BAYESIAN BI-CLUSTERING OF NEURAL SPIKING ACTIVITY WITH LATENT STRUCTURES

Ganchao Wei
Department of Statistical Science
Duke University
Durham, NC 27708, USA
ganchao.wei@duke.edu

ABSTRACT

Modern neural recording techniques allow neuroscientists to obtain spiking activity of multiple neurons from different brain regions over long time periods, which requires new statistical methods to be developed for understanding structure of the large-scale data. In this paper, we develop a bi-clustering method to cluster the neural spiking activity spatially and temporally, according to their low-dimensional latent structures. The spatial (neuron) clusters are defined by the latent trajectories within each neural population, while the temporal (state) clusters are defined by (populationally) synchronous local linear dynamics shared with different periods. To flexibly extract the bi-clustering structure, we build the model non-parametrically, and develop an efficient Markov chain Monte Carlo (MCMC) algorithm to sample the posterior distributions of model parameters. Validating our proposed MCMC algorithm through simulations, we find the method can recover unknown parameters and true bi-clustering structures successfully. We then apply the proposed bi-clustering method to multi-regional neural recordings under different experiment settings, where we find that simultaneously considering latent trajectories and spatial-temporal clustering structures can provide us with a more accurate and interpretable result. Overall, the proposed method provides scientific insights for large-scale (counting) time series with elongated recording periods, and it can potentially have application beyond neuroscience.

1 INTRODUCTION

In neuroscience, identifying types of neurons is a longstanding challenge (Nelson et al., 2006; Bota & Swanson, 2007; Zeng, 2022). Some criteria based on features such as anatomical regions, genomics and synaptic connectivity have been proposed, and there are some Bayesian approaches to integrate these features (Jonas & Kording, 2015). On the other hand, the response pattern and interactions between neural populations may change over time, especially when the experiment stimuli changes (Pooresmaeili et al., 2014; Oemisch et al., 2015; Ruff & Cohen, 2016; Steinmetz et al., 2019; Cowley et al., 2020). However, these complex dynamical observations can often be broken down into simpler units, and it can be appropriate to assume static linear dynamics within chunks of the data. Moreover, it is usually appropriate and helpful to assume similar linear dynamics can be shared by different epochs (Zoltowski et al., 2020; Glaser et al., 2020b). Here, we consider the problem of how to identify both spatial and temporal clusters of neural spikes.

The modern techniques such as the high-density probes (Jun et al., 2017; Steinmetz et al., 2021; Marshall et al., 2022) allow us to obtain large-scale multi-electrode recordings from multiple neurons across different anatomical regions over an elongated session. Several models have been developed to extract shared latent structures from simultaneous neural recordings, assuming that the activity of all recorded neurons can be described through common low-dimensional latent states (Cunningham & Yu, 2014; Gao et al., 2017). These approaches have proven useful in summarizing and interpreting high-dimensional population activity. Inferred low-dimensional latent states can provide insight into the representation of task variables (Churchland et al., 2012; Mante et al., 2013; Cunningham & Yu, 2014; Saxena & Cunningham, 2019) and dynamics of the population itself (Vyas et al., 2020). Many existing approaches are based on linear dynamical system (LDS) model (Macke et al., 2011),

which is built on the state-space model and assumes latent factors evolve with static linear dynamics. Although assuming static linear dynamics over time can be valid in some tasks and in small chunks of experiment, the assumption is not generally appropriate. To tackle the nonlinear dynamics, some variants of the LDS, such as switching-LDS (SLDS, Ghahramani & Hinton (2000); Fox (2009); Fox et al. (2008a); Murphy (2012)) and recurrent-SLDS (RSLDS, Linderman et al. (2017; 2019)) have been proposed. The non-parametric Gaussian process factor analysis (GPFA) model (Yu et al., 2009) and its variants provide a more flexible way to model non-linaer neural data, although most these methods assume independent GP and doesn't allow for interactions between latent factors. Recently, (Cai et al., 2023) proposed the dependent GP method using the kernel convolution framework (KCF, Boyle & Frean (2004); Álvarez & Lawrence (2011); Sofro et al. (2017)), but their method may not scalable for elongated neural recordings. Several methods have been developed and implemented to analyze multiple neural populations and their interactions (Semedo et al., 2019; Glaser et al., 2020b), as the interactions may occur in low-dimensional subspace (Stavisky et al., 2017; Kaufman et al., 2014; Semedo et al., 2019). But the neural populations are prespecified, and the spatio-temporal clustering structures of neural data haven't been evaluated systematically in general.

