ratio was less than 0.3; when the error ratio increased the false negative score increased to 0.08 (Fig. 6c2). The SNR_{SPK} decreased with the number of false positives, however the effect on the different units was not consistent; i.e. the

ratio of simulated false positives had different effects on the SNR_{spk} of the four tested units (Fig. 6c4).

In conclusion, the error simulations verify that the isolation score can be a measure of the extent to which noise events and spike events overlap. The simulation results also emphasize the fact that the classification error scores have a range of good predictability that is dependent on the overlap between clusters and hence on the isolation score. In this range of good predictability (e.g. for isolation scores >0.70) the false positive and negative scores should be used as a refinement of the isolation score. The results also demonstrate that SNR is misleading; it does not follow the false negatives ratio nor does it have a scale in which different units with the same ratio of false positives have the same score.

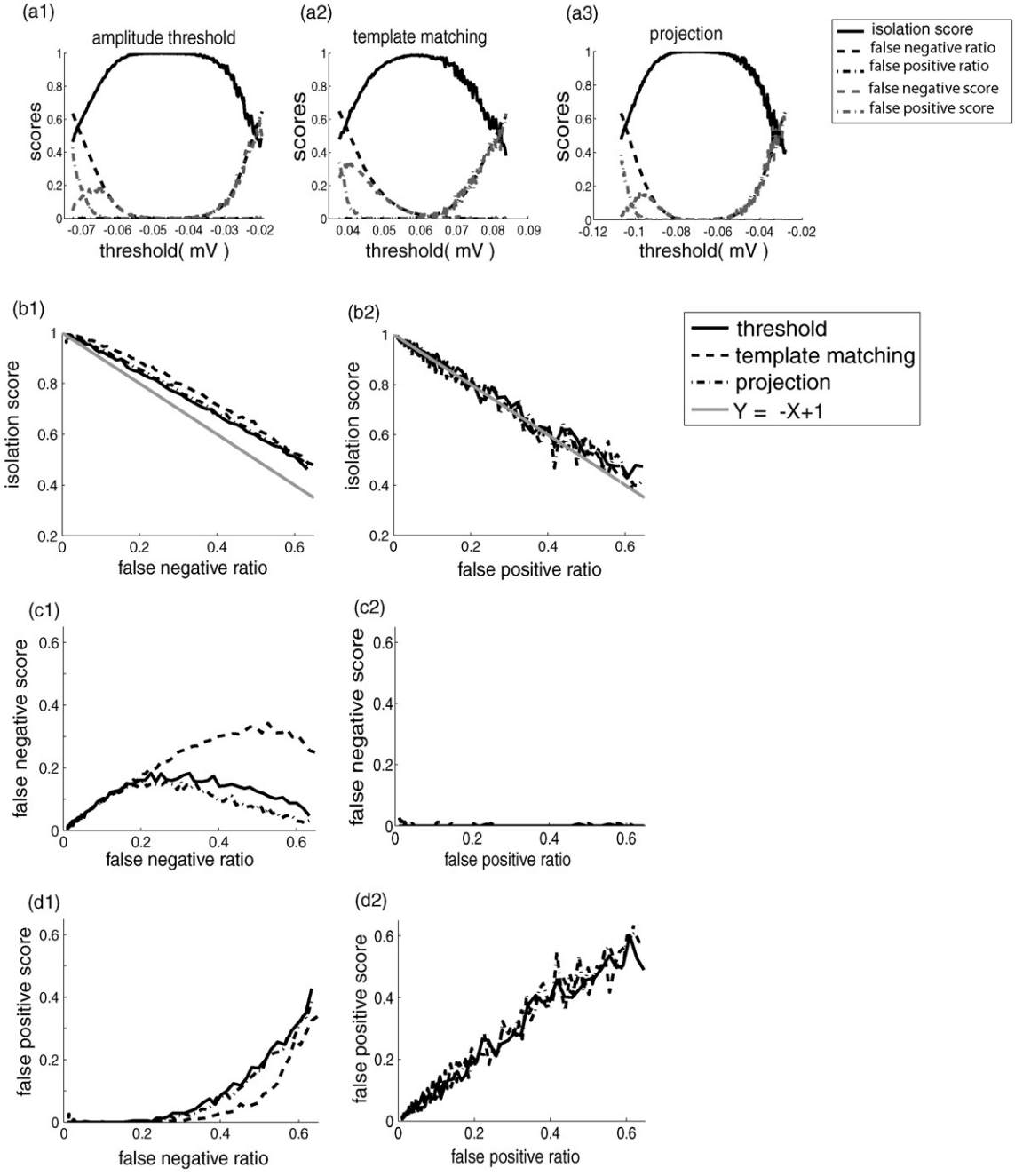

Fig. 7. Effects of different sorting algorithms on the isolation scores. We generated sorting errors using three different sorting methods on the data of unit 1 of Fig. 6. (a) The scores and the real ratio of classification errors as a function of the criterion used in the sorting method. (a1) Amplitude threshold crossing classification. (a2) Template matching algorithm. We used a training set of 200 spikes to generate a 12-point template. We then calculated the distance of all events to this template and applied a threshold to classify an event as a spike. (a3) Projection on the average template. Similar to the template matching method we projected the data on the average template (1.5 ms, 36 points) defined by a training set of 200 spikes. (b) The isolation score as a function of the ratio of real false positive and negative errors. The different lines are the scores given when using different sorting methods. The isolation score was consistent across the sorting methods. (c) False negative score as a function of the real ratio of errors. (d) False positive score as a function of the ratio of errors.

3.4.2. Sorting errors and the isolation scores

To further validate our scores under different sorting conditions we simulated clustering errors using different sorting methods. From the four well-isolated units in Section 3.4.1 we took the one with the lowest signal-to-noise ratio (Fig. 6, unit 1) and re-clustered the continuous sampled analog data using three clustering methods: (1) threshold crossing—all events that pass a threshold are marked as spikes. (2) Template matching—we used a training set to calculate the average spike waveform and then implemented an off-line algorithm similar to our real-time eight-point template matching algorithm (Section 2.1.1). Waveforms that were similar to the average of the training set have small Euclidian distances from the template and were considered as spikes from the same unit. (3) Projecting the data on the average template. The average template was generated using a training set. We then normalized this template to have a norm of 1 and convoluted it with the analog data. Spikes were detected as peaks in the resulting vector. This method is equivalent to the projection on the first principal component when the data contains only one unit (Abeles and Goldstein, 1977).

