Focused Mixture Modeling With Application to Electrophysiological Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

A new model is developed for feature learning and clustering of general data, with explicit analysis of the quantity (count) of data. The model is constituted in a "multi-task" setting, with model parameters shared across multiple data sets ("tasks"). The model is motivated by and applied to analysis of electrophysiological (ephys) data across multiple recording periods; here data from one recording period constitutes a task. It is demonstrated that joint feature (dictionary) learning and clustering allows one to perform forensics on the characteristics of the data (distinguishing single-unit spikes from non-local phenomena and artifacts). Modeling the count of data within a measurement interval addresses the time-evolving spike firing rate, and explicit modeling of the number of clusters mitigates limitations of methods like the Dirichlet process. Model properties are discussed, state-of-the-art results are presented on public data, and the methodology is demonstrated on new measured ephys data.

1 Introduction

Brain-machine interfaces utilize a sensor array to measure electrical (electrophysiological, or "ephys") activity within regions of the brain, with the ultimate goal of controlling robotic limbs [1] or muscles. When processing electrical signals from such a device, one typically (i) filters the raw sensor readings, (ii) performs thresholding to "detect" the spikes, (iii) maps each detected spike to a feature vector, and (iv) then clusters the feature vectors [14]. The complexities of real data, from moving animals, may significantly complicate the characteristics of the data, warranting reconsideration of aspects of this analysis chain.

Concerning the detection phase, not all signals that exceed a threshold are spikes (localized wavelet-like signals, henceforth also termed "single unit events"). There are biologically significant signals that may not be spikes, $e.g.$, local field potentials [6]. Some signals that excede the threshold may be due to artifacts; these may be manifested as indirect effects of movement on the recording apparatus, for example when a behaving animal collides with objects, grooms near the implant site or chews. These behaviors interfere with the electromechanical interface between the headstage (attached to the grecording cable) and implant assembly (anchored to dental acrylic on the animal's head). In addition, biological signals that occur near the reference electrode may also exceed threshold and we are especially interested in discriminating this activity from spiking. There is therefore a need for additional steps to distinguish spike-like and non-spike-like signals after step (ii), and to identify artifact signals.

Concerning feature extraction, step (iii), this is typically done prior to subsequent clustering, with principal components analysis (PCA) [14] and wavelets [13] representing popular methods. For PCA one must *a priori* select the number of principal components; while wavelets may be spike-like, they were not designed to be matched to ephys data (and as indicated above, not all signals that pass a threshold are spike-like). To infer the number of clusters based upon the observed data, mixture models have become increasingly popular, and nonparametric Bayesian methods have proven effective [8, 5]. Researchers have also recently employed mixture of factor analyzers to *jointly* perform feature extraction and clustering in a data-adaptive manner [9, 5], combining steps (iii) and (iv) above.

The use of nonparametric Bayesian methods like the Dirichlet process (DP) [8, 5] removes some of the *ad hoc* character of classical clustering methods, but there are other limitations within the context of ephys analysis. The DP and related models are characterized by a scale parameter $\alpha > 0$, and the number of clusters grows as $\mathcal{O}(\alpha \log M)$ [16], with M the number of data samples. This growth without limit in the number of clusters with increasing data is undesirable in the context of ephys analysis, for which there are typically a finite set of processes responsible for the observed data. Further, when jointly performing mixture modeling across multiple tasks, the *hierarchical* Dirichlet process (HDP) [17] shares all mixture components, which may undermine inference of subtly different clusters.

Another limitation of almost all existing ephys analysis 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 [6]. It is therefore not only desirable to model the clustering of the data, but also the rate of spike firing, ideally with these done jointly.

In this paper we integrate dictionary learning and clustering for analysis of ephys data, as in [9, 5]. However, rather than utilizing a method like DP or HDP [8, 5] for clustering, we develop a new hierarchical clustering model in which the number of clusters is modeled explicitly, and it does not grow with observed data; this implies that we model the number of underlying neural processes – 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) [10] with the Dirichlet distribution, similar to [19], but with unique characteristics. The framework explicitly models the quantity of data ($e.g.$, spikes) measured within a given recording interval. We believe that this is the first time the firing rate of ephys 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 from other signal sources, we demonstrate how this may be achieved using the proposed method, considering new measured ephys data.