The clustering of neural spikes is an important and long-lasting problem, and people put a lot effort into developing methods to uncover patterns from the complex data. The neural spiking activity is essentially a point process, and there are some methods for finding clusters in point process by, such as, Dirichlet mixture of Hawkes process (Xu & Zha, 2017), mixture of multi-level point process and (Yin et al., 2021) and group network Hawkes process (Guanhua Fang & Guan, 2023). All these methods are general and have wide rage of applications, but the defined "clusters" may not directly related to underlying mechanisms and may lose some scientific insights. To deal with this problem, some methods like mixPLDS (Buesing et al., 2014) and recent mixDPFA method (Wei et al., 2022; 2023) try to cluster neurons according to their latent structures, by using the mixture of LDS model. These approach provides a more interpretable and accurate way to clusters neurons, and may be useful for identifying "functional populations" of neurons. However, these methods assume the static linear dynamics and don't allow for the interactions between neural populations, which can limit the usage of these methods, and may bias or even fail the detection of neural populations when considering the elongated recordings, especially under different experiment conditions. On the other hand, for the clustering structures in terms of temporal states, most methods are developed based on the SLDS, by modeling the nonlinear dynamics with local linear pattern. Instead of clustering based on linear dynamics, D'Angelo et al. (2023) recently tried to cluster the experiment periods based on the distributions of spiking amplitude, using a nested formulation of mixture of finite mixtures model (MFMM), i.e., exploiting the generalized MFMM (gFMFMM, Frühwirth-Schnatter et al. (2021)) prior with common atom model (Denti et al., 2023).

In this research, we develop a bi-clustering method to cluster neural spikes both spatially (to give subject clusters) and temporally (to give state clusters), according to the latent structures of these neurons (Figure 1A). The neural population is defined via private low-dimensional latent trajectories as in mixPLDS (Buesing et al., 2014) or mixDPFA (Wei et al., 2022; 2023). For the state clusters, we assume the linear dynamics can be shared across different chunks and the state clustering structures are defined by local linear manner as in (R)SLDS. Neurons in each population is assume to have private latent trajectories, but all time series are assumed to switch between different states synchronously, to use the information from all observations. Besides extending the previous clustering method like mixDPFA to simultaneously detect the state cluster, the proposed bi-clustering method also allow for interactions between neural populations and non-stationary dynamics for neural response, using similar idea from (Glaser et al., 2020a). Simultaneously considering all these effects in the proposed bi-clustering method is necessary, since incorrect population assignments can lead to biased and inconsistent inference on the latent structure (Ventura, 2009). On the other hand, these flexibility allows for more accurate estimate of latent trajectories, and hence will lead to a more accurate estimates of the subject clustering structure.

To flexibly infer the bi-clustering structure, we model them non-parametrically to avoid prespecifying the number for subject and state clusters. Specifically, the subject clustering structure is modeled by a mixture of finite mixtures model (MFMM, Miller & Harrison (2018)) of latent trajectories and the state clustering structure is modeled by a sticky Hierarchical Dirichlet Process Hidden Markov Model (sticky- HDP-HMM, Fox et al. (2008b)). The posteriors of model parameters are sampled using a Markov Chain Monte Carlo (MCMC) algorithm, where the Pólya-Gamma data augmentation technique (Polson et al., 2013) is used to handle the counting observations for neural spikes.

The rest of this paper is structured as follows. In Section 2, we introduce the bi-clustering method for time series with counting observations, and provide brief explanations of the MCMC algorithm to sample posterior distributions of parameters. After validating the proposed bi-clustering method with a synthetic dataset in Section 3, we then apply our method to analyze multi-regional experimental recordings from a behaving mouse under different experiment settings in Section 4. Finally, in Section 5, we conclude with some final remarks and highlight some potential extensions of our current model for future research.

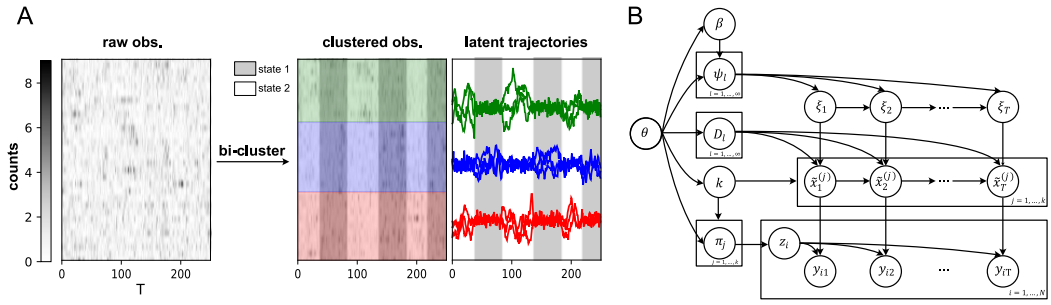

Figure 1: **Model overview. A**. The goal for our proposed model is to do clustering both spatially and temporally (i.e. "bi-clustering") for neural spike data (time series data with counting observations), according to their latent structures. The neural spiking counts are determined by a low dimensional latent factors, specific to the spatially subject clustering assignment (e.g. green ,blue and red). On the other hand, all neurons are assumed to switch between states synchronously, and are temporally clustered according to different states of linear dynamics (e.g. gray and white). **B**. Graphical model of the proposed bi-clustering model. All prior parameters are summarized as θ, and parameters such as d_i and $\boldsymbol{c}_i$ dropped for simplicity.