Each of these methods classifies events as spikes or noise by a user-defined threshold. Here we used this threshold as our independent variable and examined its effect on the isolation scores. For each threshold we estimated the real ratio of false positive and negative errors (assuming that the original classification represented the real classification) and calculated the isolation score and error classification scores. As with the random simulation of errors (e.g. random switching of spike and noise events; Section 3.4.1 and Fig. 6) the isolation scores decreased with modifications of the sorting thresholds that increased the number of classification errors. This decrease took place both when the threshold values were very conservative and led to false negative errors (Fig. 7a, left side of plots) and when the thresholds were too permissive and led to false positive errors (Fig. 7a, right side of plots). The error classification scores followed the real error ratio when it was small but suffered from biases for large real error ratios and low isolation scores. To check the dependency of the scores on the clustering method we compared the scores with the real ratio of false negatives and the ratio of false positives (Fig. 7b–d). We found that the isolation score differed slightly between the tested methods (Fig. 7b). However, when comparing these isolation scores to the isolation score obtained when errors were simulated randomly (Fig. 6) we found that for a given number of false negatives the isolation score was higher when we used different threshold levels. This over-estimation of the isolation score was probably due to the local consistency of errors induced by systemic modification of the thresholds in the sorting clustering methods. The false negative score had a range in which it is equal to the real false negative ratio (Fig. 7c). This range was larger when using template matching than when using the other sorting methods. Similarly, the false positive score (Fig. 7d) was equal to the false positive ratio when such errors existed and was biased when the number of false negative was large. This bias was smallest for the template matching algorithm. Nevertheless, as with the random simulation of errors, systemic modification of the thresholds by several sorting methods reveals that the isolation quality is a consistent and reliable estimator of the quality of the spike clustering. The classification errors can be used in cases with high levels of isolations scores (>0.8) and small levels of false positive and negative errors (<0.25).

3.4.3. Dynamic and population analysis of the isolation scores

Typical physiological experiments include long duration (>15 min) recordings of the same units. Naturally, the isolation quality may drift or change over these periods. The isolation quality tests were applied to real data recorded for periods of more than 10 min. Each recording was split into segments of 60 s (∼1000–4000 spikes in our GP data). To limit the algorithm complexity (time and place) we reduced (by random pruning) the largest cluster to a size of 1500 spikes; the other cluster was then reduced to maintain the size ratio between clusters. The length of the segment is thus a tradeoff between computational time versus effectiveness. When using a short segment the sampling of the spike and noise cluster will be more accurate due to non-stationarity and less random pruning; however the computational time will increase. After extracting these clusters they were scored as described in the previous sections. Thus, for each unit we obtained a series of scores. These series of scores can be examined for problematic recording epochs which should be scrutinized more carefully (by re-clustering or omitting these sessions). This is depicted in Fig. 8 where 43 min of consecutive real-time sorting were scored. After 35 min of recording, it can be seen that the quality of the sorting decreased rapidly despite the apparent increase in the SNR of the unit. Our recommendation is therefore to apply these tests to any prolonged extracellular recording, and then to exclude periods with low scores from the analysis database. As a rule of thumb we suggest excluding recording periods with an isolation score below 0.7–0.8.

