Multichannel Electrophysiological Spike Sorting via Joint Dictionary Learning & Mixture Modeling

Qisong Wu, David E. Carlson, Wenzhao Lian, Mingyuan Zhou, Colin R. Stoetzner, Daryl Kipke, Douglas Weber, Joshua T. Vogelstein, David B. Dunson and Lawrence Carin

***Abstract*—**[c3]**We propose a construction for joint feature learning and clustering of** [c4]**multichannel extracellular electrophysiological data across multiple recording periods**[c5] **for action potential detection and discrimination ("spike sorting").** [c6]**Our construction improves over the previous state-of-the art principally in four ways. First, via sharing information across channels, we can better distinguish between single-unit spikes and artifacts. Second, our proposed "focused mixture model" (FMM) elegantly deals with units appearing, disappearing, or *reappearing* over multiple recording days, an important consideration for any chronic experiment. Third, by jointly learning features and clusters, we improve performance over previous attempts that proceeded via a two-stage ("frequentist") learning process. Fourth, by directly modeling spike rate, we improve detection of sparsely spiking neurons. Moreover, our Bayesian construction seamlessly handles missing data. We present state-of-the-art performance without requiring manually tuning of many hyperparameters on both a public dataset with partial ground truth and a new experimental dataset.** [c7]

***Index Terms*—spike sorting, Bayesian, clustering, Dirichlet process**

I. Introduction

[c8]

SPIKE sorting of extracellular electrophysiological data is an important problem in contemporary neuroscience, with applications ranging from brain-machine interfaces [21] to neural coding [23] and beyond. Despite a rich history of work in this area [10], [31], room for improvement remains for automatic methods. An ideal tool for spike sorting achieves state-of-the-art performance and satisfies the following desiderate:

1) is fully automatic, obviating the need for the user to manually tune many "hyperparameters", especially the number of single-units,
2) benefits from multiple electrodes, multiple "sessions", and generally, more data,
3) is robust to artifactual noise, due to movement, for example,
4) handles "missing data", and
5) copes with changes in waveform shape over many days/weeks of recording, and
6) captures sparsely firing neurons.

Here we propose a Bayesian generative model and associated inference procedure; the first, to our knowledge, that satisfies all of the above desiderata.

Perhaps the most important advance in our present work over previous art is our joint feature learning and clustering strategy. More specifically, standard pipelines for processing extracellular electrophysiology data consist of the following steps: (i) filter the raw sensor readings, (ii) perform thresholding to "detect" the spikes, (iii) map each detected spike to a feature vector, and then (iv) cluster the feature vectors [20]. Our primary conceptual contribution to spike sorting methodologies is a novel unification of steps (iii) and (iv) that utilizes all available data in such a way as to satisfy all of the the above criteria. This "dictionary learning for clustering" approach improves results even for single channel and single recording experiments. More channels and more recordings simply improve our performance.

A. Previous Art

Although a comprehensive survey of previous spike sorting methods is beyond the scope of this manuscript, below we provide a summary of previous work as relevant to the above listed desiderata.

Perhaps those methods that are most similar to ours include a number of recent Bayesian methods for spike sorting [8], [13]. One can think of our method as a direct extension of theirs with a number enhancements. Most importantly, we learn features for clustering, rather than simply using principal components. We also incorporate multiple electrodes, assume a more appropriate prior over the number of clusters, and address longitudinal data.

Other popular methods utilize principal components analysis (PCA) [20] or wavelets [19] to find low-dimensional representations of waveforms for subsequent clustering. These methods typically require some manual tuning, for example, to choose the number of retained principal components. Moreover, these do not naturally handle missing data very well. Finally, these methods do not choose low-dimensional embeddings appropriate for downstream clustering, rather, they just "hope" that their low-dimensional embeddings do not discard valuable discriminating information.

Q. Wu, D. Carlson, W. Lian, M. Zhou and L. Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
C.R. Stoetzner and D. Kipke are with the Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA
D. Weber is with the Department of Biomedical Engineering, University of Pittsburgh, Pittsburgh, PA, USA
J. Vogelstein and D. Dunson are with the Department of Statistical Science, Duke University, Durham, NC, USA
Manuscript received October 27, 2012.

[c3] ~~A new model is developed for~~
[c4] *Text added.*
[c5] *Text added.*
[c6] *Text added.*
[c7] old abstract is commented out
[c8] the intro section is all new

Calabrese et al. [7] recently proposed a Mixture of Kalman Filters (MoK) model to elegantly and explicitly deals with slow changes in waveform shape. This approach also models spike rate (and even refractory period), but it does not address our other desiderata, perhaps most importantly, utilizing multiple electrodes or longitudinal data. It would be interesting to extend this work to utilize learned time-varying dictionaries rather than principal components.

Finally, several recently proposed methods address sparsely firing neurons [2], [22]. By directly incorporating firing rate into our model and inference algorithm (see Section II-C), our approach outperforms previous methods even in the absence of manual tuning (see Section III-E).

The remainder of our manuscript is organized as follows. Section II begins with a conceptual description of our model followed by mathematical details and experimental methods for new data. Section III begins by comparing the performance of our approach to several other previous state of the art methods, and then highlights the utility a number of additional features that our method includes. Section IV summarizes and provides some potential future directions. The appendix provides details of the relationships between our method and other related Bayesian models or constructions.

II. Models and Analysis

A. Model Concept

Our generative model derives from knowledge of the properties of electrophysiology signals. Specifically, we assume that each waveform can be represented by a sparse mixture of several dictionary elements, or features. Rather than presupposing a particular form of those features (as in much of harmonic analysis, for example, wavelets), we *learn* these features from the data. Importantly, we learn these features for the specific task at hand: spike sorting (aka, clustering). This is in contrast to other popular feature learning approaches, such as PCA or independent component analysis, which learn features to optimize a different objective function (for example, minimizing reconstruction error). This joint dictionary learning and clustering is a powerful idea with demonstrably good performance in a number of applications [**?**]. Moreover, statistical guarantees associated with such approaches are beginning to be understood [**?**]. Section II-B provides mathematical details for our Bayesian dictionary learning assumptions.

The above generative model requires a prior on the number of clusters. We assume a flexible prior motivated by the data. Specifically, regardless of the number of putative spikes detected, the number of different single units one could conceivably discriminate from a single electrode is upper bounded due to the conductive properties of the tissue. Thus, Bayesian non-parametric methods, which enable the number of clusters to increase to infinity as the number of threshold crossings increases, are inappropriate. Moreover, our prior is also appropriate for chronic recordings, in which single units may appear for a subset of the recording days, but also disappear and reappear intermittently. We refer to this prior as part of a mixture model a "focused mixture model" (FMM). Sections II-C and 9 provide mathematical details for the general mixture modeling case, and our specific focused mixture model assumptions.

We are also interested in multichannel recordings. When we have multiple channels that are within close proximity to one another, we can "borrow strength" across the channels to improve clustering accuracy. Moreover, we can ascertain that certain movement or other artifacts that look like spikes on a single channel, are in fact noise. For recording in awake behaving animals, such artifacts can be quite common.

Finally, we explicitly model the spike rate of each cluster. This can help address refractory issues, and perhaps more importantly, enables us detect sparsely firing neurons with high accuracy.

Because our model is fully Bayesian, we can easily generate data from our model to address missing data. Moreover, by placing relatively diffuse but informed hyperpriors on our model, our approach does not require any manual tuning. And by reformulating our priors, we can derive conjugacy which admits efficient Gibbs sampling. Section 16 provides details on these computations.

B. Bayesian dictionary learning

Consider electrophysiological data measured over a prescribed time interval. Specifically, let $\mathbf{X}_{ij} \in \mathbb{R}^{T \times N}$ represent the $j^{c1\,th}$ signal observed during interval i[c2](note that each j indexes a threshold crossing within a time intercal i; we do not indicate this dependency notationally for brevity.). The data are assumed recorded on each of N channels, from an N-element sensor array, and there are T time points associated with each detected spike waveform (the signals are aligned in time with respect to their peak value). In tetrode arrays [11], and related devices like those considered below, a single-unit event ([c3] action potential of a neuron) may be recorded on multiple adjacent channels, and therefore it is of interest to process the N signals associated with $\mathbf{X}_{ij}$ jointly; the joint analysis of all N signals is also useful for [c4]longitudinal analysis, discussed in Section III.

To constitute data $\mathbf{X}_{ij}$, [c5]we assume that threshold-based detection (or a related method) is performed on data measured from each of the N sensor channels. When a signal is detected on any of the channels, coincident data are also extracted from all N channels, within a window of (discretized) length T [c6]centered at the detection peak. On some of the channels data may be associated with a single-unit event, and on other channels the data may represent background noise. Both types of data (signal and noise) are modeled jointly, as discussed below.

Following [8], we employ dictionary learning to model each $\mathbf{X}_{ij}$; however, unlike [8] we jointly employ dictionary learning to all N channels in $\mathbf{X}_{ij}$ (rather than separately to each of the channels). The data are represented

$$\mathbf{X}_{ij} = \mathbf{D}\boldsymbol{\Lambda}\mathbf{S}_{ij} + \mathbf{E}_{ij}, \tag{1}$$

[c1] ~~th~~
[c2] *Text added.*
[c3] ~~*e.g.*~~
[c4] ~~data forensics~~
[c5] ~~it is assumed~~
[c6] *Text added.*

where $\mathbf{D} \in \mathbb{R}^{T\times K}$ represents a dictionary with K dictionary elements (columns), $\mathbf{\Lambda} \in \mathbb{R}^{K\times K}$ is a diagonal matrix with sparse diagonal elements, $\mathbf{S}_{ij} \in \mathbb{R}^{K\times N}$ represents the dictionary weights (factor scores), and $\mathbf{E}_{ij} \in \mathbb{R}^{T\times N}$ represents residual/noise. Let $\mathbf{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K)$ and $\mathbf{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N)$, with [c7] $\boldsymbol{d}_k, \boldsymbol{e}_n \in \mathbb{R}^T$. We impose [c8] priors

$$\boldsymbol{d}_k \sim \mathcal{N}(0, \frac{1}{T}\mathbf{I}_T) \ , \quad \boldsymbol{e}_n \sim \mathcal{N}(0, \text{diag}(\eta_1^{-1}, \ldots, \eta_T^{-1})), \quad (2)$$

where $\mathbf{I}_T$ is the $T\times T$ dimensional identity matrix [c1] and $\eta_t \in \mathbb{R}$ for all t.

We wish to impose that each column of $\mathbf{X}_{ij}$ lives in a linear subspace, with dimension and composition to be inferred. The composition of the subspace is defined by a selected subset of the columns of $\mathbf{D}$, and that subset is defined by the non-zero elements in the diagonal of $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$ [c2] and $\lambda_k \in \mathbb{R}$ for all k. We impose $\lambda_k \sim \nu\delta_0 + (1-\nu)\mathcal{N}_+(0, \alpha_0^{-1})$, with $\nu \sim \text{Beta}(a_0, b_0)$ and δ_0 a unit measure concentrated at zero. The hyperparameters [c3] $a_0, b_0 \in \mathbb{R}$ are set to encourage sparse $\boldsymbol{\lambda}$, and $\mathcal{N}_+(\cdot)$ represents a normal distribution truncated to be non-negative. Diffuse gamma priors are placed on $\{\eta_t\}$ and α_0.

