Journal of Neuroscience Methods 288 (2017) 82–98

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Model-based spike sorting with a mixture of drifting t-distributions

Kevin Q. Shan[a,b], Evgueniy V. Lubenov[a,b], Athanassios G. Siapas[a,b,*]

[a] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, United States
[b] Division of Engineering and Applied Science, California Institute of Technology, Pasadena, United States

HIGHLIGHTS

- A nonstationary generative model for spike sorting is proposed.
- This model tracks unit drift in chronic recordings and is robust to outliers.
- It offers improved estimates of single unit isolation in empirical data.
- An efficient software implementation is provided for fitting the model.

ARTICLE INFO

Article history:
Received 27 January 2017
Received in revised form 16 June 2017
Accepted 20 June 2017
Available online 23 June 2017

Keywords:
Clustering
Heavy tails
Chronic recording
Spike overlap
Cluster drift
Unit isolation metrics

ABSTRACT

Background: Chronic extracellular recordings are a powerful tool for systems neuroscience, but spike sorting remains a challenge. A common approach is to fit a generative model, such as a mixture of Gaussians, to the observed spike data. Even if non-parametric methods are used for spike sorting, such generative models provide a quantitative measure of unit isolation quality, which is crucial for subsequent interpretation of the sorted spike trains.
New method: We present a spike sorting strategy that models the data as a mixture of drifting t-distributions. This model captures two important features of chronic extracellular recordings—cluster drift over time and heavy tails in the distribution of spikes—and offers improved robustness to outliers.
Results: We evaluate this model on several thousand hours of chronic tetrode recordings and show that it fits the empirical data substantially better than a mixture of Gaussians. We also provide a software implementation that can re-fit long datasets in a few seconds, enabling interactive clustering of chronic recordings.
Comparison with existing methods: We identify three common failure modes of spike sorting methods that assume stationarity and evaluate their impact given the empirically-observed cluster drift in chronic recordings. Using hybrid ground truth datasets, we also demonstrate that our model-based estimate of misclassification error is more accurate than previous unit isolation metrics.
Conclusions: The mixture of drifting t-distributions model enables efficient spike sorting of long datasets and provides an accurate measure of unit isolation quality over a wide range of conditions.

1. Introduction

Chronic extracellular recordings offer access to the spiking activity of neurons over the course of days or even months. However, the analysis of extracellular data requires a process known as spike sorting, in which extracellular spikes are detected and assigned to putative sources. Despite many decades of development, there is no universally-applicable spike sorting algorithm that performs best in all situations.

Approaches to spike sorting can be divided into two categories: model-based and non-model-based (or non-parametric). In the model-based approach, one constructs a generative model (e.g. a mixture of Gaussian distributions) that describes the probability distribution of spikes from each putative source. This model may be used for spike sorting by comparing the posterior probability that a spike was generated by each source. Fitting of such models may be partially or fully automated using maximum likelihood or Bayesian methods, and the model also provides an estimate of the misclassification error.

[*] Corresponding author at: Caltech MC 139–74, 1200 E California Blvd, Pasadena, CA 91125, United States.
E-mail address: thanos@caltech.edu (A.G. Siapas).

http://dx.doi.org/10.1016/j.jneumeth.2017.06.017
0165-0270/

In the non-parametric approach, spike sorting is treated solely as a classification problem. These classification methods may range from manual cluster cutting to a variety of unsupervised learning algorithms. Regardless of the method used, scientific interpretation of the sorted spike train still requires reliable, quantitative measures of unit isolation quality. Often, these heuristics either explicitly (Hill et al., 2011) or implicitly (Schmitzer-Torbert et al., 2005) assume that the spike distribution follows a mixture of Gaussian distributions.

However, a mixture of Gaussians does not adequately model the cluster drift and heavy tails that are observed in experimental data (Fig. 1). Cluster drift is a slow change in the shape and amplitude of recorded waveforms (Fig. 1C), usually ascribed to motion of the recording electrodes relative to the neurons (Snider and Bonds, 1998; Lewicki, 1998). This effect may be small for short recordings (<1 h), but can produce substantial errors if not addressed in longer recordings (Fig. 7). Even in the absence of drift, spike residuals have heavier tails than expected from a Gaussian distribution, and may be better fit using a multivariate *t*-distribution (Figs. 1D and 6 ; see also Shoham et al., 2003; Pouzat et al., 2004).

To address these issues, we model the spike data as a mixture of drifting *t*-distributions (MoDT). This model builds upon previous work that separately addressed the issues of cluster drift (Calabrese and Paninski, 2011) and heavy tails (Shoham et al., 2003), and we have found the combination to be extremely powerful for modeling and analyzing experimental data. We also discuss the model's robustness to outliers, provide a software implementation of the fitting algorithm, and discuss some methods for reducing errors due to spike overlap.

We used the MoDT model to perform spike sorting on 34,850 tetrode-hours of chronic tetrode recordings (4.3 billion spikes) from the rat hippocampus, cortex, and cerebellum. Using these experimental data, we evaluate the assumptions of our model and provide recommended values for the model's user-defined parameters. We also analyze how the observed cluster drift may impact the performance of spike sorting methods that assume stationarity. Finally, we evaluate the accuracy of MoDT-based estimates of misclassification error and compare this to the performance of other popular unit isolation metrics in the presence of empirically-observed differences in firing rate and spike variability.