To achieve a single score for each unit we averaged the scores over all sessions. The average scores of the 155 GPe units in our database were: isolation score, 0.93 ± 0.08; false negative score, 0.1 ± 0.09; false positive score, 0.02 ± 0.04 (Fig. 8c). To compare the scores from different brain areas we calculated the scores of 87 units recorded in the primary motor cortex. Action potentials from the cortex were wider than GPe waveforms. Hence, we used 2 ms of analog recordings for each action potential. The average isolation score of these cells was 0.79 ± 0.17. The average false negative score was 0.09 ± 0.19 and the average false positive score was 0.13 ± 0.18 (Fig. 8d). All distributions were significantly different from GP scores ($p < 10^{-3}$ Kolmogorov–Smirnov test). This difference in scores is consistent with our subjective sense of the better quality of the GP data and is probably due to the difference in cell sizes and cell density in these brain areas.

3.4.4. Exploring parameter space of the isolation and classification error scores

The isolation score was designed to be insensitive to the noise cluster size; i.e. adding events to the noise cluster that are far from the spike cluster should not affect the score. The size of the noise cluster is determined by the level of the amplitude threshold used for extracting the noise cluster. To verify this insensitivity we

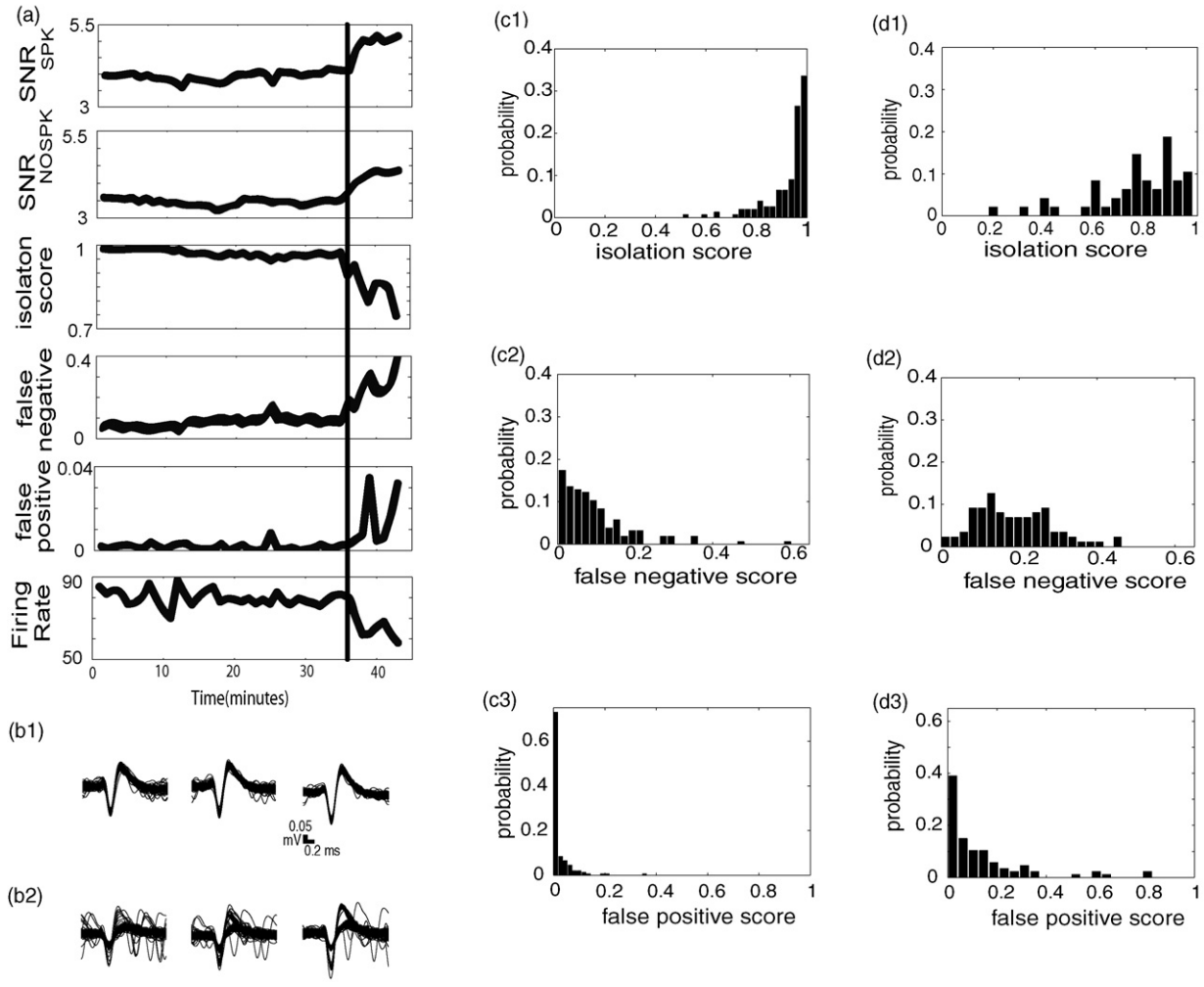

Fig. 8. Score statistics on real data. (a) Dynamic changes of the isolation scores. Data were split into segments of 60 s. For each segment we calculated the isolation and classification errors scores. (a) All five scores and spike rates were computed for 43 consecutive minutes. Although all scores were stable during the first 35 min, from the 36th minute they began to change. Both SNR scores increased; in contrast, the isolation score (which was extremely stable) rapidly decreased. Therefore, in this case the SNR scores are misleading and the isolation score indicates the moment when the quality of the sorting decreased. (b) Spike (b1) and noise (b2) events ($n = 100$, randomly selected) from 6 min of recording. The right column is from the first 6 min, the middle column from minute 19 to 25 and the left from the last 6 min of recording. In the last 6 min many noise waveforms resemble spikes. This misclassification is probably due to a slight modification in the spike