Concerning the model priors, the assumption $\boldsymbol{d}_k \sim \mathcal{N}(0, \frac{1}{T}\mathbf{I}_T)$ is consistent with a conventional ℓ_2 regularization ([c4] shrinkage[c5]) on the dictionary elements. Similarly, the assumption $\boldsymbol{e}_n \sim \mathcal{N}(0, \text{diag}(\eta_1^{-1}, \ldots, \eta_T^{-1}))$ corresponds to an ℓ_2 fit of the data to the model, with a weighting on the norm as a function of the sample point (in time) of the signal. These priors are typically employed in dictionary learning; see [35] for a discussion of the connection between such priors and optimization-based dictionary learning.

C. Mixture modeling

A mixture model is imposed for the dictionary weights $\mathbf{S}_{ij} = (\boldsymbol{s}_{ij1}, \ldots, \boldsymbol{s}_{ijN})$, with $\boldsymbol{s}_{ijn} \in \mathbb{R}^K$; $\boldsymbol{s}_{ijn}$ [c6] defines the weights on the dictionary elements for the data associated with the nth channel (nth column) in $\mathbf{X}_{ij}$. Specifically,

$$\boldsymbol{s}_{ijn} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{ij}n}, \mathbf{\Omega}_{z_{ij}n}^{-1}), \quad (3)$$

$$z_{ij} \sim \textstyle\sum_{m=1}^M \pi_m^{(i)} \delta_m, \ (\boldsymbol{\mu}_{mn}, \mathbf{\Omega}_{mn}) \sim G_0 \quad (4)$$

where G_0 is a normal-Wishart distribution *QISONG: what parameters for this?*, $\pi_m^{(i)} > 0$, $\sum_{m=1}^M \pi_m^{(i)} = 1$, and $\{\boldsymbol{s}_{ijn}\}_{n=1,N}$ are all associated with cluster z_{ij}; $z_{ij} \in \{1, \ldots, M\}$ is an indicator variable defining with which cluster $\mathbf{X}_{ij}$ is associated[c7], and M is a user-specified upper bound on the total number of clusters possible.

The use of the Gaussian model in (3) is convenient, as it simplifies computational inference, and the normal-Wishart distribution G_0 is selected because it is the conjugate prior for a normal distribution. The key novelty we wish to address in this paper concerns design of the mixture probability vector $\boldsymbol{\pi}^{(i)} = (\pi_1^{(i)}, \ldots, \pi_M^{(i)})^T$.[c8]

D. [c9] A Focused Mixture Model

[c10] The vector $\boldsymbol{\pi}^{(i)}$ defines the probability with which each of the M mixture components are employed for data recording interval i. We wish to place a prior probability distribution on $\boldsymbol{\pi}^{(i)}$, and to infer an associated posterior distribution based upon the observed data. Let $b_m^{(i)}$ be a binary variable indicating whether interval i uses mixture component m. Let $\hat{\phi}_m^{(i)}$ correspond to the relative probability of including mixture component m in interval i, which is related to the firing rate of the single-unit corresponding to this cluster during that interval. Given this, we have our prior over clusters:

$$\pi_m^{(i)} = \tfrac{1}{Z} b_m^{(i)} \hat{\phi}_m^{(i)} \quad (5)$$

[c11] where $Z = \sum_{m'=1}^M b_{m'}^{(i)} \hat{\phi}_{m'}^{(i)}$ is the normalizing constant to ensure that $\sum_m \pi_m^{(i)} = 1$. To finalize this parameterization, we further assume the following priors on $b_m^{(i)}$ and $\hat{\phi}_n^{(i)}$:

$$\hat{\phi}_m^{(i)} \sim \text{Ga}(\phi_m, p_i/(1-p_i)), \quad \phi_m \sim \text{Ga}(\gamma_0, 1), \ p_i \sim \text{Beta}(a_0, b_0) \quad (6)$$

$$b_m^{(i)} \sim \text{Bern}(\nu_m), \quad \nu_m \sim \text{Beta}(\alpha/M, 1), \ \gamma_0 \sim \text{Ga}(c_0, 1/d_0) \quad (7)$$

where $\text{Ga}(\cdot)$ denotes the gamma distribution, and $\text{Bern}(\cdot)$ the Bernoulli distribution. Note that $\{\phi_m, \nu_m\}_{m=1,M}$ are shared across all intervals i, and it is in this manner we achieve joint clustering across all [c12] time intervals[c13]. *DEC: add sentence about p_i* The reasons for the choices of these various priors is discussed in Section IV-B, when making connections to related models. For example, the choice $b_m^{(i)} \sim \text{Bern}(\nu_m)$ with $\nu_m \sim \text{Beta}(\alpha/M, 1)$ is motivated by the connection to the Indian buffet process [15] as $M \to \infty$.

[c14] Note that we explicitly model the probability of spiking for each cluster component in each time interval. This implies that data are mapped to the same cluster if they have consistent signal shape *and* if the associated firing rate is consistent (note that the firing rates for some individual neurons can vary widely—from a few spikes/sec to [c15] greater than one hundred—but motor neuron firing rates are typically much lower and less variable). Consequently both the signal shape and firing rate dictates how the data are clustered.

[c7] ~~$\boldsymbol{d}_k \in \mathbb{R}^T$ and $\boldsymbol{e}_n \in \mathbb{R}^T$~~
[c8] ~~the~~
[c1] *Text added.*
[c2] *Text added.*
[c3] ~~(a_0, b_0)~~
[c4] ~~imposition of smoothness~~
[c5] my sense is that ℓ_2 does not smooth, ℓ_2 of the gradient smoothes, a simple ℓ_2 just shrinks.
[c6] ~~refines~~
[c7] *Text added.*
[c8] moved the rest of this section to the appendix.
[c9] ~~Hierarchical count and mixture modeling~~
[c10] *Text added.*
[c11] *Text added.*
[c12] *Text added.*
[c13] ~~, like via HDP~~
[c14] ~~Note that we explicitly model the number of clusters and quantity of data within a given cluster.~~
[c15] ~~100+;~~

E. [c16] Inference

The posterior distribution of model parameters is approximated via Gibbs sampling. Most of the update equations for the model are relatively standard due to conjugacy of consecutive distributions in the hierarchical model; these "standard" updates are not repeated here (see [8]). Perhaps the most important update equation is for ϕ_m, as we found this to be a critical component of the success of our inference. To perform such sampling we utilize the following lemma.

Lemma II.1. *Denote $s(n,j)$ as the Sterling numbers of the first kind [18] and $F(n,j) = (-1)^{n+j}s(n,j)/n!$ as their normalized and unsigned representations, with $F(0,0) = 1$, $F(n,0) = 0$ if $n > 0$, $F(n,j) = 0$ if $j > n$ and $F(n+1,j) = \frac{n}{n+1}F(n,j) + \frac{1}{n+1}F(n,j-1)$ if $1 \leq j \leq n$. Assuming $n \sim \text{NegBin}(\phi, p)$ is a negative binomial distributed random variable, and it is augmented into a compound Poisson representation [3] as*

$$n = \sum_{l=1}^{\ell} u_l,\ u_l \sim \text{Log}(p),\ \ell \sim \text{Pois}(-\phi \ln(1-p)) \tag{8}$$

where $\text{Log}(p)$ is the logarithmic distribution [3] with probability generating function $G(z) = \ln(1-pz)/\ln(1-p)$, $|z| < p^{-1}$, then we have

$$\Pr(\ell = j|n,\phi) = R_\phi(n,j) = F(n,j)\phi^j \Big/ \sum_{j'=1}^{n} F(n,j')\phi^{j'} \tag{9}$$

for $j = 0, 1, \cdots, n$.

The proof is provided in the Appendix.

Concerning sampling ϕ_m, since $\phi_m \propto \prod_{i:b_m^{(i)}=1} \text{NegBin}(n_{im}^*; \phi_m, p_i)\text{Ga}(\phi_m; \gamma_0, 1)$ [c1](see Appendix IV-B [c2] for details), using Lemma II.1, we can first sample a latent count variable ℓ_{im} for each n_{im}^* as

$$\Pr(\ell_{im} = l|n_{im}^*, \phi_m) = R_{\phi_m}(n_{im}^*, l),\quad l = 0, \cdots, n_{im}^*. \tag{10}$$

Since $\ell_{im} \sim \text{Pois}(-\phi_m \ln(1-p_i))$, using the conjugacy between the gamma and Poisson distributions, we have

$$\phi_m|\{\ell_{im}, b_m^{(i)}, p_i\} \sim \text{Ga}\left(\gamma_0 + \textstyle\sum_{i:b_m^{(i)}=1} \ell_{im},\ \frac{1}{1-\sum_{i:b_m^{(i)}=1}\ln(1-p_i)}\right). \tag{11}$$

Notice that marginalizing out ϕ_m in $\ell_{im} \sim \text{Pois}(-\phi_m \ln(1-p_i))$ results in $\ell_{im} \sim \text{NegBin}(\gamma_0, \frac{-\ln(1-p_i)}{1-\ln(1-p_i)})$, therefore, we can use the same data augmentation technique by sampling a latent count $\tilde{\ell}_{im}$ for each ℓ_{im} and then sampling γ_0 using the gamma Poisson conjugacy as

$$\Pr(\tilde{\ell}_{im} = l|\ell_{im}, \gamma_0) = R_{\gamma_0}(\ell_{im}, l),\quad l = 0, \cdots, \ell_{im} \tag{12}$$

$$\gamma_0|\{\tilde{\ell}_{im}, b_m^{(i)}, p_i\} \sim \text{Ga}\left(c_0 + \textstyle\sum_{i:b_m^{(i)}=1} \tilde{\ell}_{im},\ \frac{1}{d_0 - \sum_{i:b_m^{(i)}=1}\ln\left(1-\frac{-\ln(1-p_i)}{1-\ln(1-p_i)}\right)}\right).$$

Another important parameter is $b_m^{(i)}$. Since $b_m^{(i)}$ can only be zero if $n_{im}^* = 0$ and when $n_{im}^* = 0$, $\Pr(b_m^{(i)} = 1|-) \propto \text{NegBin}(0; \phi_m, p_i)\pi_m$ and $\Pr(b_m^{(i)} = 0|-) \propto (1-\pi_m)$, we have

$$b_m^{(i)}|\pi_m, n_{im}^*, \phi_m, p_i \sim \text{Bernoulli}\left(\delta(n_{im}^* = 0)\tfrac{\pi_m(1-p_i)^{\phi_m}}{\pi_m(1-p_i)^{\phi_m} + (1-\pi_m)} + \delta(n_{im}^* > 0)\right).$$

A large p_i thus indicates a large variance-to-mean ratio on n_{im}^* and M_i. Note that when $b_m^{(i)} = 0$, the observed zero count $n_{im}^* = 0$ is no longer explained by $n_{im}^* \sim \text{NegBin}(r_m, p_i)$, this satisfies the intuition that the underlying beta-Bernoulli process is governing whether a cluster would be used or not, and once it is activated, it is r_m and p_i that control how much it would be used.