2 Bi-clustering Method for Neural Spikes

In this section, we introduce a bi-clustering model for neural spiking activity, i.e., the time series data with counting observations. The goal for the proposed model is to cluster neural spikes both spatially (for subject cluster) and temporally (for state cluster), based on the multi-population and -state latent structures. To flexibly capture the clustering structures, we build model non-parametrically. The graphical representation of the model is summarized in Figure 1B. After introducing the model, we briefly describe how a MCMC algorithm is used to infer model parameters.

2.1 Multi-population and -state Linear Dynamic Model

Assume we can observe spiking activity of N neurons up to recording length T. Denote the number of counts for neuron $i \in \{1, \ldots, N\}$ at time bin $t \in \{1, \ldots, T\}$ as $y_{it} \in \mathbb{Z}_{\geq 0}$, and the cluster indicator of neuron i as z_i (i.e. the "subject indicator"). Assume the count y_{it} follows a negative-binomial distribution, where the log-mean response is modeled by a linear combination of subject baseline d_i, population baseline $\mu_t^{(z_i)}$ and $p-$dimensional latent factor $\boldsymbol{x}_t^{(z_i)} \in \mathbb{R}^p$ (Here we assume all populations have the same latent dimension for convenience). In other words, the observation equation is as follows:

$$\begin{aligned} y_{it} &\sim \text{NB}(r_i, \mu_{it}) \\ \log \mu_{it} &= d_i + \mu_t^{(z_i)} + \boldsymbol{c}_i' \boldsymbol{x}_t^{(z_i)} \end{aligned} \tag{1}$$

, where $\text{NB}(r, \mu)$ denotes the negative-binomial distribution (NB) with mean μ and variance $\mu + \mu^2/r$, and $\boldsymbol{c}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. The NB distribution can be replaced by a Poisson distribution when it's appropriate to assume equi-dispersion, for ease of model inference. If we further denote $\tilde{\boldsymbol{x}}_t^{(j)} = (\mu_t^{(j)}, \boldsymbol{x}_t'^{(j)})'$ and $\tilde{\boldsymbol{c}}_i = (1, \boldsymbol{c}_i')'$, then $\log \mu_{it} = d_i + \tilde{\boldsymbol{c}}_i' \tilde{\boldsymbol{x}}_t^{(z_i)}$. To save words and notations, if not specified, we refer $\tilde{\boldsymbol{x}}_t^{(j)}$ as "latent factor" for cluster j, which also includes the population baseline.

Although each neural population is modeled with private latent factors, there usually exist some interactions between clusters (Musall et al., 2019; Stringer et al., 2019), and these interactions can change over time (Ruff & Cohen, 2016; Steinmetz et al., 2019; Cowley et al., 2020), especially when the external condition changes. On the other hand, interactions between neural populations and receiving common inputs for all neurons suggest that neurons in different clusters may synchronize the response states over time. Therefore, to allow for the interactions between populations and model the synchronous state switching, we stack the latent factors for all clusters together, and assume all latent factors evolve in a conditional linear manner, given the discrete latent states ξ_t shared across the cluster, as in Glaser et al. (2020b). In other words, the state clustering structure is defined by the local linear dynamics of latent factors, by assuming complex dynamics can be decomposed into simple linear unit and the small chunks of the neural response can be sufficiently described by the LDS model. Specifically, assume there are k unique clusters, i.e., $|\{z_i\}_{i=1}^N| = k$, the cluster-stacked latent factors (including population baseline) is denoted as $\tilde{\boldsymbol{X}}_t = (\tilde{\boldsymbol{x}}_t'^{(1)}, \ldots, \tilde{\boldsymbol{x}}_t'^{(k)})' \in \mathbb{R}^{k(p+1)}$. To capture temporal dynamics of the data, we further put AR(1) structure onto the latent factors $\tilde{\boldsymbol{X}}_t$. In other words, given the discrete latent state at t as ξ_t (i.e. the "state indicator" shared across the subject cluster), $\tilde{\boldsymbol{X}}_t$ is assumed evolve linearly with a Gaussian noise as follows:

$$\tilde{\boldsymbol{X}}_{t+1} = \boldsymbol{b}_{\xi_t} + \boldsymbol{A}_{\xi_t}\tilde{\boldsymbol{X}}_t + \boldsymbol{\epsilon}_{\xi_t} \tag{2}$$