waveform (reflected by the SNR) that was not identified by the semi-automatic template matching algorithm. (c) Distributions of the scores of 155 GPe units. Scores from different sessions were averaged. (c1) Isolation score. (c2) False negative score. (c3) False positive score. (d) Distribution of scores of 87 cortex units. (d1) Isolation score. (d2) False negative score. (d3) False positive score.

modified the size of the noise cluster by changing the fraction of events from the spike cluster used to calculate the threshold (these were the low amplitude spikes, hence fewer spikes means a closer to zero threshold). The distribution of the isolation score was independent of the threshold used for noise cluster extraction ($p > 0.86$ one-way ANOVA, $p > 0.79$ Kruskal–Wallis non-parametric ANOVA).We then compared the isolation score of all GPe units ($n = 155$) and found that the scores calculated with different noise clusters were highly correlated (Fig. 9a). Hence, our methods are insensitive to the size of the noise cluster. As described above, in order to reduce computation time we used a random sample from the spike and noise clusters. As a result each time we calculated the isolation score we used different events. The fact that we obtained the same scores when using different random samples from the same distribution further demonstrates the stability of our method.

The isolation score depends on the λ parameter that sets the gain of the distance stretch. To check the dependency of the isolation score on this parameter we modified this parameter and calculated the isolation score of all 155 GPe units (Fig. 9b1). When λ was larger than 5 the isolation score was highly correlated with the scores calculated with our default value of $\lambda = 10$ ($R > 0.946$). However, when λ was equal to 1 the scores were not as highly correlated ($R = 0.72$). This is expected since small values of λ mean that the Euclidian distance between events is not stretched, and therefore distant events influence the score. To further investigate the influence of λ on our scores we simulated classification errors by applying different thresholds when sorting the data (as described in Section 3.4.2). We modified λ and calculated the isolation score as a function of the false negatives and positives ratio (Fig. 9b2–3). We found that when λ is small the score over-estimates the number of false positives (Fig. 9b3).