F. Data Acquisition and Pre-processing

In this work we use two datasets, the popular "hc-1" dataset [c2] and a new dataset based upon experiments we have performed with freely moving rats (institutional review board approvals were obtained). These data will be made available to the research community. NeuroNexus$^{\text{TM}}$ sensors (Figure 3(a)) were humanely placed in the motor cortex, and electrophysiological data were measured during one-hour periods on eight consecutive days, starting on the day after implant (data were collected for additional days, but the signal quality degraded after 8 days, as discussed below). Note that nearby sensors are close enough to record the signal of a single or small group of neurons, termed a single-unit event. However, all eight sensors in a line are too far separated to simultaneously record a single-unit event on all eight.

The data were bandpass filtered (0.3-3 kHz), and then all signals 3.5 times the standard deviation of the background signal were deemed detections. The peak of the detection was placed in the center of a 1.3 msec window, which corresponds to $T = 40$ samples at the recording rate. The signal $\mathbf{X}_{ij} \in \mathbb{R}^{T\times N}$ corresponds to the data measured simultaneously across all N channels within this window. Here $N = 8$, with a concentration on the data measured from the 8 channels of the zoomed-in Figure 3(a).[c3]

seems like one of the reviewers wants more details here. i don't have them though. DEC has emailed colin, i believe, to get them.

III. Results

For these experiments we used a truncation level of $K = 40$ dictionary elements, and the number of mixture components was truncated to $M = 20$ (these truncation levels are upper bounds, and within the analysis a subset of the possible dictionary elements and mixture components are utilized). In dictionary learning, the gamma priors for $\{\eta_t\}$ and α_0 were set as $\text{Ga}(10^{-6}, 10^{-6})$. In the context of the [c4]focused mixture model, we set $a_0 = b_0 = 1$, $c_0 = 0.1$ and $d_0 = 0.1$. Prior $\text{Ga}(10^{-6}, 10^{-6})$ was placed on parameter α

[c16] ~~Computations~~
[c1] *Text added.*
[c2] *Text added.*
[c2] available from http://crcns.org/data-sets/hc/hc-1
[c3] i don't know what a "concentration of the data" means here.
[c4] ~~hierarchical count and mixture modeling~~

related to the [c5]Indian Buffet Process (see Appendix IV-B [c6]for details). None of these parameters have been tuned, and many related settings yield similar results. In all examples we ran 6,000 Gibbs samples, with the first 3,000 discarded as burn-in (however, typically high-quality results are inferred with far fewer samples, offering the potential for computational acceleration).

A. Real data with partial ground truth

We first consider publicly available dataset[c0] hc-1. These data consist of both extracellular recordings and an intracellular recording from a nearby neuron in the hippocampus of an anesthetized rat [16]. Intracellular recordings give clean signals on a spike train from a specific neuron, providing accurate spike times for that neuron. Thus, if we detect a spike in a nearby extracellular recording within a close time period (< 0.5ms) to an intracellular spike, we assume that the spike detected in the extracellular recording corresponds to the known neuron's spikes.

For the accuracy analysis, we determine one cluster that corresponds to the known neuron. We consider a spike to be correctly sorted if it is a known spike and is in the known cluster or if it is an unknown spike in [c1]an unknown cluster. [c2]Formally, we define:

$$\text{Accuracy} = \left\{1 - \frac{F_p + F_n}{\text{Total number of waveforms}}\right\} \times 100\% \tag{13}$$

[c3]where F_p and F_n are the total number of false positives and negatives, respectively.

We considered the widely used data d533101 and the same preprocessing from [7]. These data consist of 4-channel extracellular recordings and 1-channel intracellular recording. We used 2491 detected spikes and 786 of those spikes came from the known neuron. Accuracy of cluster results based on multiple methods are shown in Figure 1. [c4] We several different clustering schemes and two different strategies for learning low-dimensional representations of the data. Specifically, we learn low-dimensional representations using either: dictionary learning (DL) or the first two principal components (PC). Given this representation, we consider several different clustering strategies: (i) Matrix Dirichlet Process (MDP), which implements a DP on the X_{ij} *matrices*, as opposed to previous DP approaches on vectors [8], [13], (ii) focused mixture model (as described above), (iii) Hierarchical DP (HDP), (iv) independent DP (both DPs are from [8]), (v) Mixture of Kalman filters (MoK) [7], (vi) Gaussian mixture models (GMM) [6], and (vii) K-means (KMEANS) [**?**]. Although we do not consider all pairwise comparisons, we do consider many options. Note that all of the DL approaches are novel. It should be clear from Figure 1 that dictionary learning enhances performance over principal components for each clustering approach. Moreover, sharing information across channels, as in MDP and FMM (both novel constructions), seems to further improve performance.

Note that in Figure 1, in the context of PCA features, we considered the two principal components (similar results were obtained with the three principal components); when we considered the 20 principal components, for comparison, the results deteriorated, presumably because the higher-order components correspond to noise. An advantage of the proposed approach is that we model the noise explicitly, via the residual $\mathbf{E}_{ij}$ in (1); with PCA the signal and noise are not distinguished[c5] as well.

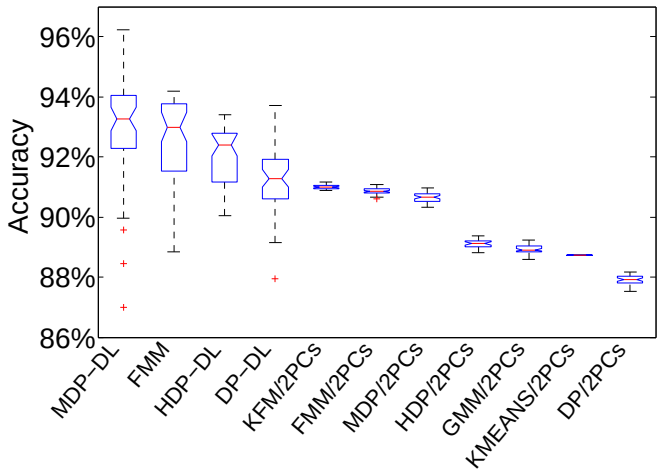


Fig. 1. *Accuracy of the various methods on d533101 data [16]. All abbreviations are explained in the main text (Section III-A). Note that dictionary learning dominates performance over principal components. Moreover, modeling multiple channels (as in MDP and FMM) dominates performance over operating on each channel separately.*

B. Handling missing data

The quantity of data acquired by a neural recording system is enormous, and therefore in many systems one first performs spike detection ([c6]for example, based on a threshold), and then a signal is extracted about each detection (a temporal window is placed around the peak of a given detection). This step is often imperfect, and significant portions of many of the spikes may be missing due to the windowed signal extraction (and the missing data are not retainable, as the original data are discarded). Conventional feature-extraction methods typically cannot be applied to such temporally clipped signals.

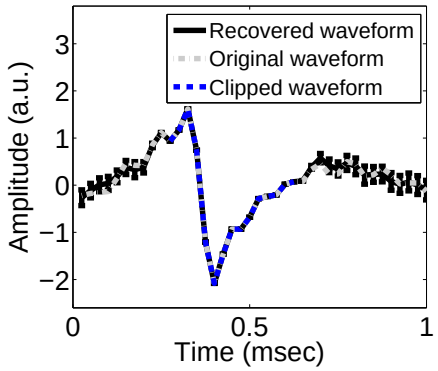
Returning to (1), this implies that some columns of the data $\mathbf{X}_{ij}$ may have missing entries. Conditioned on $\mathbf{D}$, $\mathbf{\Lambda}$, $\mathbf{S}_{ij}$, and $(\eta_1, \ldots, \eta_T)$, we have $\mathbf{X}_{ij} \sim \mathcal{N}(\mathbf{D}\mathbf{\Lambda}\mathbf{S}_{ij}, \text{diag}(\eta_1^{-1}, \ldots, \eta_T^{-1})$. The missing entries of $\mathbf{X}_{ij}$ may be treated as random variables, and they are integrated out analytically within the Gaussian likelihood function. Therefore, for the case of missing data in $\mathbf{X}_{ij}$, we simply evaluate (1) at the points of $\mathbf{X}_{ij}$ for which data

[c5] ~~IBP~~
[c6] *Text added.*
[c0] available from http://crcns.org/data-sets/hc/hc-1
[c1] ~~the~~
[c2] *Text added.*
[c3] *Text added.*
[c4] the rest of this paragraph is new, the old one is commented out
[c5] *Text added.*
[c6] ~~*e.g.*~~

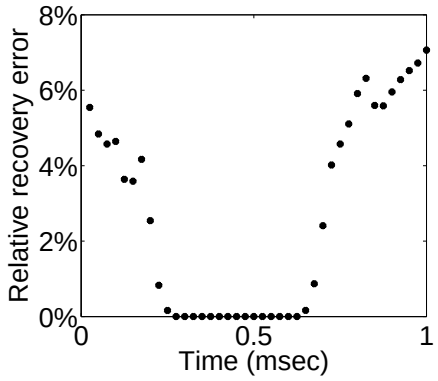
are observed. The columns of the dictionary $\mathbf{D}$ of course have support over the entire signal, and therefore given the inferred $\mathbf{S}_{ij}$ (in the presence of missing data), one may impute the missing components of $\mathbf{X}_{ij}$ via $\mathbf{D}\boldsymbol{\Lambda}\mathbf{S}_{ij}$. As long as, across all $\mathbf{X}_{ij}$, the same part of the signal is not clipped away (lost) for all observed spikes, by jointly processing all of the [c7]retained data (all spikes) we may infer $\mathbf{D}$, and hence infer missing data.

In practice we are less interested in observing the imputed missing parts of $\mathbf{X}_{ij}$ than we are in simply clustering the data, in the presence of missing data. By evaluating $\mathbf{X}_{ij} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\Lambda}\mathbf{S}_{ij}, \text{diag}(\eta_1^{-1}, \ldots, \eta_T^{-1})$ only at points for which data are observed, and via the mixture model in (4), we directly infer the desired clustering, in the presence of missing data (even if we are not explicitly interested in subsequently examining the imputed values of the missing data).

To examine the ability of the model to perform clustering in the presence of missing data, we reconsider the publicly available data from Section III-A. For the first 10% of the spike signals (300 spike waveforms), we impose that a fraction of the beginning and end of the spike is absent. The original signals are of length $T = 40$ samples. As a demonstration, for the "clipped" signals, the first 10 and the last 16 samples of the signals are missing. A clipped waveform example is shown in Figure 2(a); we compare the mean estimation of the signal, and the error bars reflect one standard deviation from the full posterior on the signal. In the context of the analysis, we processed all of the data as before, but now with these "damaged"/clipped signals. We observed that 94.11% of the non-damaged signals were clustered properly (for the one neuron for which we had truth), and 92.33% of the damaged signals were sorted properly. The recovered signal in Figure 2(a) is typical, and is meant to give a sense of the accuracy of the recovered missing signal. The ability of the model to perform spike sorting in the presence of substantial missing data is a key attribute of the dictionary-learning-based framework.