2 Models and Analysis

2.1 Bayesian dictionary learning and clustering

Consider ephys data measured over a prescribed time interval. Specifically, let $\mathbf{X}_{ij} \in \mathbb{R}^{T\times N}$ represent the jth signal observed during interval i. 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 signal. In tetrode arrays [7], and related devices, a single-unit event ($e.g.$, 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 data forensics, discussed in Section 4.

Following [5], we employ dictionary learning to model each $\mathbf{X}_{ij}$; however, unlike [5] 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 $\mathbf{X}_{ij} = \mathbf{D}\mathbf{\Lambda}\mathbf{S}_{ij} + \mathbf{E}_{ij}$, where $\mathbf{D} \in \mathbb{R}^{T\times K}$ represents a dictionary, $\mathbf{\Lambda} \in \mathbb{R}^{K\times K}$ 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 $\boldsymbol{d}_k \in \mathbb{R}^T$ and $\boldsymbol{e}_n \in \mathbb{R}^T$. We impose the priors $\boldsymbol{d}_k \sim \mathcal{N}(0, \text{diag}(\gamma_1^{-1}, \ldots, \gamma_T^{-1}))$ and $\boldsymbol{e}_n \sim \mathcal{N}(0, \text{diag}(\eta_1^{-1}, \ldots, \eta_T^{-1}))$.

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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$. 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 (a_0, b_0) 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 $\{\gamma_t\}$, $\{\eta_t\}$, and α_0.

A mixture model is imposed for the dictionary weights $\mathbf{S}_{ij} = (\boldsymbol{s}_{ij1}, \ldots, \boldsymbol{s}_{ijN})$, with $\boldsymbol{s}_{ijn} \in \mathbb{R}^K$ (the indices j are clustered in the mixture model). Specifically,

$$\boldsymbol{s}_{ijn} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{ij}n}, \boldsymbol{\Omega}_{z_{ij}n}^{-1}) \ , \qquad z_{ij} \sim \sum_{m=1}^{M} \pi_m^{(i)} \delta_m \ , \qquad (\boldsymbol{\mu}_{mn}, \boldsymbol{\Omega}_{mn}) \sim G_0 \tag{1}$$

where G_0 is a normal-Wishart distribution, $\pi_m^{(i)} > 0$, $\sum_{m=1}^{M} \pi_m^{(i)} = 1$, and $\{\boldsymbol{s}_{ijn}\}_{n=1,N}$ are all associated with cluster z_{ij}.

A typical prior for the mixture weights $\boldsymbol{\pi}^{(i)} = (\pi_1^{(i)}, \ldots, \pi_M^{(i)})^T$ is a Dirichlet distribution [9], and in the limit $M \to \infty$ the Dirichlet process [8, 5]. Rather than drawing each $\boldsymbol{\pi}^{(i)}$ independently, we may consider the hierarchical Dirichlet process (HDP) [17], which imposes dependencies between $\{\boldsymbol{\pi}^{(i)}\}$, yielding a sharing of data across the multiple measurement intervals. Below we develop a framework that yields hierarchical clustering like HDP, but the number of clusters and the data count ($e.g.$, spike rate) are modeled explicitly.

2.2 Hierarchical count and mixture modeling

Let the total set of data measured during interval i be represented $\boldsymbol{\mathcal{D}}_i = \{\mathbf{X}_{ij}\}_{j=1,M_i}$. 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 $m \in \{1, \ldots, M\} = \mathcal{S}$, 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

$$\hat{\phi}_m^{(i)} \sim \text{Gamma}(\phi_m, p_i/(1-p_i)) \ , \quad b_m^{(i)} \sim \text{Bernoulli}(\nu_m) \ , \quad p_i \sim \text{Beta}(a_0, b_0) \tag{2}$$

$$\phi_m \sim \text{Gamma}(\gamma_0, 1) \ , \quad \nu_m \sim \text{Beta}(\alpha/M, 1) \ , \quad \gamma_0 \sim \text{Gamma}(c_0, 1/d_0) \tag{3}$$

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 intervals, like via HDP. However, here we explicitly model the number of clusters and quantity of data. Also note that $n_{im}^* = 0$ when $b_m^{(i)} = 0$, and therefore $\boldsymbol{b}^{(i)} = (b_1^{(i)}, \ldots, b_M^{(i)})^T$ defines indicator variables identifying the active subset of clusters $\mathcal{S}_i$ for representation of $\boldsymbol{\mathcal{D}}_i$. Marginalizing out $\hat{\phi}_m^{(i)}$, $n_{im}^* \sim \text{NegBin}(b_m^{(i)} \phi_m, p_i)$.