, where $\boldsymbol{\epsilon}_{\xi_t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_{\xi_t})$ and $\tilde{\boldsymbol{X}}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{k(p+1)})$. Here, the dynamic parameters $(\boldsymbol{b}_{\xi_t}, \boldsymbol{A}_{\xi_t}, \boldsymbol{Q}_{\xi_t})$ summarize dynamics (state changes) within and across population. To make the model identifiable, we further assume 1) $\tilde{\boldsymbol{x}}^{(j)}$ is zero-centered, i.e. $\sum_{t=1}^T \tilde{\boldsymbol{x}}_t^{(j)} = \boldsymbol{0}$ and 2) $\tilde{\boldsymbol{x}}_{1:T}^{(j)}\tilde{\boldsymbol{x}}_{1:T}'^{(j)}$ is diagonal, where $\tilde{\boldsymbol{x}}_{1:T}^{(j)} = (\tilde{\boldsymbol{x}}_1^{(j)}, \ldots, \tilde{\boldsymbol{x}}_T^{(j)}) \in \mathbb{R}^{(p+1)\times T}$ (Fokoué & Titterington, 2003). The identifiability issue under signed-permutation is handled by alignment to samples in early stage of the chain. For more details on model constraints, see Section A.1. In summary, given the neuron i belonging to cluster $z_i = j$, the counting series is generated by a negative-binomial linear model $\mathcal{M}$ defined in equation 1 as $(y_{i1}, \ldots, y_{iT})' \sim \mathcal{M}(d_i, \boldsymbol{c}_i, \tilde{\boldsymbol{x}}_{1:T}^{(j)})$, where the prior for $\tilde{\boldsymbol{x}}_{1:T}^{(j)}$ is denoted as $\mathcal{H}$. The within- and between-population linear dynamics at $t-$th step are captured by dynamical parameters $(\boldsymbol{b}_{\xi_t}, \boldsymbol{A}_{\xi_t}, \boldsymbol{Q}_{\xi_t})$, where ξ_t is the state indicator at t. To do clustering both spatially (subject cluster) and temporally (state cluster) in a flexible way, we model each of these two clustering structures non-parametrically as follows.

2.2 Subject Clustering Model

Since the number of neural populations is finite but unknown, we put prior on number of subject cluster $|\{z_i\}_{i=1}^N| = k$ as in (Wei et al., 2023), which leads to the mixture of the finite mixtures model (MFMM) as follows:

$$\begin{aligned}
K &\sim f_k, && f_k \text{ is a p.m.f. on} \{1, 2, \ldots\}, \\
\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k) &\sim \text{Dir}_k(\gamma, \ldots, \gamma) && \text{given } K = k, \\
z_1, \ldots, z_N &\overset{i.i.d.}{\sim} \boldsymbol{\pi} && \text{given } \boldsymbol{\pi}, \\
\tilde{\boldsymbol{x}}_{1:T}^{(1)}, \ldots, \tilde{\boldsymbol{x}}_{1:T}^{(k)} &\overset{i.i.d.}{\sim} \mathcal{H} && \text{given } k, \\
(y_{i1}, \ldots, y_{iT})' &\sim \mathcal{M}(d_i, \boldsymbol{c}_i, \tilde{\boldsymbol{x}}_{1:T}^{(z_i)}) && \text{given } d_i, \boldsymbol{c}_i, \tilde{\boldsymbol{x}}_{1:T}^{(z_i)}, z_i, \text{for } i = 1, \ldots, N,
\end{aligned} \tag{3}$$

, where p.m.f denotes the probability mass function. By using the MFMM, we can integrate the field knowledge about the number of clusters into our analysis, by specifying the f_k. In the analysis of this paper, we assume k follows a geometric distribution, i.e., $k \sim \text{Geometric}(\zeta)$ with p.m.f. defined as $f_k(k \mid \zeta) = (1-\zeta)^{k-1}\zeta$ for $k = 1, 2, \ldots$, and $\gamma = 1$. For general use of the proposed method to some problems where the number of subject cluster number can potentially grow to infinity, using the mixture model such as the Dirichlet process mixtures model (DPMM) maybe conceptually more appropriate. See Miller & Harrison (2018) for more detailed discussion.

2.3 State Clustering Model

For state clustering structure, as the number of states can potentially shoot to infinity, we model the discrete state ξ_t by a sticky Hierarchical Dirichlet Process Hidden Markov Model (sticky-HDP-

HMM) proposed by (Fox et al., 2008b) to encourage the re-occurrence of states as follows:

$$
\begin{aligned}
\beta &\sim \text{GEM}(\eta), \\
\psi_l &\overset{i.i.d.}{\sim} \text{DP}(\alpha + m, \frac{\alpha\beta + m\delta_l}{\alpha + m}) \\
\xi_t &\sim \psi_{\xi_{t-1}}, \\
(\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l) &\overset{i.i.d.}{\sim} \mathcal{S}, && \text{for } l = 1, 2, \ldots, \\
\tilde{\boldsymbol{X}}_{t+1} &\sim \mathcal{N}(\boldsymbol{b}_{\xi_t} + \boldsymbol{A}_{\xi_t}\tilde{\boldsymbol{X}}_t, \boldsymbol{Q}_{\xi_t}) && \text{for } t = 1, \ldots, T
\end{aligned}
\tag{4}
$$

, where GEM denotes the stick breaking process (Sethuraman, 1994), δ_i denotes the indicator function at index i, DP denotes the Dirichlet process and $\mathcal{S}$ denotes the normal-inverse-Wishart prior of $(\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l)$. The details of $\mathcal{S}$ can be found in the appendix A.1, when introducing the MCMC algorithm. The sticky HDP-HMM extends the HDP-HMM with a "sticky" parameter $m > 0$ to encourage longer state duration, and hence can handle the rapid-switching problem to some degree. Some more careful methods for modeling the state duration and state transition is further discussed in the Section 5.