(a)



(b)

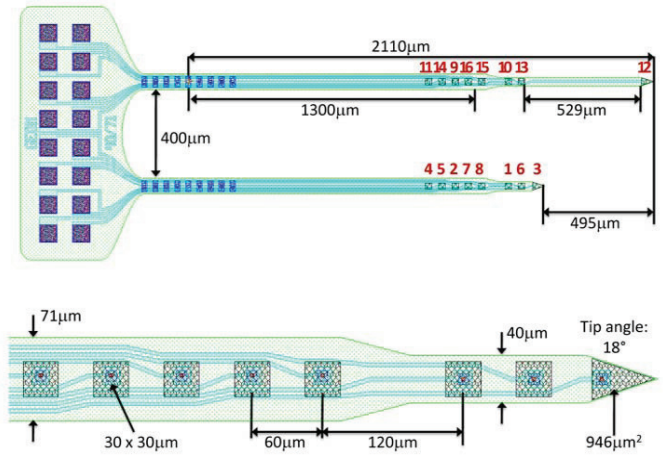
Fig. 2. *Our generative model easily addresses missing data. (a) Example of a clipped waveform from the publicly available data (blue), original waveform (gray) and recovery waveform (black); the error bars reflect one standard deviation from the posterior distribution on the underlying signal. (b) Relative errors (with respect to the mean estimated signal).*

C. [c1]*Longitudinal analysis of* *electrophysiological data*

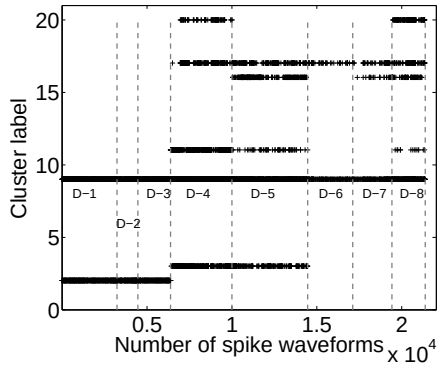
[c6]Figure 3(b)[c7](a) shows the recording probe used for the analysis of the rat motor cortex data. [c8] Figure 3(b) [c9]shows assignments of data to each of the possible clusters, for data measured across the 8 days, as computed by the proposed model ([c10]for example, for the first three days, two clusters were inferred). Results are shown for the maximum-likelihood collection sample. As a comparison to the proposed focused mixture model (FMM) of Section 9, we also considered the simplified HDP construction discussed in Section IV-B, with the $\boldsymbol{b}^{(i)}$ set to all ones (in both cases we employ the same form of dictionary learning, as in Section II-B). From Figure 3(c), it is observed that on held-out data the FMM yields improved results relative to the traditional HDP.

In fact, the proposed model was developed specifically to address the problem of multi-day [c11]longitudinal analysis of electrophysiological data, as a consequence of observed limitations of HDP (which are only partially illuminated by Figure 3(c)). Specifically, while the focused nature of the FMM allows learning of specialized clusters that occur over limited days, the "non-focused" HDP tends to merge similar but distinct clusters. This yields HDP results that are characterized by fewer total clusters, and by cluster characteristics that are less revealing of detailed neural processes. Patterns of observed neural activity may shift over a period of days due to many reasons, including cell death, tissue encapsulation, or device movement; this shift necessitates the FMM's ability to focus on subtle but important differences in the data properties over days. This ability to infer subtly different clusters is related to the focused topic model's ability [32] to discern distinct topics that differ in subtle ways. The study of large quantities of data (8 days) makes the ability to distinguish subtle differences in clusters more challenging (the DP-DL-based model works well

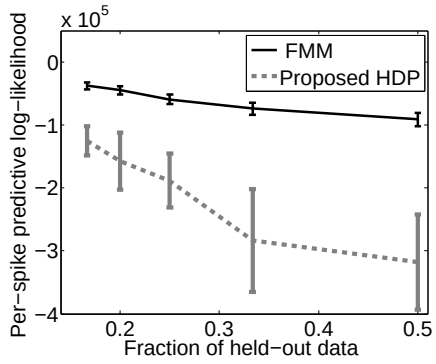
[c7] *Text added.*
[c1] ~~Forensic analysis of new longitudinal~~
[c6] *Text added.*
[c7] *Text added.*
[c8] ~~In~~
[c9] ~~are shown~~
[c10] ~~*e.g.*~~
[c11] ~~forensic~~



(a)



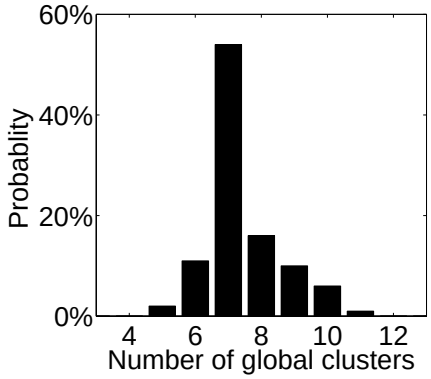
(b)



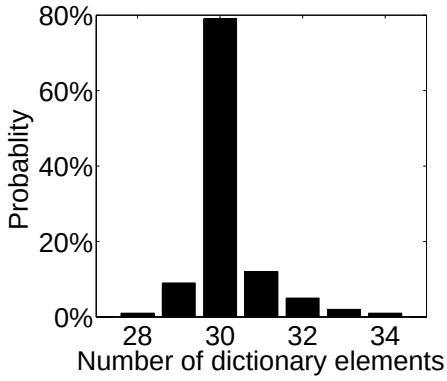
(c)

Fig. 3. [c4] *Longitudinal data analysis of the rat motor cortex data. (a) Schematic of the neural recording array that was placed in the rat motor cortex. The red numbers identify the sensors, and a zoom-in of the bottom-eight sensors is shown. The sensors are ordered by the order of the read-out pads, at left. The presented data are for sensors numbered 1 to 8, corresponding to the zoomed-in region. (b) From the maximum-likelihood collection sample, the apportionment of data among mixture components (clusters). Results are shown for 45 sec recording periods, on each of 8 days. For example, D-4 reflects data on day 4. Note that while the truncation level is such that there are 20 candidate clusters (vertical axis in (b)), only an inferred subset of clusters are actually used on any given day. (c) Predictive likelihood of held-out data. The horizontal axis represents the fraction of data held out during training.* [c5] *FMM-DL dominates HDP-DL on these data.*

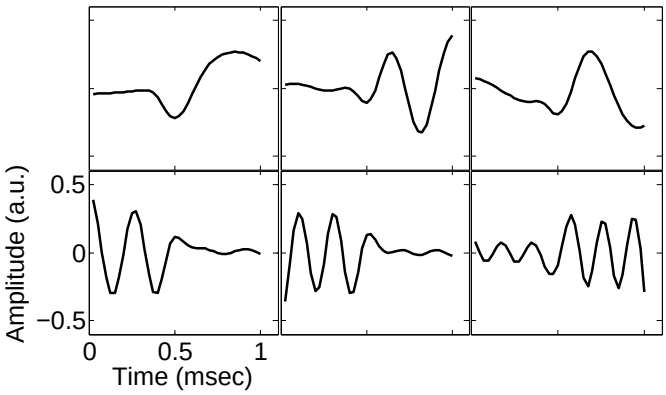
when observing data from one recording session, like in Figure 1, but the analysis of multiple days of data is challenging for HDP).



(a)



(b)



(c)

Fig. 4. [c3] *Posteriors and dictionaries from rat motor cortex data (the same data as in Figure 3).(a) Approximate posterior distribution on the number of global clusters (mixture components). (b) Approximate posterior distribution of the number of dictionary elements. (c) Examples inferred dictionary elements*[c4]*; amplitudes of dictionary elements are unit less.*

Note from Figure 3(b) that the number of detected signals is different for different recording days, despite the fact that the recording period reflective of these data (45 secs) is the same for all days. This highlights the need to allow modeling of different signal rates, as in our model but not emphasized

in these results.

Among the parameters inferred by the model are approximate posterior distributions on the number of clusters across all days, and on the required number of dictionary elements. These approximate posteriors are shown in Figures 4(a)[c1] and 4(b), and [c2] Figure 4(c) [c3]shows example dictionary elements. Although not shown for brevity, the $\{p_i\}$ had posterior means in excess of 0.9.

To better represent insight that is garnered from the model, [c4] Figure 5 [c5]depicts the inferred properties of three of the clusters, from Day 4 (D-4 in Figure 3(b)). Shown are the *mean* signal for the 8 channels in the respective cluster (for the 8 channels at the bottom of Figure 3(a)), and the error bars represent one standard deviation, as defined by the estimated posterior. Note that the cluster in [c6]the top row of Figure 5[c7] corresponds to a localized single-unit event, presumably from a neuron (or a coordinated small group of neurons) near the sensors associated with channels 7 and 8. The cluster in [c8]the middle row of Figure 5[c9] similarly corresponds to a single-unit event situated near the sensors associated with channels 3 and 6. Note the proximity of sensors 7 and 8, and sensors 3 and 6, from Figure 3(a). The HDP model uncovered the cluster in [c10]the top row of Figure 5[c11], but not that in [c12]the middle row of Figure 5[c13] [c14](not shown).[c15]

Note [c16]the bottom row of Figure 5[c17], in which the mean signal across all 8 channels is approximately the same (HDP also found related clusters of this type). This cluster is deemed to *not* be associated with a single-unit event, as the sensors are too physically distant across the array for the signal to be observed simultaneously on all sensors from a single neuron. This class of signals is deemed associated with an artifact or some global phenomena, due to (possibly) movement of the device within the brain, and/or because of charges that build up in the device and manifest signals with animal motion. Note that in [c18]the top two rows of Figure[c19] 5[c20] the error bars are relatively tight with respect to the strong signals in the set of eight, while the error bars in Figure 5(c) are more pronounced (the mean curves look [c21]smooth[c22], but this is based upon averaging thousands of signals).

[c1] -
[c2] ~~in~~
[c3] ~~are shown~~
[c4] ~~in~~
[c5] ~~are depicted~~
[c6] *Text added.*
[c7] ~~(a)~~
[c8] *Text added.*
[c9] ~~(b)~~
[c10] *Text added.*
[c11] ~~(a)~~
[c12] *Text added.*
[c13] ~~(b)~~
[c14] *Text added.*
[c15] QISONG: why do we not show this?
[c16] *Text added.*
[c17] ~~(c)~~
[c18] *Text added.*
[c19] ~~s~~
[c20] ~~(a)-(b)~~
[c21] ~~clean~~
[c22] i assume you meant smooth, i don't know what "clean" means.