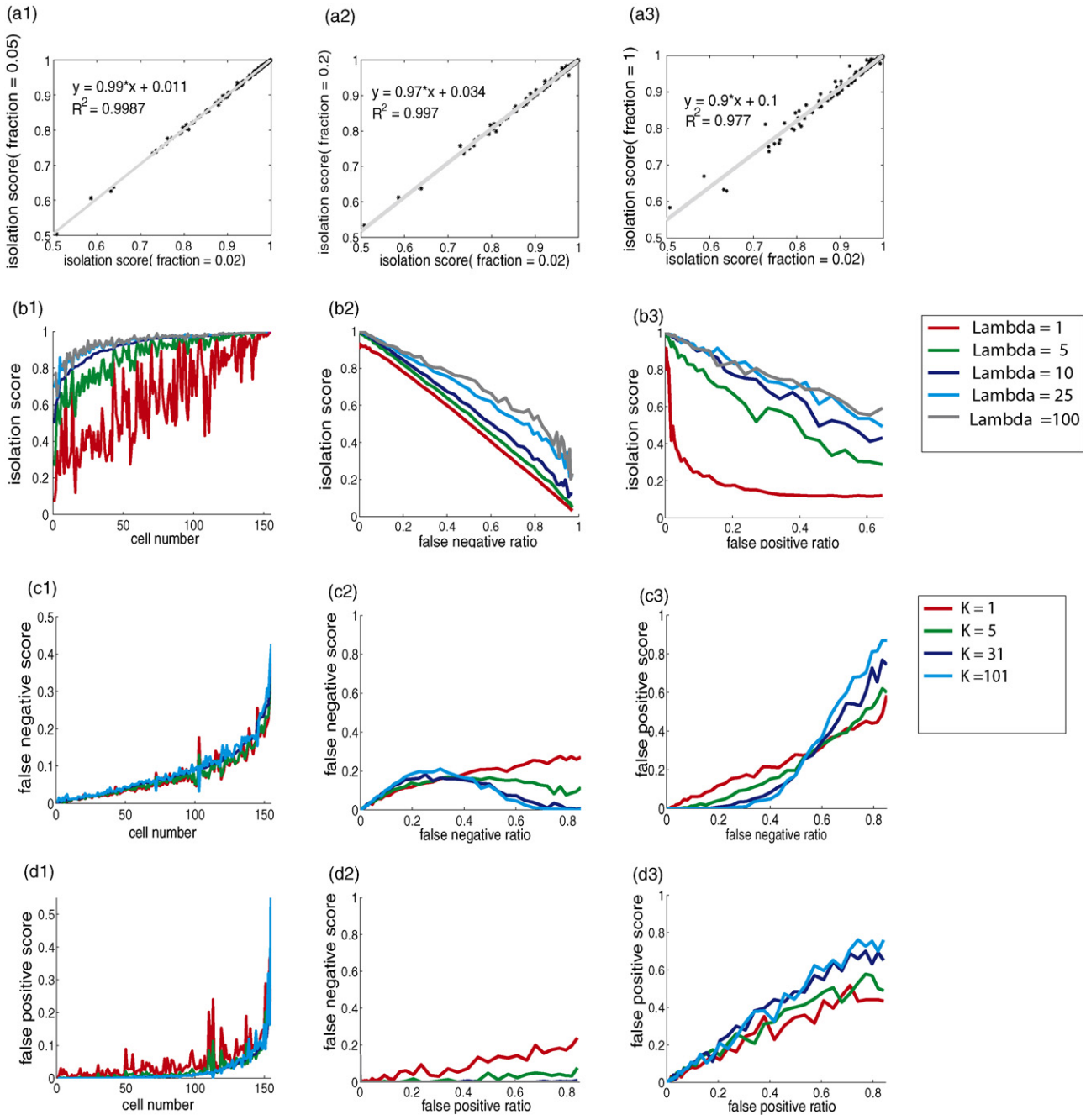

Fig. 9. Investigating parameter space of the isolation and classification-error scores. We modified the parameters used for calculating the score and compared the scores of real units and the scores of a unit with errors simulated by re-clustering the data using a threshold crossing method. (a) We modified the fraction of the spike clusters we used to calculate the threshold (these were the low amplitude spikes; hence fewer spikes means a closer to zero threshold) and compared the isolation scores when using 2% of the spike cluster. (a1) 2% vs. 5%. (a2) 2% vs. 20%. (a3) 2% vs. 100%. (b) Comparison of isolation score when modifying λ. (b1) Real data results. Units were sorted by the isolation score when using $\lambda = 10$. (b2) Isolation score as a function of the ratio of simulated false negatives. (b3) Isolation score as a function of the ratio of simulated false positives. (c) Comparison of false negative score when modifying K. (c1) Real data results. Units were sorted by the false negative score when using $K = 31$. (c2) False negative score as a function of the ratio of false negatives. (c3) False negative score as a function of false positives. (d) Same as (c) for the false positive score.

In addition we found that as λ increases the false negative score tends to increase. However, this increase is bounded. We conclude that our selection of $\lambda = 10$ does not suffer from biases that occur when λ is small and it is within the large range in which the isolation score follows the ratio of classification errors.

The KNN algorithm used for the calculation of the classification error scores depends on the K we use. We modified this parameter and calculated the scores of all 155 GPe units (Fig. 9c1 and d1). The units were sorted by the scores when using the default value of $K = 31$. The false negative score changed only slightly when modifying K (Fig. 9c1); on the other hand the false positive score was sensitive to the K used (Fig. 9d1). We simulated clustering errors (as we simulated errors when modifying λ) and calculated the error classification score as a function of