In (3) we simply drew $\nu_m \sim \text{Beta}(\alpha/M, 1)$. In practice one often truncates M to a large value (we do that in our computations). In this context it is preferable to associate large-valued ν_m with small indices m, so only negligibly small ν_m are discarded after truncation. We therefore employ the stick-breaking representation for ν_m [18]: $\nu_m = \prod_{l=1}^{m} \tilde{\nu}_l$, with $\tilde{\nu}_l \sim \text{Beta}(\alpha, 1)$.

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 [21] for a proof of equivalence): first draw $M_i \sim \text{Poisson}(\sum_{m=1}^{M} b_m^{(i)} \hat{\phi}_m^{(i)})$, and then

$$(n_{i1}^*, \ldots, n_{iM}^*) \sim \text{Mult}(M_i; \pi_1^{(i)}, \ldots, \pi_M^{(i)}) \ , \qquad \pi_m^{(i)} = b_m^{(i)} \hat{\phi}_m^{(i)} / \sum_{m'=1}^{M} b_{m'}^{(i)} \hat{\phi}_{m'}^{(i)} \tag{4}$$

with $\hat{\phi}_m^{(i)}$, $\{\phi_m\}$, $\{b_m^{(i)}\}$, and $\{p_i\}$ constituted as in (2)-(3). Note that we have $M_i \sim \text{NegBin}(\sum_{m=1}^{M} b_m^{(i)} \phi_m, p_i)$ by marginalizing out $\hat{\phi}_m^{(i)}$.

Rather than drawing $(n_{i1}^*, \ldots, n_{iM}^*) \sim \text{Mult}(M_i; \pi_1^{(i)}, \ldots, \pi_M^{(i)})$, 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 $j \in \{1, \ldots, M_i\}$ 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 (4) is now used within the mixture model in (1).

In the context of modeling and analyzing ephys data, recent work on clustering models has accounted for refractory-time violations [8, 5], 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, 5] 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 ephys data are spike related (which are associated with refractory-time issues). As demonstrated in Section 4, a key component of the proposed method is that it allows us to distinguish single-unit (spike) events from other phenomena.

2.3 Relationship to existing models

By the definition of the Dirichlet distribution, for *any* $p_i \in (0, 1)$, we may also write $\boldsymbol{\pi}^{(i)} \sim \text{Dir}(b_1^{(i)}\phi_1, \ldots, b_M^{(i)}\phi_M)$. Considering the limit $M \to \infty$, and upon marginalizing out the $\{\nu_m\}$, the binary vectors $\{\boldsymbol{b}^{(i)}\}$ may be drawn via 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), and therefore for finite α the number of non-zero components in $\boldsymbol{b}^{(i)}$ is finite, even when $M \to \infty$. Consequently $\text{Dir}(b_1^{(i)}\phi_1, \ldots, b_M^{(i)}\phi_M)$ is well defined even when $M \to \infty$ 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 [19], 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 0.5 as in [19], 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 ephys data, the rate at which neurons fire plays an important role in information content [6]. 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 ephys data, and likely for other clustering problems of this type.

The proposed model is also related to the beta-gamma-gamma-Poisson factor model in [21]. However, the data and application considered in [21] is distinct from that addressed here.

In an HDP [17] representation for data of this type, each interval i is characterized by a probability measure $G_i \sim \text{DP}(\hat{\alpha}_1 G)$ and $G \sim \text{DP}(\hat{\alpha}_0 G_0)$. We have $G = \sum_{m=1}^{\infty} \hat{\pi}_m \delta_{\omega_m}$, where ω_m corresponds to parameters that represent the mth mixture component (here characterized by $\{\boldsymbol{\mu}_{mn}, \boldsymbol{\Omega}_{mn}\}_{n=1,N}$). The parameters $\{\hat{\pi}_m\}$ characterize the "global popularity" of dish/cluster across all intervals.