2.4 Model Inference

We do Bayesian inference on the proposed bi-clustering model by an efficient MCMC algorithm. In each sampling iteration, there are approximately four steps: 1) sample dynamical latent factors $\tilde{\boldsymbol{x}}_{1:T}^{(z_i)}$, 2) sample remaining subject-specific parameters in observation equation equation 1, including subject baseline d_i, factor loading $\tilde{\boldsymbol{c}}_i$ and dispersion r_i, 3) sample the temporal states ξ_t and corresponding dynamical parameters $(\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l)$ for each sampled state, and 4) sample the subject cluster indices z_i. The details of sampling procedures can be found in the appendix Section A.1, and we briefly introduce the key sampling method for each step here.

In step 1), the full conditional distribution of latent factors $\tilde{\boldsymbol{x}}_{1:T}^{(j)}$ is equivalent to the posterior distribution of the negative-binomial dynamic GLM (NB-DGLM), which has no closed form. However, the NB distribution falls within a Pólya-Gamma (PG) augmentation scheme (Polson et al., 2013; Windle et al., 2013; Linderman et al., 2016), therefore we can sample them in closed form by introducing the PG augmented variables. Conditioning on the auxiliary variables ω_{it}, the transformed "effective" observations $\hat{y}_{it}$ has Gaussian likelihood, and hence we can sample the posterior of $\tilde{\boldsymbol{x}}_{1:T}^{(j)}$ using the forward-filtering-backward-sampling (FFBS, Carter & Kohn (1994); Frühwirth-Schnatter (1994)) algorithm. For Poisson observation model, we can treat the data as coming from the NB distribution, use the samples as proposal and add one more Metropolis-Hasting (MH) step to accept or reject the proposal. In Poisson case, the dispersion r_i becomes the tuning parameter to achieve desirable acceptance rate (Wei et al., 2022; 2023).

In step 2), the sampling of d_i and $\tilde{\boldsymbol{c}}_i$ is regular NB regression problem, and we again use the PG data augmentation technique to sample them. The dispersion parameter r_i is updated via a Gibbs sampler, using the method described in Zhou et al. (2012), as the gamma distribution is the conjugate prior to the r_i under the compound Poisson representation.

In step 3), the discrete states ξ_t are sampled by a weak-limit Gibbs sampler for sticky HDP-HMM as in Fox et al. (2008b). The weak-limit sampler constructs a finite approximation to the HDP transitions prior with finite Dirichlet distributions, as the infinite limit converges in distribution to a true HDP. Given the latent factors $\tilde{\boldsymbol{x}}_{1:T}^{(j)}$ and state indicator ξ_t, we can update dynamical parameters $(\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l)$ for each state separately in closed form.

In step 4), given the parameters in observation equation equation 1, we sample the subject cluster indices z_i using the algorithm for MFMM proposed by Miller & Harrison (2018), which is analogous to the partition-based algorithm for DPMM (Neal, 2000). The label switching issue of z_i is handled by the Equivalence Classes Representatives (ECR) algorithm (Papastamoulis & Iliopoulos, 2010), by using the sample from early stage of the chain as pivot allocation. When sampling the subject clustering assignments z_i in such a high dimensional time series data with large T, evaluating the full likelihood given samples of $\boldsymbol{c}_i$ as in Gaussian MFA (Fokoué & Titterington, 2003) leads to a poor mixing chain. Instead, we evaluate the marginalized likelihood by integrating out the subject-

specific loading $\boldsymbol{c}_i$, similar to Wei et al. (2023). The marginalized likelihood is evaluated by Laplace approximation.

The Python implementation of the NB and Poisson bi-clustering model is available in `https://github.com/weigcdsb/bi_clustering` and supplementary material. The additional details for MCMC sampling can be found in appendix Section A.1.