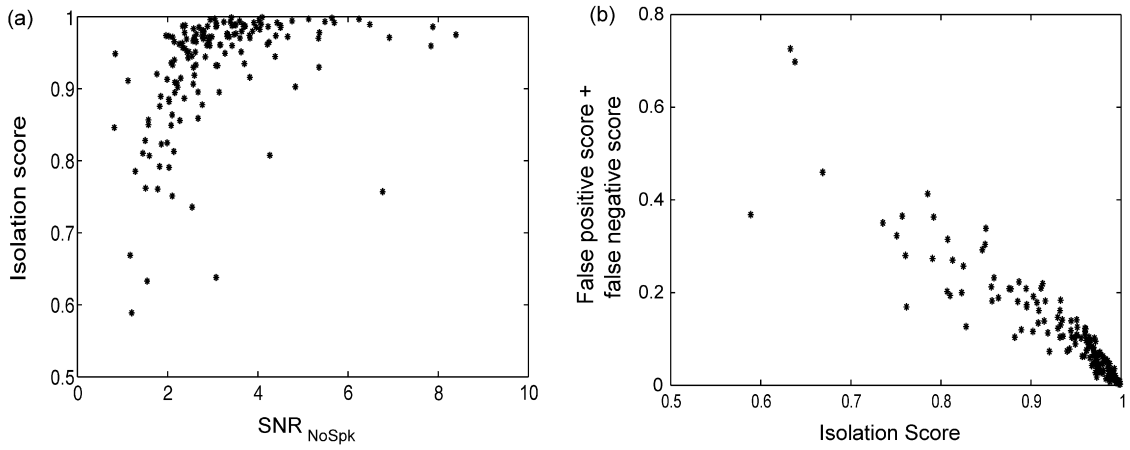

Fig. 10. Comparing the scores. The scores were compared using the data from 155 GP units from three different monkeys. (a) Isolation score vs. $SNR_{No\ Spk}$. When the SNR is small the isolation score tends to be small, and when the SNR is large the isolation score is usually close to 1. But this relation is not linear. Therefore, any SNR threshold will either include units that are poorly isolated (low isolation score) or exclude units that are well isolated (high isolation score). Furthermore, the outliers (high SNR with low isolation score) reveal the weaknesses of the spike sorting process. (b) False positive + false negative scores vs. isolation score. As the isolation score decreases the variability increases. Hence, the error type scores should only be used when the isolation score is high.

error ratio for different K values. We found that although K can bias the scores when the error ratios is large, there is a range of good predictability.

3.4.5. Comparing the scores

Our SNR, isolation and classification error scores were not designed to be independent. To determine the degree of dependency we compared the different scores (Figs. 3d and 10). First we compared the two SNR scores (Fig. 3d). The underlying reasons for the differences between these scores were described in Section 3.1.2; nevertheless we found the SNR scores to be highly correlated ($R^2 = 0.94$).

We then compared the isolation score and the SNR score and found that as expected, in most cases units with a high SNR score had high isolation scores and units with low SNR scores had low isolation scores (Fig. 10a). On the other hand, the connection between these scores was non-linear and had outliers in which the isolation score was low although the SNR was high (e.g. last 6 min of Fig. 8). Due to these properties, any exclusion/inclusion criteria of units using a threshold based on the SNR scores will lead either to inclusion of units with a low isolation score or to exclusion of units with a high isolation score. Finally we compared the isolation score and the sum of false positive and negative scores (Fig. 10b) and found that when the isolation score was high, the variability of the sum was low and when the score was low the variability of the sum of the classification error scores was high. This again shows that the classification error scores are a refinement of the isolation score only when it is high, and further that when the isolation score is low these scores are less reliable and should not be used.

4. Discussion

We quantified the quality of spike detection and sorting using signal-to-noise ratios (SNR), isolation scores, and classification error scores. We then simulated errors for validating the scores, compared several spike sorting algorithms, and investigated the parameter space of the scores.

4.1. Related studies

Some previous studies have quantified the quality of clustering of recordings from multi-channel electrodes. These methods can be adapted to single channel recordings. In their study, Pouzat et al. (2002) assumed a Gaussian distribution of the noise, which they used to evaluate the variability of the spike waveforms. Shoham et al. (2003) have argued that the Gaussian assumption is inaccurate, and that the t-distribution is a better fit for the data. Furthermore, the distribution of noise, in general, is not sufficient for estimating the variability of signal statistics (Fee et al., 1996b). Even modeling the variability generated by the cells' intrinsic properties is not always sufficient because it does not predict the variability caused by changes in the relative position of electrodes and recorded neurons (Fig. 1). Schmitzer-Torbert et al. (2005) used the χ^2 distribution as a distance measure of the noise events from the spike cluster in a feature space. In their method the distance between a noise event and the spike cluster was treated in a global manner; i.e. the score of each noise event depended on its distance from the center of the spike cluster. By contrast, our scores are based on the local properties of the spike cluster. While their approach focused on the contribution of the noise events, our scores iterate over the events in the spike cluster. As a result of these differences, our isolation score captures phenomena found in non-homogenous spike clusters (i.e. clusters containing false positives), which the χ^2 distance does not. In addition, we can obtain an estimate of the number of false positive and false negative errors that is not available with previous methods. Harris et al. (2001) and Schmitzer-Torbert et al. (2005) introduced the *isolation distance* which quantifies the quality of clustering by the minimal distance where the number of spike events and noise events is equal. Although this score is "self consistent"; i.e. the score will

decrease in the same recording with a reduction of the quality of the sorting, it does not have a global scale to differentiate between well and poorly isolated units. For example, a well-isolated unit with a low SNR can have the same score as a poorly isolated cluster with a high SNR. A major advantage of our isolation score is its intuitive range of zero to one, which enables easy comparison of units recorded at different times, and even by different research groups.