To make a connection between the proposed model and the HDP, motivated by (2)-(3), 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 $\phi_m \sim \text{Gamma}(\gamma_0, 1)$, we may consider drawing $\tilde{\alpha}_1 \sim \text{Gamma}(M\gamma_0, 1)$, 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

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

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 $M \to \infty$, with $\boldsymbol{b}^{(i)}$ all ones, this corresponds to the HDP, with $\hat{\alpha}_1 \sim \text{Gamma}(\hat{\alpha}_0, 1)$. 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 $\phi_m \sim \text{Gamma}(\gamma_0, 1)$, 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 4, this focusing nature of the proposed model, which we term a focused mixture model (FMM), is important in the context of ephys data, and likely other applications.

3 Computations

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. 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 3.1. *Denote $s(n, j)$ as the Sterling numbers of the first kind [12] 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 $1 \le j \le n$. Assuming $n \sim \text{NegBin}(\phi, p)$ is a negative binomial distributed random variable, and it is augmented into a compound Poisson representation [2] as*

$$n = \sum_{l=1}^{\ell} u_l, \; u_l \sim \text{Log}(p), \; \ell \sim \text{Pois}(-\phi \ln(1-p)) \tag{5}$$

where $\text{Log}(p)$ is the logarithmic distribution [2] with probability generating function $G(z) = \ln(1-pz)/\ln(1-p)$, $|z| < p^{-1}$, then we have

$$\Pr(\ell = j | n, \phi) = R_\phi (n, j) = F(n, j) \phi^j \Big/ \sum_{j'=1}^{n} F(n, j') \phi^{j'}, \;\; j = 0, 1, \cdots, n. \tag{6}$$

The proof is provided in Supplementary Material.

Concerning sampling ϕ_m, since $\phi_m \propto \prod_{i: b_m^{(i)}=1} \text{NegBin}(n_{im}^*; \phi_m, p_i) \text{Gamma}(\phi_m; \gamma_0, 1)$, using Lemma 3.1, we can first sample a latent count variable ℓ_{im} for each n_{im}^* as

$$\Pr(\ell_{im} = l | n_{im}^*, \phi_m) = R_{\phi_m}(n_{im}^*, l), \;\; l = 0, \cdots, n_{im}^*. \tag{7}$$

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

$$\phi_m | \{\ell_{im}, b_m^{(i)}, p_i\} \sim \text{Gamma}\left(\gamma_0 + \sum_{i: b_m^{(i)}=1} \ell_{im}, \frac{1}{1 - \sum_{i: b_m^{(i)}=1} \ln(1-p_i)} \right). \tag{8}$$

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

$$\Pr(\tilde{\ell}_{im} = l | \ell_{im}, \gamma_0) = R_{\gamma_0}(\ell_{im}, l), \;\; l = 0, \cdots, \ell_{im} \tag{9}$$

$$\gamma_0 | \{\tilde{\ell}_{im}, b_m^{(i)}, p_i\} \sim \text{Gamma}\left(c_0 + \textstyle\sum_{i: b_m^{(i)}=1} \tilde{\ell}_{im}, \frac{1}{d_0 - \sum_{i: b_m^{(i)}=1} \ln\left(1 - \frac{-\ln(1-p_i)}{1-\ln(1-p_i)}\right)} \right). \tag{10}$$

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

$$b_m^{(i)} | \pi_m, n_{im}^*, \phi_m, p_i \sim \text{Bernoulli}\left(\delta(n_{im}^* = 0) \frac{\pi_m (1-p_i)^{\phi_m}}{\pi_m (1-p_i)^{\phi_m} + (1-\pi_m)} + \delta(n_{im}^* > 0) \right). \tag{11}$$

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.