3 SIMULATION

To validate and illustrate the proposed bi-clustering method, we simulate neural spikes from the NB bi-clustering generative model defined in observation equation 1 and system equation 2. In this simulation, we generate 3 clusters with 10 neurons in each cluster ($N = 30$ in total). The recording length is $T = 500$ and the dimension for $\boldsymbol{x}_{1:T}^{(j)}$ are all $p = 2$. For each neuron, the individual baseline is generated by $d_i \sim N(0, 0.5^2)$, the factor loading is generated by $\boldsymbol{c}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_2)$ and dispersion are all $r_i = 10$. For latent factors of these three clusters $\{\tilde{\boldsymbol{x}}_{1:T}^{(j)}\}_{j=1}^3$, they are generated from two discrete states, and the state indicator $\xi_t = 1, 2$ is generated from a semi-Markov chain (Sansom & Thomson, 2001; Yu, 2010), to encourage longer state duration. These states correspond two sets of linear dynamics: 1) independent state, where $\boldsymbol{A} \in \mathbb{R}^9$ is diagonal and 2) interactive state, where $\boldsymbol{A}$ is a random rotation of an orthogonal matrix, and hence there are interactions between clusters. The bias term is $\boldsymbol{b} = \boldsymbol{0}$ and noise covariance is $\boldsymbol{Q} = \boldsymbol{I}_9 \cdot 10^{-2}$ for both states.

We then apply the proposed bi-clustering methods to the simulated data, by setting the maximum number of states be 10 for the weak-limit sampler. Here, we run a MCMC chain for 10,000 iterations, starting from 1 subject cluster and 10 uniformly distributed temporal states. The results shown here summarize the posterior samples from iteration 2500 to 10,000. The trace plots for several parameters are shown in appendix (Figure 4), and they don't suggest significant convergence issues. First, to evaluate the inferred clustering structure, we check the similarity matrices for both state (Figure 2A) and subject cluster (Figure 2B). The entry (i, j) for a similarity matrix is the posterior probability that data points i and j belong to the same cluster. Both subject and state similarity matrices are sorted according to true clustering assignments, and hence if the algorithm can recover the simulated cluster structures, the diagonal blocks will have high posterior probability, which is the pattern shown in Figure 2A and 2B. The histograms of posterior samples (Figure 2C)show that our method can successfully recover the number of subject and state cluster. To represent the clustering structure in temporal state more intuitively, we also provide the single point estimates of ξ_t (Figure 2D) by maximizing the posterior expected adjusted Rand index (maxEPAR, Fritsch & Ickstadt (2009)), which performs better than other points estimates such as the MAP estimates. All these results show that we can successfully recover the clustering structures on spatial and temporal dimension, including the number of clusters for each.

One the other hand, the advantage for proposed bi-clustering method is that it can simultaneously provide unbiased estimates of the latent trajectories for each cluster, which can be very helpful for scientific interpretation and insights. Here, we show the latent trajectories for the cluster that most neuron 1-10 belong to in Figure 2E (subject 1-10 also forms the a maxPEAR subject cluster). The samples for $\boldsymbol{x}_t^j$ (excluding μ_t^j) in Figure 2E are rotated by a single matrix, to match the posterior mean to ground truth (the rotation matrix is found least square between posterior mean and ground truth). The trace plots for L2/ Frobenius norms of latent trajectories (show both raw and rotated traces for $\boldsymbol{x}_t^j$) for each detected cluster is shown in Figure 4C-D. The fitting results show that simultaneously considering subject and state clustering structure is necessary for estimation of latent structures. We also show the performance of Poisson bi-clustering model (i.e. replace the NB distribution in observation equation 1 by Poisson distribution). The clustering results are summarized by similarity matrices (Figure 2F) and maxPEAR estimate of state (the third bar in Figure 2D). Since in this simulation example, the over-dispersion is not severe ($r_i = 10$), the Poisson version can also recover the true clustering structure, but with some bias in the estimation of latent trajectories. However, for data with large over-dispersion (which is common for real data), the wrong assumption on equi-dispersion will hugely influence the clustering structures, and it would be necessary to use the more flexible NB bi-clustering model.