In summary, we propose the isolation score, which is a measure of the separation between two groups (clusters); and then we present the two classification scores using a "one-class classification problem" approach. There are few other metrics for group separation (usually as evaluations of clustering techniques) and classification problems (Trevor et al., 2001), e.g. metrics based on Euclidian distance, city block, etc. However, we feel that the isolation and classification scores provide better metrics for spike data due to their insensitivity to noise cluster size.

4.2. Relationship between scores

The scores show inter-dependence. A low isolation score is likely when the SNR is low, because low recording quality leads to cluster errors. On the other hand, large SNR values that appear with low isolation scores indicate problems with the clustering algorithm. There are many possible reasons for such isolation failure. These include assumptions in the clustering algorithm that may not have been fulfilled; e.g. the statistical model was wrong, the data were non-stationary or human errors were made. In this case (high SNR, low isolation score) we suggest re-clustering the data.

4.3. The score under different conditions

Our simulation of spike errors using different sorting algorithms has shown that under different sorting conditions the scores are consistent and follow the number of simulated classification errors. We showed that the isolation score decreases with the ratio of classification errors and the classification error scores have a range in which they follow the real error ratio. However, applying sorting algorithms that directly reduce the scores may lead to a bias; i.e. a high isolation score despite a large ratio of classification errors. Nonetheless such an algorithm requires local consistency of spike clusters. Hence, we suggest using our scores when the sorting algorithms are based on global parameters (such as template matching and PCA based methods). Furthermore we suggest that local consistency algorithms should be used for post processing of sorting algorithms (see below).

We compared the scores of two different brain areas. To enable this comparison we adjusted the time interval slightly for representation of events. We found that the isolation scores of GPe units were significantly larger than cortex units. This was consistent with our subjective impression that GPe units were better isolated. A major difference between GP and cortical recordings is that GP recordings are usually of only one cell per electrode, whereas two to three units are typically recorded by a single cortical electrode. Our isolation score methods do not distinguish between recordings of several cells on one electrode versus single cell recordings. In both cases given a cluster of spikes we extract the noise cluster and calculate our scores. As a result our scores reflect the quality of each unit and not the overall quality of the all units recorded from a given electrode.

A preliminary condition for quality assessment is the insensitivity of the isolation score to the exact size of the noise cluster. By using different thresholds for extracting the noise cluster we showed that once the noise cluster contains the events that are close to the spike cluster, the score depends only slightly on the exact size of the noise cluster. As a result our methods can be applied to systems with intermittent sampling conditioned by the extraction and analog sampling only of putative spikes. In such systems it is possible to use other spike clusters, if they exist, such as the noise reference; however this may lead to over-estimation of data quality.

4.4. Future directions

To enhance the reliability of the results of studies based on extracellular recordings we suggest using the isolation score for preliminary analysis and exclusion of units or periods with very low isolation scores from the study data-base. We suggest that the findings be first verified on the recordings with high isolation scores and then extended to the entire data base. We suggest excluding units with isolation scores below 0.8 in studies whose conclusions may be influenced by the isolation quality of the recorded units. However, we believe that more testing is needed for setting this threshold and hope that such a threshold will emerge after future work is done in different recording settings and neuronal areas. In any case, this should not limit the report of the isolation score even when it is not used as a criterion for excluding data.

An additional benefit of classification error scores is that they identify likely misclassified events. Our KNN approach can be used as a post-processing tool to optimize the original spike sorting, by flipping the classification for these missed events. An even more promising approach would be to use our isolation score algorithm to recluster these missed events, by using the $P(X)$ values (Fig. 4). Recall that this value is akin to the probability that event X belongs to the spike cluster. The re-clustering could simply flip the classification for events X, for which their $P(X)$ value is greater than some threshold.

In this study we did not attempt to develop a method for finding the optimal value of K in the K nearest neighbors approach, but only constrained it. A data-driven approach, where K depends on various parameters of the spike and noise clusters (e.g. number of elements, overlap of the two clusters as measured by the isolation score, etc.) may be pursued. One may consider using two values for K, one for detecting false negatives and one for detecting false positives.