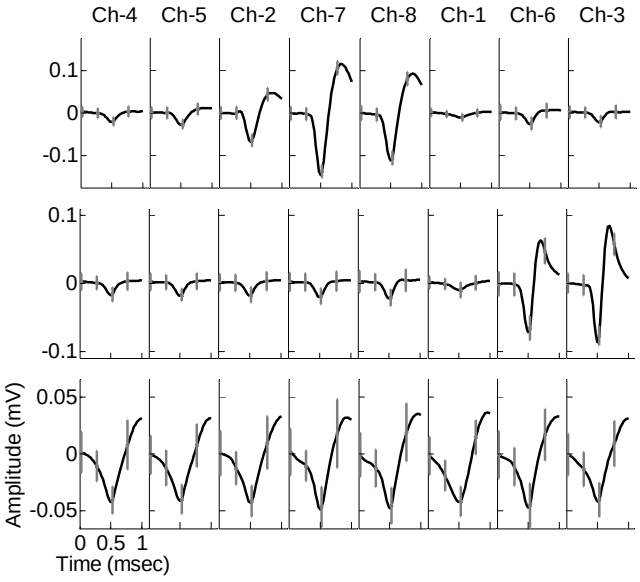


Fig. 5. *Example clusters inferred for data on the bottom 8 channels of Fig. 3(a). (a)-(b) Example of single-unit events. (c) Example of a cluster* not *attributed to a single-unit-event. The 8 signals are ordered from left to right consistent with the numbering of the 8 channels at the bottom of Figure 3(a). The black curves represent the mean, and the error bars are one standard deviation.*

In addition to recording the electrophysiological data, video was recorded of the rat throughout[c23] the experiment. Robust PCA [33] was used to quantify the change in the video from frame-to-frame, with high change associated with large motion by the animal (this automation is [c24]useful because one hour of data are collected on each day; direct human viewing is tedious and unnecessary). On Day 4, the model infers that in periods of high animal activity, 20% to 40% of the detected signals are due to single-unit events (depending on which portion of data are considered); during periods of relative rest 40% to 70% of detected signals are due to single-unit events[c25]. This suggests that animal motion causes signal artifacts, as discussed in Section I

In these studies the total fraction of single-unit events, even when at rest, diminishes with increasing number of days from sensor implant; this may be reflective of changes in the system due to the glial immune response of the brain [5], [24]. The discerning ability of the proposed FMM to distinguish subtly different signals, and analysis of data over multiple days, has played an important role in this analysis. Further, [c26]longitudinal analyses like that in Figure 5 were the principal reason for modeling the data on all $N = 8$ channels jointly (the ability to distinguish single-unit events from anomalies is predicated by this multi-channel analysis).

[c23] *Text added.*
[c24] ~~required~~
[c25] LARRY: this suggests that even during low-movement, we get about 30% false negatives? how would we know this?
[c26] ~~forensic~~

D. Model tuning

As constituted in Section II, the model is essentially parameter free. All of the hyperparameters are set in a relatively diffuse manner (see the discussion at the beginning of Section III), and the model infers the number of clusters and their composition with no parameter tuning required. While this may generally be viewed as a strength, there are situations for which a neuroscientist may wish to favor particular kinds of clusterings, and to have an adjustable parameter with which different solutions may be considered. All of the results presented above were manifested without any model tuning. We now discuss how one may constitute a single "knob" (parameter) that a neuroscientist may "turn" to examine different kinds of results.

In Section II-B the variance of additive noise $(e_1, \cdots, e_n)$ are controlled by the covariance $\text{diag}(\eta_1^{-1}, \cdots, \eta_T^{-1})$. If we set $\text{diag}(\eta_1^{-1}, \cdots, \eta_T^{-1}) = \omega_0^{-1}\mathbf{I}_T$, then parameter ω_0 may be tuned to control the variability (diversity) of spikes. The cluster diversity encouraged by setting different values of ω_0 in turn manifests different numbers of clusters, which a neuroscientist may adjust as desired. As an example, we consider the publicly available data from Section III-A, and clusterings (color coded) are shown for two settings of ω_0 [c1]in Figure 6. In this figure[c2], each spike is depicted in two-dimensional principal component (PC) space, taking the dominant two components; this is simply for display purposes, as here feature learning is done via dictionary learning, and in general more than two dictionary components are utilized to represent a given waveform.

The value of ω_0 defines how much of a given signal is associated with noise $\mathbf{E}_{ij}$, and how much is attributed to the term $\mathbf{D}\boldsymbol{\Lambda}\mathbf{S}_{ij}$ characterized by a summation of dictionary elements (see (1)). If ω_0 is large, then the noise contribution to the signal is small (because the noise variance is imposed to be small), and therefore the variability in the observed data is associated with variability in the underlying signal (and that variability is captured via the dictionary elements). Since the clustering is performed on the dictionary usage, if ω_0 is large we expect an increasing number of clusters, with these clusters capturing the greater diversity/variability in the underlying signal. By contrast, if ω_0 is relatively small, more of the signal is attributed to noise $\mathbf{E}_{ij}$, and the signal components modeled via the dictionary are less variable (variability is attributed to noise, not signal). Hence, as ω_0 diminishes in size we would expect fewer clusters. This phenomenon is observed in the example in Figure 6, with this representative of behavior we have observed in a large set of experiments[c3] on the rat motor cortex data.

E. Sparsely Firing Neurons[c6]

Recently, several manuscripts have directly addressed spike sorting in the present of sparsely firing neurons [2], [22]. Based on reviewer recommendations, we assessed the performance of

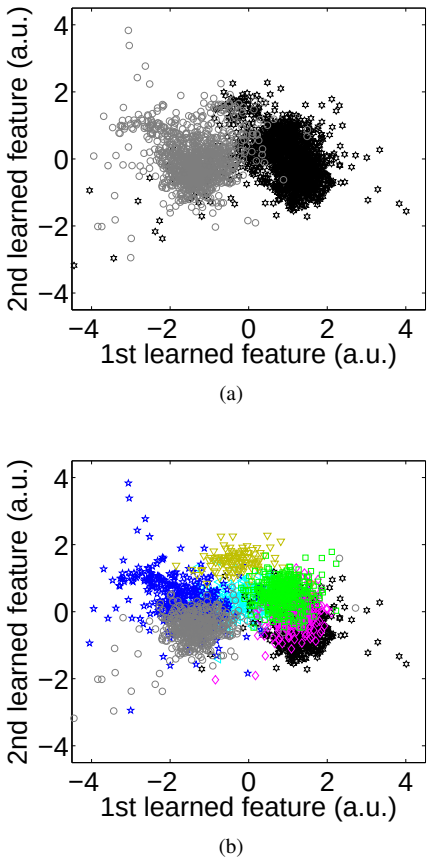


Fig. 6. [c5]*Effect of manually tuning ω_0 to obtain a different number of features for the rat motor cortex data. (a) 2-D plot in the first two learned features based on cluster result with $\omega_0 = 10^6$, and the number of inferred clusters is two. (b) 2-D plot in the first two learned features based on cluster result with $\omega_0 = 10^8$, and the number of inferred clusters is seven.*

FMM in such regimes utilizing the following synthetic data. First, we extracted spike waveforms from four clusters from the Pittsburgh dataset. We excluded all waveforms that did not clearly separate (Figure III-E(a1)) to obtain clear clustering criteria (Figure III-E(a2)). Then, we added independent and identically distributed Gaussian noise to each waveform at two different levels to obtain increasingly noisy and less-well separated clusters (Figure III-E(b1), (b2), (c1), and (c2)). We applied FMM, Waveclus and Waveclus focus [**?**] and ISOMAP dominant sets [**?**] to all three signal-to-noise ratio (SNR) regimes to assess our relative performance with the following results.

The third column of Figure III-E shows the posterior estimate of the number of clusters for each of the three scenarios. As long as SNR is relatively good, for example, higher than 2 in this simulation, the posterior number of clusters inferred by FMM correctly has its maximum at four clusters. Similarly, for the good and moderate SNR regimes, the confusion matrix is essentially a diagonal matrix, indicating that FMM assigns spikes to the correct cluster. Only in the poor SNR regime (SNR=1.5), does the posterior move away from the truth. This occurs because Unit 1 becomes over segmented, as depicted in

[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*
[c6] this whole subsection is new

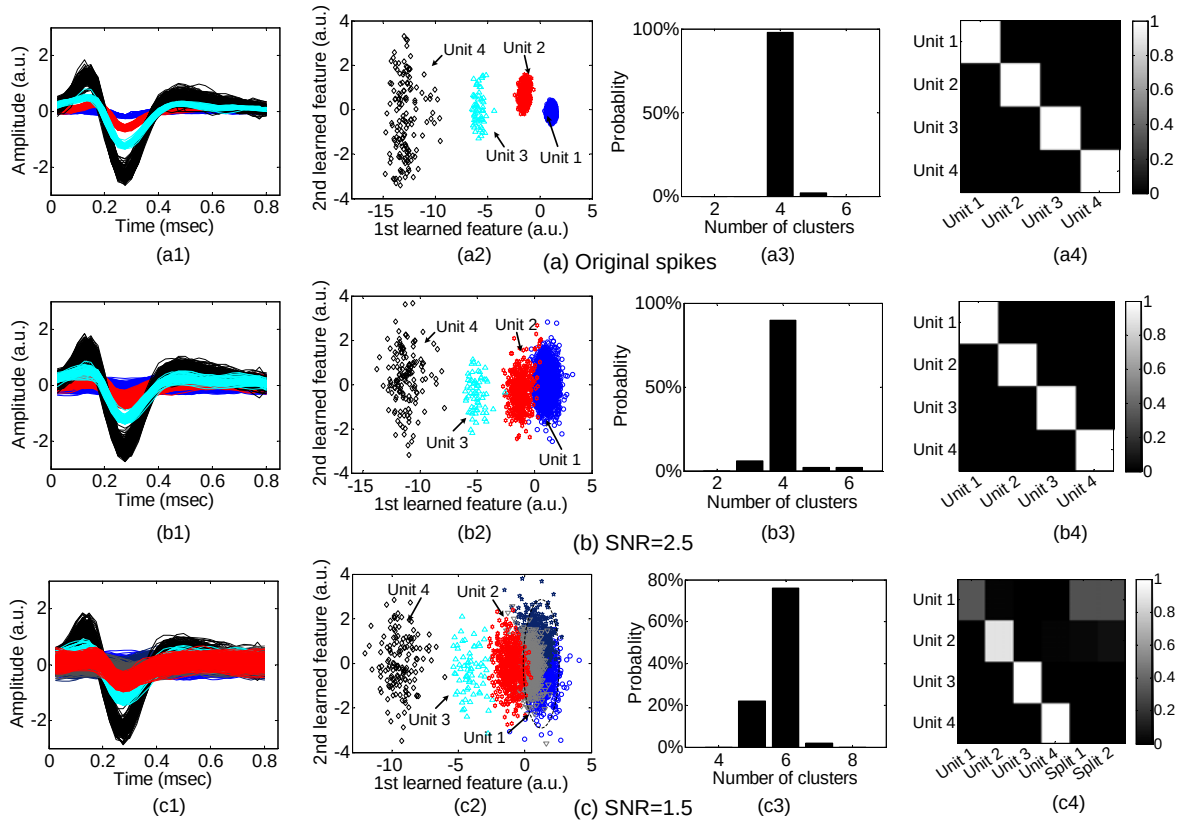


Fig. 7. *Sparse firing results on synthetic data based on the Pittsburgh dataset. The three rows correspond to three different signal-to noise ratio (SNR) levels: (a) 1, (b) 1.5, and (c) 2.5. The four columns correspond to: (1) cluster results of spike waveforms with colors representing different clusters, (2) plots of learned features based on cluster result, (3) approximate posterior distribution of cluster numbers, and (4) confusion matrix heatmap, Split 1 and Split 2 in (c4) are the clusters split from unit 1. Note that we accurately recover all the sparsely spiking neurons except the sparsest one in the noisiest regime.*

(c2). (c4) shows that only this unit struggles with assignment issues, suggestive of the possibility of a post-hoc correction if desired.