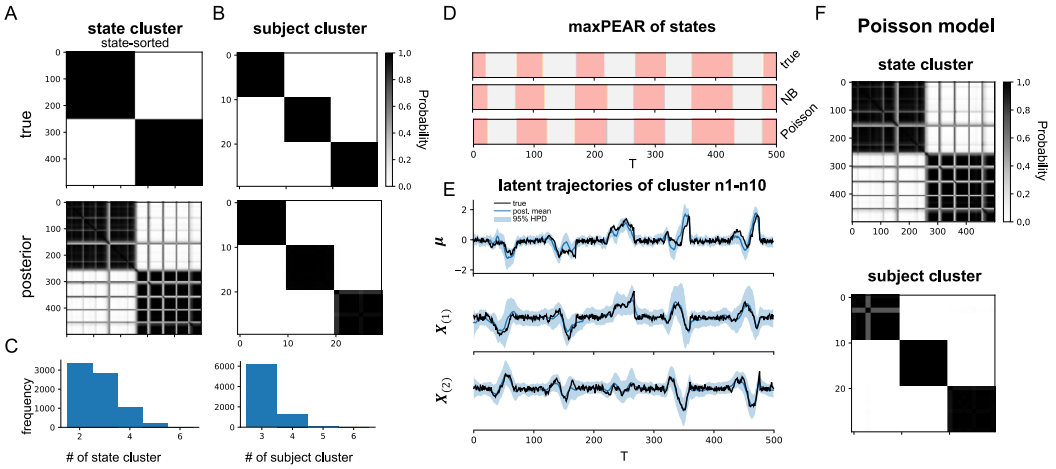

Figure 2: **Simulations.** Here, we show the results for posterior samples from iteration 2500 to 10,000 for each MCMC chain on the simulated dataset. These results are from NB bi-clustering model if not specified. **A**. The posterior similarity matrix for temporal states are ordered according to ground true states, representing the inferred clustering structures relative to ground truth. **B**. Spatially, the similarity matrix for subject are ordered according true subject clusters. **C**. The histograms of posterior samples on number of state cluster (true = 2) and subject cluster (true = 3). **D**. The maxPEAR estimates (point estimates) of the discrete states for NB and Poisson bi-clustering model, comparing to the true temporal states. **E**. The inferred latent trajectories for the detected cluster that most subject 1-10 belong to. Here $\boldsymbol{\mu}$ denotes $\mu_{1:T}$, $\boldsymbol{X}_{(i)}$ denotes i-th row of $\boldsymbol{x}_{1:T}$, and samples of $\boldsymbol{X}_{(i)}$ are transformed by a single matrix, to match posterior mean and ground truth (by least square). The black lines are truths, blue lines are posterior means and shaded light blue regions are 95% highest posterior density (HPD) regions. **F**. The similarity matrices of state and subject cluster for Poisson bi-clustering model, sorted using the same order as in panel **A** and **B** respectively.

4 APPLICATION

We then apply our bi-clustering method to Allen Institute Visual Coding Neuropixels dataset. The dataset contains neural spiking activity from multiple brain regions of an awake mouse, under different visual stimuli. See Siegle et al. (2021) for more detailed data description. Here, we use the electrophysiology session 719161530 to investigate the bi-clustering structures of neurons from three anatomical sites, under three consecutive experimental epochs. After excluding neurons having less than 1Hz response rate, 78 neurons are contained in the following analysis. Among these neurons, 37 originate from the hippocampal CA1, 20 from the lateral posterior nucleus of the thalamus (LP), and 21 from the primary visual cortex (VISp). The neural spikes are recorded when the mouse is exposed to three consecutive visual stimuli : spontaneous (S, durates 30.025s), natural movie (N, durates 300.251s) and again spontaneous (S, durates 30.025s). These three epochs are chosen, according to the hypothesis that spontaneous connectivity patterns in visual cortex are shaped by habitual coactivations produced by natural visual stimulation (Fiser et al., 2010; Harmelech & Malach, 2013; Sadaghiani & Kleinschmidt, 2013), and the similarity in connectivity patterns between spontaneous and natural movie are found in human visual cortex (Wilf et al., 2017). Here, we rebin the data with 500ms, and hence $T = 720$. For formal application, we may need a smaller bin size for the higher resolution. The binned spiking counts for these 78 neurons are shown in Figure 3A.

Then, we fit the data with both NB bi-clustering and Poisson bi-clustering model, and run two independent chains for each. The results from all four chains can be found in the Section A.3, Figure 6. Although the formal analysis requires us to tune some parameters such as latent dimension p and the sticky parameter m in equation 4, we here run 10,000 iterations using $p = 2$ and $m = 10$, simply to illustrate the usage of proposed method on real data. Since these neurons come from three brain regions, we set the prior for number of subject cluster as $k \sim \text{Geometric}(0.415)$, such that

$P(k \leq 3) = 0.8$. The trace plots for several parameters are provided in appendix (Figure 5B-C), which don't suggest significant convergence issues.

For subject clustering structure, the NB bi-clustering model detects around 10 clusters (histogram and trace in Figure 5A-B), and the posterior similarity matrix sorted by maxPEAR estimate is shown in Figure 3C-i. Generalluy, the method detects a large neural population with a high "confidence", with several weak clusters. We further sort the similarity matrix according to the anatomical labels, to examine the relationship between subject clustering results and anatomy (Figure 3-ii). The re-sorted result show that most neurons of the detected largest cluster come from CA1, while some neurons in LP and VISp are also included. Moreover, although most identified subject clusters are neurons from the same anatomical area, there are some mismatches between these two criteria. Especially, some neurons in CA1 are grouped into the same cluster with neurons in VISp, and also between LP and VISp. This may imply that there are some "functional interactions" between CA1 and VISp, and LP and VISp. We also compare the subject clustering results from Poisson bi-clustering model and mixDPFA, and sort the similarity matrices using the same order as in Figure 3C-ii. When assuming the equi-dispersion and fit the Poisson bi-clustering model, there are more mismatches between the detected clusters and anatomy (Figure 3C-iii), which may suggest some spurious interactions are detected when ignoring the over-dispersion. The mixDPFA assumes Poisson distributed spikes and further ignores the potential state changes along the time. The detected clustering structures are more noisy, and especially can hardly distinguish between VISp and LP (Figure 3C-iv). The results are consistent with previous finding of the mixDPFA, which shows that the neural population may change under different experimental settings, if the static dynamics is assumed (Wei et al., 2023). Overall, these results suggest that it is necessary to consider the over-dispersion and time-varying nonlinear dynamics, to obtain unbiased estimate of clustering structures.