Finally, our methods are based on analyzing the waveform of extracellular events and did not take spike train properties into account such as the firing rate or refractory periods. These properties are valuable for assessing spike sorting quality and thus can be used independently or could be incorporated into our

scores. For example, we could introduce a progressive penalty for units with detected spikes in their estimated refractory period.

In summary, we have developed methods for quantifying the isolation quality of extracellularly recorded action potentials and compared these different methods. The scoring methods were applied directly to the spike waveform; however they may be used on other representations of the spike, e.g. PCA or wavelet-based representations. Isolation quality quantifications are a necessary step in interpreting studies based on extracellular recording. The conclusions of many single-units studies are more dependent on their unit isolation quality than on the power of the statistical and analytical methods used for their spike-train analysis. Nevertheless, in most cases, objective criteria are used and reported for the later stage but not for the first stages of the data acquisition process. We encourage research groups to use isolation measures, as developed in this manuscript, rather than more commonly used phrases such as "only well-isolated units were included in our study".

Acknowledgement

This study was partly supported by a Center of Excellence grant administered by the ISF and HUNA's "Fighting against Parkinson" grant.

References

Abeles M, Goldstein MHJ. Multispike train analysis. IEEE Trans Biomed Eng 1977;65:762–73.

Bar-Gad I, Ritov Y, Bergman H. Failure in identification of overlappig spikes from multiple neuron activity causes artificial correlations. J Neurosci Methods 2001;107:1–13.

Bergman H, DeLong MR. A personal computer-based spike detector and sorter: implementation and evaluation. J Neurosci Methods 1992;41:187–97.

Elias S, Joshua M, Goldberg JA, Heimer G, Arkadir D, Morris G, et al. Statistical properties of pauses of the high-frequency discharge neurons in the external segment of the globus pallidus. J Neurosci 2007;27:2525–38.

Fee MS, Mitra PP, Kleinfeld D. Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability. J Neurosci Methods 1996a;69:175–88.

Fee MS, Mitra PP, Kleinfeld D. Variability of extracellular spike waveforms of cortical neurons. J Neurophysiol 1996b;76:3823–33.

Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighbourhood component analysis. Neural Inform Process Syst (NIPS'04) 2004;17:513–20.

Harris KD, Hirase H, Leinekugel X, Henze DA, Buzsaki G. Temporal interaction between single spikes and complex spike bursts in hippocampal pyramidal cells. Neuron 2001;32:141–9.

Heimer G, Bar-Gad I, Goldberg JA, Bergman H. Dopamine replacement therapy reverses abnormal synchronization of pallidal neurons in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine primate model of parkinsonism. J Neurosci 2002;22:7850–5.

Lewicki MS. Bayesian modeling and classification of neural signals. Neural Comp 1994;6:1005–30.

Lewicki MS. A review of methods for spike sorting: the detection and classification of neural action potentials. Network 1998;9:R53–78.

Likhtik E, Pelletier JG, Paz R, Pare D. Prefrontal control of the amygdala. J Neurosci 2005;25:7429–37.

Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 2004;43:133–43.

Nenadic Z, Burdick JW. Spike detection using the continuous wavelet transform. IEEE Trans Biomed Eng 2005;52:74–87.

Pare D, Gaudreau H. Projection cells and interneurons of the lateral and basolateral amygdala: distinct firing patterns and differential relation to theta and delta rhythms in conscious cats. J Neurosci 1996;16:3334–50.

Pouzat C, Delescluse M, Viot P, Diebolt J. Improved spike-sorting by modeling firing statistics and burst-dependent spike amplitude attenuation: a Markov chain Monte Carlo approach. J Neurophysiol 2004;91:2910–28.

Pouzat C, Mazor O, Laurent G. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. J Neurosci Methods 2002;122:43–57.

Quiroga RQ, Nadasdy Z, Ben Shaul Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 2004;16:1661–87.

Schmitzer-Torbert N, Jackson J, Henze D, Harris K, Redish AD. Quantitative measures of cluster quality for use in extracellular recordings. Neuroscience 2005;131:1–11.

Shoham S, Fellows MR, Normann RA. Robust, automatic spike sorting using mixtures of multivariate *t*-distributions. J Neurosci Methods 2003;127:111–22.

Trevor H, Robert T, Jerome F. The elements of statistical learning: data mining, inference and prediction. New York: Springer Verlag; 2001.

Vapnik VN. Statistical learning theory. New York: Wiley; 1998.

Wood F, Black MJ, Vargas-Irwin C, Fellows M, Donoghue JP. On the variability of manual spike sorting. IEEE Trans Biomed Eng 2004;51:912–8.

Worgotter F, Daunicht WJ, Eckmiller R. An on-line spike form discriminator for extracellular recordings based on an analog correlation technique. J Neurosci Methods 1986;17:141–51.