Figure 8(a) compares the performance of FMM to previously proposed methods. Even after fine-tuning the Waveclus method to obtain its optimal performance on these data, FMM yields a better accuracy. In addition to obtaining better point-estimates of spiking, via our Bayesian generative model, we also obtain posteriors over all random variables of our model, including number of spikes per unit. Figure **??**(b) and (c) show such posteriors, which may be used by the experimentalist to assess data quality.

F. Computational requirements

The software used for the tests in this paper were written in (non-optimized) Matlab, and therefore computational efficiency has not been a focus. The principal motivating focus of this study concerned [c1]interpretation of [c2]longitudinal spike waveforms, as discussed in Section 1, for which computation speed is desirable, but there is not a need for real-time processing ([c3]for example, for a prosthetic). Nevertheless, to give a sense of the computational load for the model, it takes about 20 seconds for each Gibbs sample, when considering analysis of 170800 spikes across $N = 8$ channels; computations were performed on a PC, specifically a Lenevo T420 (CPU is Inter(R) Core (TM) i7 M620 with 4 GB RAM). Significant computational acceleration may be manifested by coding in C, and via development of online methods for Bayesian inference ([c4]for example, see [29]). In the context of such online Bayesian learning one typically employs approximate variational Bayes inference rather than Gibbs sampling, which typically manifests significant acceleration [29].

IV. Discussion

A. Summary

A new focused mixture model (FMM) has been developed, motivated by real-world studies with longitudinal electrophysiological data, for which traditional methods like the hierarchical Dirichlet process have proven inadequate. In addition to performing "focused" clustering, the model jointly performs feature learning, via dictionary learning, which significantly improves performance over principal components. We explicitly model the count of signals within a recording period [c5]by

[c1] ~~forensics/~~
[c2] *Text added.*
[c3] ~~*e.g.*~~
[c4] ~~*e.g.*~~
[c5] *Text added.*

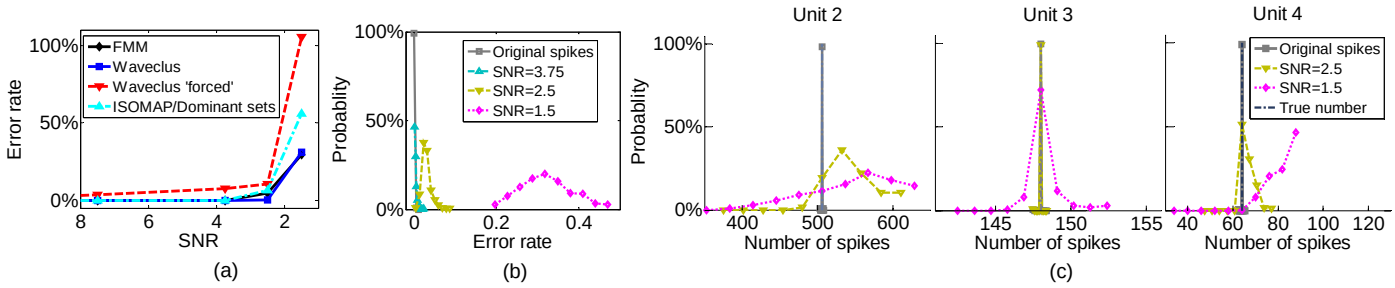


Fig. 8. *Performance analysis in the sparsely firing neuron case on synthetic data based on the Pittsburgh dataset. (a) Accuracy comparisons based on the cluster results under the various SNR. (b) Approximate posterior distributions of error rate for FMM-DL in the different SNR levels. (c) Approximate posterior distributions of spike waveform number for the unit 2, unit 3. and unit 4 under the various SNR regimes.*

p_i. The rate of neuron firing constitutes a primary information source [9], and therefore it is desirable that it be modeled. This rate is controlled here by a parameter [c6] $\phi_m^{(i)}$, and this was allowed to be unique for each recording period i.

B. Future Directions

In future research one may constitute a mixture model on [c1] $\phi_m^{(i)}$, with each mixture component reflective of a latent neural (firing) state; one may also explicitly model the time dependence of [c2] $\phi_m^{(i)}$ [c3], as in the MoK work [7]. Inference of this state could be important for decoding neural signals and controlling external devices or muscles. In future work one may also wish to explicitly account for covariates associated with animal activity [28], which may be linked to the firing rate we model here (we may regress p_i to observed covariates).

In the context of modeling and analyzing electrophysiological data, recent work on clustering models has accounted for refractory-time violations [7], [8], [13], which occur when two or more spikes that are sufficiently proximate are improperly associated with the same cluster/neuron (which is impossible physiologically due to the refractory time delay required for the same neuron to re-emit a spike). The methods developed in [8], [13] may be extended to the class of mixture models developed above. We have not done so for two reasons: (i) in the context of everything else that is modeled here (joint feature learning, clustering, and count modeling), the refractory-time-delay issue is a relatively minor issue in practice; and (ii) perhaps more importantly, an important issue is that not all components of electrophysiological data are spike related (which are associated with refractory-time issues). As demonstrated in Section III, a key component of the proposed method is that it allows us to distinguish single-unit (spike) events from other phenomena.

The perhaps the most important feature of spike sorting methods that we have not explicitly included in this model is "overlapping spikes" [1], [4], [12], [17], [27], [30], [34]. Preliminary analysis of our model in this regime (not shown), inspired by reviewer comments, demonstrated to us that while the FMM as written is insufficient to address this issue, a minor modification to FMM will enable "demixing" overlapping spikes. We are currently pursuing this avenue. Neuronal bursting—which can change the waveform shape of a neuron—is yet another possible avenue for future work.

[c6] ~~p_i~~
[c1] ~~p_i~~
[c2] ~~p_i~~
[c3] *Text added.*

APPENDIX

A. Connection to Previous Bayesian Non-Parametrics

The use of nonparametric Bayesian methods like the Dirichlet process (DP) [8], [13] removes some of the *ad hoc* character of classical clustering methods, but there are other limitations within the context of electrophysiological data analysis. The DP and related models are characterized by a scale parameter $\alpha > 0$, and the number of clusters grows as $\mathcal{O}(\alpha \log S)$ [25], with S the number of data samples. This growth without limit in the number of clusters with increasing data is undesirable in the context of electrophysiological data, for which there are a finite set of processes responsible for the observed data. Further, when jointly performing mixture modeling across multiple tasks, the *hierarchical* Dirichlet process (HDP) [26] shares all mixture components, which may undermine inference of subtly different clusters.

[c4] In this paper we integrate dictionary learning and clustering for analysis of electrophysiological data, as in [8], [14]. However, as an alternative to utilizing a method like DP or HDP [8], [13] for clustering, we develop a new hierarchical clustering model in which the number of clusters is modeled explicitly; this implies that we model the number of underlying [c5] neurons—or clusters—separately from the firing rate, with the latter controlling the total number of observations. This is done by integrating the Indian buffet process (IBP) [15] with the Dirichlet distribution, similar to [32], but with unique characteristics. The IBP is a model that may be used to *learn* features representative of data, and each potential feature is a "dish" at a "buffet"; each data sample (here [c6] a neuronal

[c4] ~~Another limitation of almost all existing electrophysiological data methods is that they only focus on clustering the observed data. While assigning data to a cluster is important, such frameworks do not address one of the most significant aspects of spike data: recent research indicates that a major portion of the information content related to neural spiking is carried in the spike *rate*, in terms of the number of spikes within a defined interval. It is therefore not only desirable to model the clustering of the data, but also the rate of spike firing, ideally with these modeled jointly.~~
[c5] ~~neural processes~~
[c6] *Text added.*

spike) selects which features from the "buffet" are most appropriate for its representation. The Dirichlet distribution is used for clustering data, and therefore here we jointly perform feature learning and clustering, by integrating the IBP with the Dirichlet distribution. The proposed framework explicitly models the quantity of data ([c7]for example, spikes) measured within a given recording interval. We believe that this is the first time the firing rate of electrophysiological data is modeled jointly with clustering (and, here, jointly with feature/dictionary learning). The model demonstrates state-of-the-art clustering performance on publicly available data. Further, concerning distinguishing single-unit-events, we demonstrate how this may be achieved using the proposed method, considering new measured (experimental) electrophysiological data.

B. Relationship to existing models

A typical prior for $\boldsymbol{\pi}^{(i)}$ is a symmetric Dirichlet distribution [14],

$$\boldsymbol{\pi}^{(i)} \sim \text{Dir}(\tilde{\alpha}_0/M, \dots, \tilde{\alpha}_0/M). \tag{14}$$

In the limit[c1], $M \rightarrow \infty$[c2], this reduces to a draw from a Dirichlet process [8], [13], represented $\boldsymbol{\pi}^{(i)} \sim \text{DP}(\tilde{\alpha}_0 G_0)$, with G_0 the "base" distribution defined in (4). Rather than drawing each $\boldsymbol{\pi}^{(i)}$ independently [c3]from DP($\tilde{\alpha}_0 G_0$), we may consider the hierarchical Dirichlet process (HDP) [26] as

$$\boldsymbol{\pi}^{(i)} \sim \text{DP}(\tilde{\alpha}_1 G) \ , \qquad G \sim \text{DP}(\tilde{\alpha}_0 G_0) \tag{15}$$

The HDP construction imposes that the $\{\boldsymbol{\pi}^{(i)}\}$ share the same set of "atoms" $\{\boldsymbol{\mu}_{mn}, \boldsymbol{\Omega}_{mn}\}$, implying a sharing of the different types of clusters across the time intervals i at which data are collected. A detailed discussion of the HDP formulation is provided in [8].

These models have limitations in that the inferred number of clusters grows with observed data (here the clusters are ideally connected to [c4]neurons, the number of which will not necessarily grow with [c5]longer samples). Further, the above clustering model assumes the number of samples is given, and hence is not modeled (the information-rich firing rate is not modeled). Below we develop a framework that yields hierarchical clustering like HDP, but the number of clusters and the data count ([c6]for example, spike rate) are modeled explicitly.

Let the total set of data measured during interval i be represented $\boldsymbol{\mathcal{D}}_i = \{\mathbf{X}_{ij}\}_{j=1}^{M_i}$[c7], where M_i is the total number of events during interval i. In the experiments below, a "recording interval" corresponds to a day on which data were recorded for an hour (data are collected separately on a sequence of days), and the set $\{\mathbf{X}_{ij}\}_{j=1}^{M_i}$ defines all signals that exceeded a threshold during that recording period. In addition to modeling M_i, we wish to infer the number of distinct clusters C_i characteristic of $\boldsymbol{\mathcal{D}}_i$, and the relative fraction (probability) with which the M_i observations are apportioned to the C_i clusters.