2. Methods

2.1. Mixture of drifting t-distributions (MoDT) model

Spike sorting begins with spike detection and feature extraction. During these preprocessing steps, spikes are detected as discrete events in the extracellular voltage trace and represented as points $\boldsymbol{y}_n$ in some *D*-dimensional feature space.

The standard mixture of Gaussians (MoG) model treats this spike data $\boldsymbol{y}_n$ as samples drawn from a mixture distribution with PDF given by

$$f_{\text{MoG}}\left(\boldsymbol{y}_n;\phi\right) \quad = \sum_{k=1}^{K} \alpha_k f_{\text{mvG}}\left(\boldsymbol{y}_n;\boldsymbol{\mu}_k,\boldsymbol{C}_k\right),$$

where $\phi = \{\ldots, \alpha_k, \boldsymbol{\mu}_k, \boldsymbol{C}_k, \ldots\}$ is the set of fitted parameters, *K* is the number of mixture components, α_k are the mixing proportions, and $f_{\text{mvG}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C})$ is the PDF of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$:

$$f_{\text{mvG}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C}) \quad = \frac{1}{(2\pi)^{D/2}|\mathbf{C}|^{1/2}} \exp\left[-\frac{1}{2}\delta^2(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C})\right].$$

Table 1
Mathematical notation. Lowercase bold letters ($\boldsymbol{y}_n$, $\boldsymbol{\mu}_{kt}$) denote *D*-dimensional vectors, and uppercase bold letters ($\boldsymbol{C}_k$, $\boldsymbol{Q}$) denote $D \times D$ symmetric positive definite matrices.

Dimensions	
D	Number of feature space dimensions
N	Number of spikes
K	Number of clusters
T	Number of time frames
Given data	
$\boldsymbol{y}_n$	Observed spike *n*
t_n	Time frame in which spike *n* occurred
w_n	Weighting of spike *n* (multiplier applied to log-likelihood)
User-defined constants	
ν	*t*-distribution degrees-of-freedom parameter
$\boldsymbol{Q}$	Drift regularization parameter
Fitted model parameters	
α_k	Mixing proportion for cluster *k*
$\boldsymbol{\mu}_{kt}$	Location parameter for cluster *k* in time frame *t*
$\boldsymbol{C}_k$	Scale parameter for cluster *k*
Latent variables introduced by EM procedure	
z_{nk}	Posterior probability that spike *n* belongs to cluster *k*
u_{nk}	Scaling variable introduced in formulating the *t*-distribution as a Gaussian-Gamma compound distribution

For notational convenience, let δ^2 denote the squared Mahalanobis distance

$$\delta^2\left(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C}\right) \quad = \left(\boldsymbol{y}-\boldsymbol{\mu}\right)^{\top}\boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right).$$

We make two changes to this model. First, we replace the multivariate Gaussian distribution with the multivariate *t*-distribution. The PDF for this distribution, parameterized by location $\boldsymbol{\mu}$, scale $\boldsymbol{C}$, and degrees-of-freedom ν, is given by

$$f_{\text{mvt}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C},\nu) = \frac{1}{(\nu\pi)^{D/2}|\mathbf{C}|^{1/2}}\frac{\Gamma((\nu+D)/2)}{\Gamma(\nu/2)}\left[1+\frac{1}{\nu}\delta^2(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{C})\right]^{-(\nu+D)/2}$$

Second, we break up the dataset into *T* time frames (we used a frame duration of 1 min) and allow the cluster location $\boldsymbol{\mu}$ to change over time. The mixture distribution becomes

$$f_{\text{MoDT}}\left(\boldsymbol{y}_n;\phi\right) \quad = \sum_{k=1}^{K} \alpha_k f_{\text{mvt}}\left(\boldsymbol{y}_n;\boldsymbol{\mu}_{kt_n},\boldsymbol{C}_k,\nu\right),$$

where $t_n \in \{1, \ldots, T\}$ denotes the time frame for spike *n*. We use a common ν parameter for all components and have chosen to treat it as a user-defined constant. The fitted parameter set is thus $\phi = \{\ldots, \alpha_k, \boldsymbol{\mu}_{k1}, \ldots, \boldsymbol{\mu}_{kT}, \boldsymbol{C}_k, \ldots\}$.

In order to enforce consistency of the component locations across time, we introduce a prior on the location parameter that penalizes large changes over consecutive time steps. This prior has a joint PDF proportional to

$$f_{\text{prior}}(\boldsymbol{\mu}_{k1}, \ldots, \boldsymbol{\mu}_{kT}) = \prod_{t=2}^{T} f_{\text{mvG}}(\boldsymbol{\mu}_{kt}-\boldsymbol{\mu}_{k(t-1)};0,\boldsymbol{Q}), \qquad (1)$$

where $\boldsymbol{Q}$ is a user-defined covariance matrix that controls how much the clusters are expected to drift.

2.2. EM algorithm for model fitting

Assuming independent spikes and a uniform prior on the other model parameters, we can obtain the maximum *a posteri-*