Table 1: *Results from testing on d533101 data [11]. KFM represent Kalman Filter Mixture method [4]. All results, except the two results of the proposed method, are taken from [5]. 2 PCs denotes using the top 2 principle components, and results were indistinguishable from using the top 3 principle components. For the proposed model, dictionary learning was done as in Sec. 2.1, and "FMM" corresponds to the focused mixture-model of Sec. 2.2*

K-means	GMM	K-means w/ 2 PCs	GMM w/ 2 PCs	KFM w/ 2 PCs	DP w/ 2 PCs	HDP w/ 2 PCs	DP-DL	HDP-DL	Proposed w/ DP	Proposed w/ FMM
88.08%	87.42%	88.44%	89.04%	88.36%	88.07%	89.54%	91.89%	93.05%	94.4%	93.7%

4 Results

For these experiments we used a truncation level of $K = 40$ dictionary elements, and the number of mixture components was truncated to $M = 20$. In dictionary learning, the gamma priors of $\{\gamma_t\}$, $\{\eta_t\}$ and α_0 were set as Gamma$(10^{-6}, 10^{-6})$. In the context of the hierarchical count and mixture modeling, we set $a_0 = b_0 = 1$, $c_0 = 0.1$ and $d_0 = 0.1$. Prior Gamma$(10^{-6}, 10^{-6})$ was placed on parameter α related to the IBP. 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.

4.1 Real data with partial ground truth

We first consider publicly available dataset[1] hc-1. These data consist of both extracellular recordings and an intracellular recording from a nearby neuron in the hippocampus of an anesthetized rat [11]. Intracellular recordings give clean signals on a spike train from a specific neuron, giving accurate spike times for that neuron. Thus, if we detect a spike in a nearby extracellular recording within a close time period (<.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 the unknown cluster.

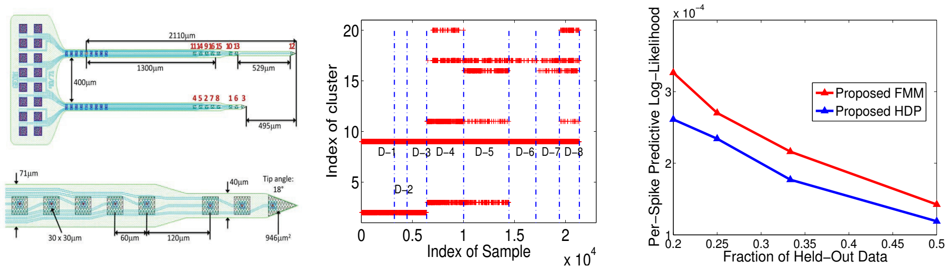
We considered the widely used data d533101 and the same preprocessing from [4]. This data consists of a 4-channel extracellular recordings and 1-channel intracellular recording. We used 2491 detected spikes and 786 of those spikes came from the known neuron. The results are shown in Table 1. The DP-DL and HDP-DL results correspond to dictionary learning applied separately to each channel (from [5]), and the proposed model corresponds to joint dictionary learning on all 4 channels. These data are relatively simple, with two clear mixture components and with the spikes observed simultaneously across all 4 channels; the DP-based and focused mixture model (FMM) form of the model therefore yield similar results, with the gain in these results relative to DP-DL and HDP-DL deemed manifested as a result of joint dictionary learning across all channels.

4.2 Forensic analysis of new longitudinal ephys data

The next dataset is new, 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 1(a)) were humanely placed in the motor cortex, and ephys 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 2.6 msec window, which corresponds to $T = 80$ 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 1(a).

[1] available from http://crcns.org/data-sets/hc/hc-1

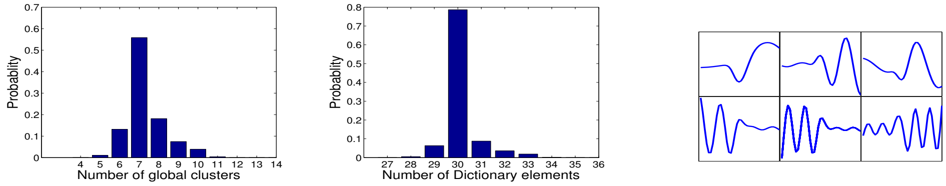


(a) (b) (c)

Figure 1: *(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. (c) Predictive likelihood of held-out data. The horizontal axis represents the fraction of data held out during training.*

In Figure 1(b) are shown assignments of data to each of the possible clusters, for data measured across the 8 days, as computed by the proposed model ($e.g.$, for the first three days, two clusters were inferred). Results are shown for the maximum-likelihood collection sample. Rather than employing the proposed focused mixture model (FMM) of Section 2.2, we also considered the simplified HDP construction discussed in Section 2.3, with the $\boldsymbol{b}^{(i)}$ set to all ones (in both cases we employ the same form of dictionary learning, as in Section 2.1). From Figure 1(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 forensic analysis of ephys data, as a consequence of observed limitations of HDP (which are only partially illuminated by Figure 1(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 [19] 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-based model works well when observing data from one recording session, like in Table 1, but the analysis of multiple days of data is challenging for HDP).