Let n_{im}^* represent the number of data samples in $\boldsymbol{\mathcal{D}}_i$ that are apportioned to cluster $m \in \{1, \dots, M\} = \mathcal{S}$, with $M_i = \sum_{m=1}^{M} n_{im}^*$. The set $\mathcal{S}_i \subset \mathcal{S}$, with $C_i = |\mathcal{S}_i|$, defines the *active* set of clusters for representation of $\boldsymbol{\mathcal{D}}_i$, and therefore M serves as an upper bound ($n_{im}^* = 0$ for $m \in \mathcal{S} \setminus \mathcal{S}_i$).

We impose $n_{im}^* \sim \text{Poisson}(b_m^{(i)} \hat{\phi}_m^{(i)})$ with

$$\hat{\phi}_m^{(i)} \sim \text{Ga}(\phi_m, p_i/(1-p_i)), \tag{16}$$

$$\phi_m \sim \text{Ga}(\gamma_0, 1), \ p_i \sim \text{Beta}(a_0, b_0), \tag{17}$$

$$b_m^{(i)} \sim \text{Bern}(\nu_m), \nu_m \sim \text{Beta}(\alpha/M, 1), \ \gamma_0 \sim \text{Ga}(c_0, 1/d_0) \tag{18}$$

Note that $n_{im}^* = 0$ when $b_m^{(i)} = 0$, and therefore $\boldsymbol{b}^{(i)} = (b_1^{(i)}, \dots, b_M^{(i)})^T$ defines indicator variables identifying the active subset of clusters $\mathcal{S}_i$ for representation of $\boldsymbol{\mathcal{D}}_i$. Marginalizing out $\hat{\phi}_m^{(i)}$, $n_{im}^* \sim \text{NegBin}(b_m^{(i)} \phi_m, p_i)$. This emphasize another motivation for the form of the prior: the negative binomial modeling of the counts (firing rate) is more flexible than a Poisson model, as it allows the mean and variance on the number of counts to be different (they are the same for a Poisson model).

While the above construction yields a generative process for the number, n_{im}^*, of elements of $\boldsymbol{\mathcal{D}}_i$ apportioned to cluster m, it is desirable to explicitly associate each member of $\boldsymbol{\mathcal{D}}_i$ with one of the clusters (to know not just *how many* members of $\boldsymbol{\mathcal{D}}_i$ are apportioned to a given cluster, but also *which* data are associated with a given cluster). Toward this end, consider the alternative equivalent generative process for $\{n_{im}^*\}_{m=1,M}$ (see Lemma 4.1 in [36] for a proof of equivalence): first draw $M_i \sim$ Poisson($\sum_{m=1}^{M} b_m^{(i)} \hat{\phi}_m^{(i)}$), and then

$$(n_{i1}^*, \dots, n_{iM}^*) \sim \text{Mult}(M_i; \pi_1^{(i)}, \dots, \pi_M^{(i)}) \tag{19}$$

$$\pi_m^{(i)} = b_m^{(i)} \hat{\phi}_m^{(i)} / \textstyle\sum_{m'=1}^{M} b_{m'}^{(i)} \hat{\phi}_{m'}^{(i)} \tag{20}$$

with $\hat{\phi}_m^{(i)}$, $\{\phi_m\}$, $\{b_m^{(i)}\}$, and $\{p_i\}$ constituted as in (6)-(7). Note that we have $M_i \sim$ NegBin($\sum_{m=1}^{M} b_m^{(i)} \phi_m, p_i$) by marginalizing out $\hat{\phi}_m^{(i)}$.

Rather than drawing $(n_{i1}^*, \dots, n_{iM}^*) \sim \text{Mult}(M_i; \pi_1^{(i)}, \dots, \pi_M^{(i)})$, for each of the M_i data we may draw indicator variables $z_{ij} \sim \sum_{m=1}^{M} \pi_m^{(i)} \delta_m$, where δ_m is a unit measure concentrated at the point m. Variable z_{ij} assigns data sample $j \in \{1, \dots, M_i\}$ to one of the M possible clusters, and $n_{im}^* = \sum_{j=1}^{M_i} 1(z_{ij} = m)$, with $1(\cdot)$ equal to one if the argument is true, and zero otherwise. The probability vector $\boldsymbol{\pi}^{(i)}$ defined in (20) is now used within the mixture model in (4).

As a consequence of the manner in which $\hat{\phi}_m^{(i)}$ is drawn in (6), and the definition of $\boldsymbol{\pi}^{(i)}$ in (20), for *any* $p_i \in (0, 1)$, the proposed model imposes

$$\boldsymbol{\pi}^{(i)} \sim \text{Dir}(b_1^{(i)} \phi_1, \dots, b_M^{(i)} \phi_M) \tag{21}$$

Hence the proposed model is a generalization of (14). Considering the limit $M \rightarrow \infty$, and upon marginalizing out the $\{\nu_m\}$, the binary vectors $\{\boldsymbol{b}^{(i)}\}$ are drawn from the Indian buffet process (IBP), denoted $\boldsymbol{b}^{(i)} \sim \text{IBP}(\alpha)$. The number of non-zero components in each $\boldsymbol{b}^{(i)}$ is drawn from Poisson(α),

[c7] ~~e.g.~~
[c1] *Text added.*
[c2] *Text added.*
[c3] ~~$\boldsymbol{\pi}^{(i)} \sim \text{DP}(\tilde{\alpha}_0 G_0)$~~
[c4] ~~neural processes~~
[c5] ~~increasing data~~
[c6] ~~e.g.~~
[c7] *Text added.*

and therefore for finite α the number of non-zero components in $\boldsymbol{b}^{(i)}$ is finite, even when $M \rightarrow \infty$. Consequently $\text{Dir}(b_1^{(i)}\phi_1, \ldots, b_M^{(i)}\phi_M)$ is well defined even when $M \rightarrow \infty$ since, with probability one, there are only a finite number of non-zero parameters in $(b_1^{(i)}\phi_1, \ldots, b_M^{(i)}\phi_M)$. This model is closely related to the compound IBP Dirichlet (CID) process developed in [32], with the following differences.

Above we have explicitly derived the relationship between the negative binomial distribution and the CID, and with this understanding we recognize the importance of p_i; the CID *assumes* $p_i = 1/2$, but there is no theoretical justification for this. Note that $M_i \sim \text{NegBin}(\sum_{m=1}^M b_m^{(i)}\phi_m^{(i)}, p_i)$. The mean of M_i is $(\sum_{m=1}^M b_m^{(i)}\phi_m)p_i/(1-p_i)$, and the variance is $(\sum_{m=1}^M b_m^{(i)}\phi_m)p_i/(1-p_i)^2$. If p_i is fixed to be [c1] 1/2 as in [32], this implies that we believe that the variance is two times the mean, and the mean and variance of M_i are the same for all intervals i and i' for which $\boldsymbol{b}^{(i)} = \boldsymbol{b}^{(i')}$. However, in the context of electrophysiological data, the rate at which neurons fire plays an important role in information content [9]. Therefore, there are many cases for which intervals i and i' may be characterized by firing of the same neurons (*i.e.*, $\boldsymbol{b}^{(i)} = \boldsymbol{b}^{(i')}$) but with very different rates ($M_i \neq M_{i'}$). The modeling flexibility imposed by inferring p_i therefore plays an important practical role for modeling electrophysiological data, and likely for other clustering problems of this type.

To make a connection between the proposed model and the HDP, motivated by (6)-(7), consider $\bar{\boldsymbol{\phi}} = (\bar{\phi}_1, \cdots, \bar{\phi}_M) \sim \text{Dir}(\gamma_0, \cdots, \gamma_0)$, which corresponds to $(\phi_1, \ldots, \phi_M)/\sum_{m'=1}^M \phi_{m'}$. From $\bar{\boldsymbol{\phi}}$ we yield a *normalized* form of the vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)$. The normalization constant $\sum_{m=1}^M \phi_m$ is lost after drawing $\bar{\boldsymbol{\phi}}$; however, because $\phi_m \sim \text{Ga}(\gamma_0, 1)$, we may consider drawing $\tilde{\alpha}_1 \sim \text{Ga}(M\gamma_0, 1)$, and *approximating* $\boldsymbol{\phi} \approx \tilde{\alpha}_1\bar{\boldsymbol{\phi}}$. With this approximation for $\boldsymbol{\phi}$, $\boldsymbol{\pi}^{(i)}$ may be drawn approximately as $\boldsymbol{\pi}^{(i)} \sim \text{Dir}(\tilde{\alpha}_1 b_1^{(i)}\bar{\phi}_1, \ldots, \tilde{\alpha}_1 b_M^{(i)}\bar{\phi}_M)$. This yields a simplified and approximate hierarchy

$$\boldsymbol{\pi}^{(i)} \sim \text{Dir}(\tilde{\alpha}_1(\boldsymbol{b}^{(i)} \odot \bar{\boldsymbol{\phi}})) \qquad (22)$$
$$\bar{\boldsymbol{\phi}} = (\bar{\phi}_1, \cdots, \bar{\phi}_M) \sim \text{Dir}(\gamma_0, \cdots, \gamma_0),\ \tilde{\alpha}_1 \sim \text{Ga}(M\gamma_0, 1)$$

with $\boldsymbol{b}^{(i)} \sim \text{IBP}(\alpha)$ and $\odot$ representing a pointwise/Hadamard product. If we consider $\gamma_0 = \hat{\alpha}_0/M$, and the limit $M \rightarrow \infty$, with $\boldsymbol{b}^{(i)}$ all ones, this corresponds to the HDP, with $\hat{\alpha}_1 \sim \text{Ga}(\hat{\alpha}_0, 1)$. Therefore, the proposed model is intimately related to the HDP, with three differences: (i) p_i is not restricted to be 1/2, which adds flexibility when modeling counts; (ii) rather than drawing $\bar{\boldsymbol{\phi}}$ and the normalization constant $\tilde{\alpha}_1$ separately, as in the HDP, in the proposed model $\boldsymbol{\phi}$ is drawn directly via $\phi_m \sim \text{Ga}(\gamma_0, 1)$, with an explicit link to the count of observations $M_i \sim \text{NegBin}(\sum_{m=1}^M b_m^{(i)}\phi_m, p_i)$; and (iii) the binary vectors $\boldsymbol{b}^{(i)}$ "focus" the model on a sparse subset of the mixture components, while in general, within the HDP, all mixture components have non-zero probability of occurrence for all tasks i. As demonstrated in Section III, this focusing nature of the proposed model is important in the context of electrophysiological data.