(a) (b) (c)

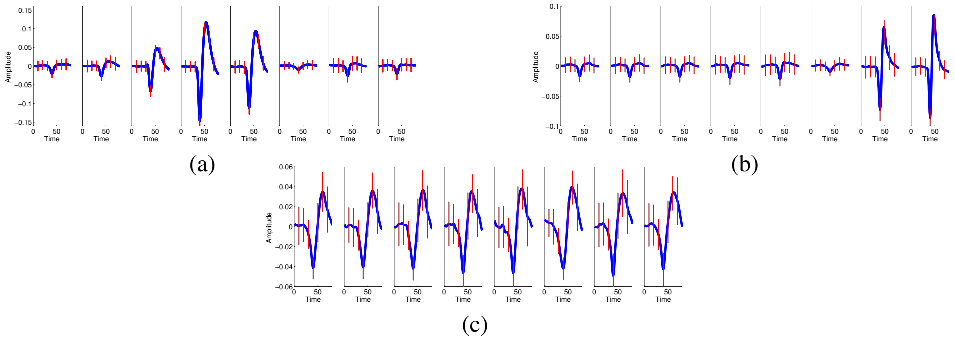
Figure 2: *(a) Approximate posterior distribution on the number of global clusters (mixture components). (b) Approximate posterior distribution on the number of required dictionary elements. (c) Example inferred dictionary elements.*

Note from Figure 1(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 2(a)-2(b), and in Figure 2(c) are shown 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, in Figure 3 are depicted the inferred properties of three of the clusters, from Day 4 (D-4 in Figure 1(b)). Shown are the *mean* signal for the 8 channels in the respective cluster (for the 8 channels at the bottom of Figure 1(a)), and the error bars represent one standard deviation, as reflected in the posterior. Note that the cluster in Figure 3(a) 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 Figure 3(b) 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 1(a). The HDP model uncovered the cluster in Figure 3(a), but not that in Figure 3(b).

Note Figure 3(c), 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 Figures 3(a)-3(b) the error bars are relatively tight with respect to the strong signals in the set of eight, while the error bars in Figure 3(c) are more pronounced (the mean curves look clean, but this is based upon averaging thousands of signals).



(a)

(b)

(c)

Figure 3: *Example clusters inferred for data on the bottom 8 channels of Fig. 1(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 1(a). The blue curves represent the mean, and the error bars are one standard deviation.*

In addition to recording the ephys data, video was recorded of the rat throughout. Robust PCA [20] 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 required 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. This suggests that animal motion causes signal artifacts, as discussed in Section 1

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 [3, 15]. 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.

5 Conclusions

A new focused mixture model (FMM) has been developed, motivated by real-world studies with longitudinal ephys 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. The model explicitly models the count of signals within a recording period. The rate of neuron firing constitutes a primary information source [6], and therefore it is desirable that it be modeled. This rate is controlled here by a parameter p_i, and this was allowed to be unique for each recording period i. In future research one may constitute a mixture model on p_i, with each mixture component reflective of a latent neural (firing) state. Inference of this state could be important for decoding neural signals and controlling external devices or muscles.

References

[1] A. Abbott. Mind-controlled robot arms show promise. *Nature*, 2012.

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

[3] 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.

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

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

[6] 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.

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

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

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

[10] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

[11] 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.

[12] N.L. Johnson, A.W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

[13] J. C. Letelier and P. P. Weber. Spike sorting based on discrete wavelet transform coefficients. *J. Neuroscience Methods*, 2000.

[14] M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 1998.

[15] 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 micro-machined silicon devices. *Brain Res.*, 2003.

[16] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *J. Am. Stat. Ass.*, 2006.

[18] Y.W. Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction for the indian buffet process. In *AISTATS*, 2007.

[19] 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.

[20] 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.

[21] M. Zhou, L.A. Hannah, D.B. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.