[c1] ~~0.5~~

C. Proof of Lemma 3.1

Proof: Denote $w_j = \sum_{l=1}^j u_l$, $j = 1, \cdots, m$. Since w_j is the summation of j iid $\text{Log}(p)$ distributed random variables, the probability generating function of w_j can be expressed as $G_{W_j}(z) = [\ln(1-pz)/\ln(1-p)]^j$, $|z| < p^{-1}$, thus we have

$$\begin{aligned}\Pr(w_j = m) &= G_{W_j}^{(m)}(0)/m! = \tfrac{d^m}{dz^m}[\ln(1-pz)/\ln(1-p)]^j \\ &= (-1)^m p^j j! s(m,j)/[\ln(1-p)]^j \end{aligned} \qquad (23)$$

where we use the property that $[\ln(1+x)]^j = j!\sum_{n=j}^\infty \frac{s(n,j)x^n}{n!}$ [18]. Therefore, we have

$$\begin{aligned}\Pr(\ell = j|-) &\propto \Pr(w_j = n)\text{Pois}(j; -r\ln(1-p)) \\ &\propto (-1)^{n+j}s(n,j)/n!r^j = F(n,j)r^j. \end{aligned} \qquad (24)$$

■

The values $F(n,j)$ can be iteratively calculated and each row sums to one, e.g., the 3rd to 5th rows are

$$\begin{pmatrix} 2/3! & 3/3! & 1/3! & 0 & 0 & 0 & \cdots \\ 6/4! & 11/4! & 6/4! & 1/4! & 0 & 0 & \cdots \\ 24/5! & 50/5! & 35/5! & 10/5! & 1/5! & 0 & \cdots \end{pmatrix}.$$

To ensure numerical stability when $\phi > 1$, we may also iteratively calculate the values of $R_\phi(n,j)$.

REFERENCES

[1] Dimitrios A Adamos, Nikolaos A Laskaris, Efstratios K Kosmidis, and George Theophilidis. NASS: an empirical approach to spike sorting with overlap resolution based on a hybrid noise-assisted methodology. *Journal of neuroscience methods*, 190(1):129–42, June 2010.

[2] Dimitrios A Adamos, Nikolaos A Laskaris, Efstratios K Kosmidis, and George Theophilidis. In quest of the missing neuron: spike sorting based on dominant-sets clustering. *Computer methods and programs in biomedicine*, 107(1):28–35, July 2012.

[3] F. J. Anscombe. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, 1949.

[4] Izhar Bar-Gad, Ya'acov Ritov, Eilon Vaadia, and Hagai Bergman. Failure in identification of overlapping spikes from multiple neuron activity causes artificial correlations. *Journal of Neuroscience Methods*, 107(1-2):1–13, May 2001.

[5] R. Biran, D.C. Martin, and P.A. Tresco. Neuronal cell loss accompanies the brain tissue response to chronically implanted silicon microelectrode arrays. *Exp. Neurol.*, 2005.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[7] A. Calabrese and L. Paniski. Kalman filter mixture model for spike sorting of non-stationary data. *J. Neuroscience Methods*, 2010.

[8] B. Chen, D.E. Carlson, and L. Carin. On the analysis of multi-channel neural spike data. In *NIPS*, 2011.

[9] J.P. Donoghue, A. Nurmikko, M. Black, and L.R. Hochberg. Assistive technology and robotic control using mortor cortex ensemble-based neural interface systems in humans with tetraplegia. *J. Physiol.*, 2007.

[10] Gaute T Einevoll, Felix Franke, Espen Hagen, Christophe Pouzat, and Kenneth D Harris. Towards reliable spike-train recordings from thousands of neurons with multielectrodes. *Current opinion in neurobiology*, 22(1):11–7, February 2012.

[11] A.A. Emondi, S.P. Rebrik, A.V. Kurgansky, and K.D. Miller. Tracking neurons recorded from tetrodes across time. *J. Neuro. Meth.*, 2004.

[12] Felix Franke, Michal Natora, Clemens Boucsein, Matthias H J Munk, and Klaus Obermayer. An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes. *Journal of computational neuroscience*, 29(1-2):127–48, August 2010.

[13] J. Gasthaus, F. Wood, D. Gorur, and Y.W. Teh. Dependent Dirichlet process spike sorting. *In Advances in Neural Information Processing Systems*, 2009.

[14] D. Gorur, C. Rasmussen, A. Tolias, F. Sinz, and N. Logothetis. Modelling spikes with mixtures of factor analysers. *Pattern Recognition*, 2004.

[15] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.
[16] D. A. Henze, Z. Borhegyi, J. Csicsvari, A. Mamiya, K. D. Harris, and G. Buzsaki. Intracellular feautures predicted by extracellular recordings in the hippocampus in vivo. *J. Neurophysiology*, 2010.
[17] Joshua A Herbst, Stephan Gammeter, David Ferrero, and Richard H R Hahnloser. Spike sorting with hidden Markov models. *Journal of neuroscience methods*, 174(1):126–34, September 2008.
[18] N.L. Johnson, A.W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.
[19] J. C. Letelier and P. P. Weber. Spike sorting based on discrete wavelet transform coefficients. *J. Neuroscience Methods*, 2000.
[20] M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 1998.
[21] Miguel A L Nicolelis and Mikhail A Lebedev. Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nature Reviews Neuroscience*, 10(7):530–540, 2009.
[22] Carlos Pedreira, Juan Martinez, Matias J Ison, and Rodrigo Quian Quiroga. How many neurons can we see with current spike sorting algorithms? *Journal of neuroscience methods*, 211(1):58–65, October 2012.
[23] F Rieke, D Warland, R De Ruyter Van Steveninck, and W Bialek. *Spikes: Exploring the Neural Code*, volume 20. MIT Press, 1997.
[24] D.H. Szarowski, M.D. Andersen, S. Retterer, A.J. Spence, M. Isaacson, H.G. Craighead, J.N. Turner, and W. Shain. Brain responses to micromachined silicon devices. *Brain Res.*, 2003.
[25] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.
[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *J. Am. Stat. Ass.*, 2006.
[27] Carlos Vargas-Irwin and John P Donoghue. Automated spike sorting using density grid contour clustering and subtractive waveform decomposition. *Journal of neuroscience methods*, 164(1):1–18, August 2007.
[28] V. Ventura. Automatic spike sorting using tuning information. *Neural Computation*, 2009.
[29] C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. *Artificial Intelligence and Statistics*, 2011.
[30] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. *The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.
[31] Bruce C Wheeler. Multi-unit neural spike sorting. *Annals of Biomedical Engineering*, 19(5), 1991.
[32] S. Williamson, C. Wang, K.A. Heller, and D.M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.
[33] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *Neural Information Processing Systems (NIPS)*, 2009.
[34] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection. *Proceedings of Neural Information Processing Systems*, 2004.
[35] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Trans. Image Processing*, 2012.
[36] M. Zhou, L.A. Hannah, D.B. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

REBUTTAL

General comments

We would like to thank both reviewers for their very helpful comments. Below, we have appended the reviewer comments, along with our responses in *red italics*. Quotes from the revised manuscript appear in *black italics*.

A. Response to Reviewer 1

Thank you for your very helpful comments. We address each of your concerns in order.

1) Major Concerns:

a. While the keywords might indicate that the the paper is dealing with spike sorting, the term spike sorting appears only once on page 7.
Thank you for pointing out this omission. We have updated the title, abstract, introduction, and methods to make "spike sorting" more central to the text. We highlight some of those changes below:
- **Title:** Electrophysiological Spike Sorting via Joint Dictionary Learning & Mixture Modeling
- **Abstract:** A new model is developed for feature learning and clustering of extracellular electrophysiological data across multiple recording periods for action potential detection and discrimination ("spike" sorting).
- **Introduction:** Spike sorting is used throughout our revised Introduction.

b. I think that spike sorting papers should follow the following outline...
a) A review of existing methods and what is the problem with them.
b) A piece of data that demonstrates the problem in a clear way.
c) An outline of the method developed.
d) A formal description of the method.
e) Evaluation of the method on artificial data or data with ground truth should be presented.
f) A comparison of the method with other methods which are simpler and provide good results in other systems.
Thank you for suggesting an improved outline. We have modified the outline of the text to reflect your suggestion. Please see the attached table of contents for details.

c. I do not think PCA can be the unique method to compare with as it has many pitfalls.
We agree with this comment whole-heartedly. For this reason, we have compared the performance of our method with many state of the art algorithms. For details, please see both Figures 1 and 8.

2) Minor Concerns:

a. The acronym ephys should be omitted. The use of this acronym is quite annoying.
We have striken "ephys" from the record.

b. I do not see how the word Forensic fits into this paper. It does not fit any of the dictionary definitions of the word.
Thank for pointing this out. We have replaced "forensic" with "longitudinal analysis" throughout.

c. The method section should describe the experiments in a proper way.
We have added a section entitled "Data Acquisition and Pre-processing" to the methods section.

d. The second to last paragraph of the introduction (In this paper) should be rewritten. It is too confusing in its present form. Too many buzzwords and very little information.
We have substantially re-written the introduction based largely on your suggestions. Note that we have added a subsection entitled, "Connection to Previous Bayesian Non-Parametrics" to highlight which paragraphs highlight relationship to previous spike sorting methods, and which highlight relationships to previous Bayesian non-parametric methods.

B. Response to reviewer 2

1) Main Concern:

a. **Overlapping Spikes** Traditional spike sorting methods fail to identify spikes from multiple neurons, when they overlap due to occurrence within a short time interval. It has already been reported that this failure may cause artificial correlations in brain areas with high firing rate or increased firing synchrony [2]. Recently, a number of different approaches have appeared in the spike sorting literature trying to tackle this problem [3-8].
We have added a section to the Discussion section of our manuscript directly addressing this concern. In brief, FMM does not elegantly handle overlapping spikes in its current incarnation, but we are actively pursuing such a generalization.

b. **Sparsely firing neurons** Very recently, the importance of the identification of this type of neurons (neurons with low probability of firing) and its limitations to contemporary algorithms has been highlighted in the spike sorting domain [9-10].
Thank you for this suggestion. We have now added a synthetic data analysis section to the manuscript devoted exclusively to stressing out the performance of our model in this sparse-firing regime.

2) Other Concerns:

a. **Neuronal bursting** Could the proposed model tackle the slight progressive changes in the spike waveforms due to neuronal bursting activity (well presented in [11])? If yes, it would be important to be stressed out.
We conjecture that our model might address neuronal bursting better than previously proposed methods. It is now part of our future extensions.

b. **Literature** The authors could enrich their literature references (mostly introduction section), in comparison to their corresponding conference paper.
We have beefed up our references in the introduction.

3) Minor Concerns:

a. (Page 2).. recent research indicates that a major portion of the information content related to neural spiking is carried in the spike rate , in terms of the number of spikes within a defined interval [6] (Page 4).. However, in the context of ephys data, the rate at which neurons fire plays an important role in information content [6]. Unless the authors concept is tailored to brain interfaces, it would be more appropriate to use a more standard reference for rate coding. See, for example, [11-12] and associated literature.

b. (Page 5)The DP-DL and HDP-DL results correspond to dictionary learning applied separately to each channel (from [5]), and the Matrix-DP (M-DP) and matrix-FMM (M-FMM) with the top 2 principle components without dictionary learning correspond to mixture models with the spikes observed simultaneously across all 4 channels, and the proposed model corresponds to joint dictionary learning all 4 channels, we compare DPDL and FMM based mixture modeling (here both models employ the proposed form of dictionary learning, with the differences manifested in how the mixture component of the model is performed). Difficult to follow. Please split to smaller sentences.

c. (Page 7) This highlights the need to allow modeling of different signal rates.. Do you mean neural firing rates?