For the state clustering structures (histogram and trace in Figure 5A-B), the algorithm detects around 13 clusters, and we show the similarity matrix (Figure 3B) and the maxPEAR estimate (Figure 3D). The stage changes faster for spontaneous (S) than natural movie (N), but the temporal states don't show a clear pattern as in subject cluster in NB-biclustering model. This may suggest similarity between spontaneous and natural visual stimulation for temporal response, in addition to spatial connectivity pattern (Wilf et al., 2017). However, the Poisson bi-clustering model suggests the earlier and later neurons have different stages, which may be spurious by ignoring the over-dispersion. Further application to data with "ground-truth" states, such as sleep/ wake, may help us study effects of over-dispersion in different models.

Finally, we also show the details of the largest maxPEAR subject cluster (traces of L2 norm for latent trajectoreis in Figure 5C). The largest maxPEAR cluster has 21 neurons, which contains 9 neurons from CA1, 7 neurons from LP and 5 neurons from VISp. The spiking counts of these neurons (Figure 3E) may suggest periodic pattern, i.e., alternating strong and weak response, in the middle portion of the natural movie epoch. The observed pattern is captured by the latent trajectories that most of these neurons belong to (Figure 3E).

5 Discussion

In this paper, we introduce a Bayesian nonparametric method to cluster the neural spiking activity spatially and temporally, according to the latent structures (trajectories) for each neural population. Compared to other clustering method for time series (e.g. distance-based methods), the clustering structures defined by latent trajectories can provide scientific insights for the large-scale complicated time series data. Moreover, simultaneously consider the subject and state clustering structures can provide us unbiased and consistent estimates of latent structures (e.g. trajectories) and vice versa.

Although the proposed method can successfully bi-cluster the neural spikes, there are some potential improvements. First, the subject clustering structures are modeled by MFMM, which consider the nature for number of neural populations. However, the uncertainty of clustering results can be large in some cases, and hence it may be better to consider the generalized MFMM (gMFMM), which can provide greater efficiency in the cluster estimation (Frühwirth-Schnatter et al., 2021). Moreover, the common atom specification of gMFMM (Denti et al., 2023; D'Angelo et al., 2023) can provide flexiblity in partitions estimations, resolve the degeneracy isuues and more importantly can allow us borrow information from different neural populations. Second, we currently pre-specify and assume all clusters share the same dimension of latent factors p for convenience. However, this

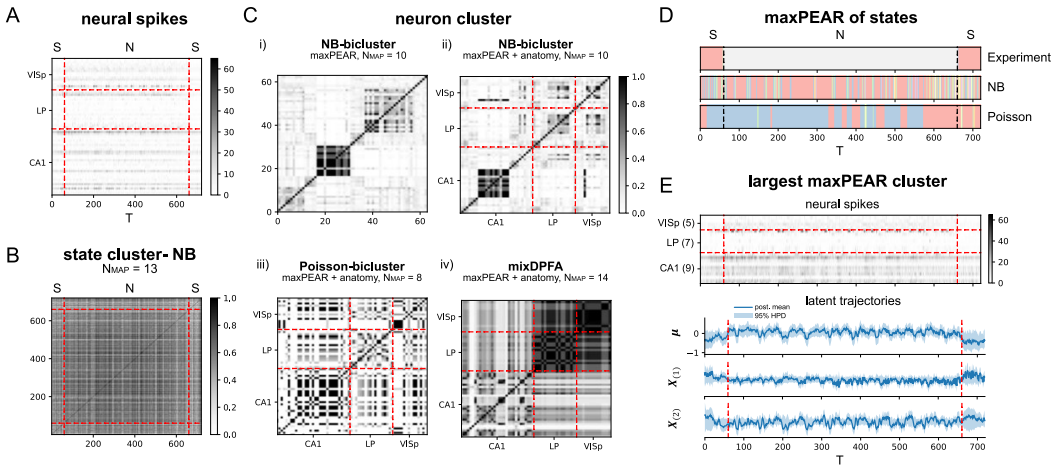

Figure 3: **Application in multi-regional neural data under different experiment epochs. A**. Here, we apply our method to multi-regional Neuropixels data, which contains neural spikes from 3 regions (CA1, LP and VISp) across 3 periods with different visual stimuli: spontaneous (S), natural movie (N) and spontaneous (S). The results from iteration 2500 to 10,000 for each chain are shown here. **B**. The similarity matrix of state cluster for NB bi-clustering model. **C**. The similarity matrices of neuron cluster sorted by maxPEAR estimate for NB bi-clustering model (NB-maxPEAR, upper-left). The clustering results sorted by both NB-maxPEAR and anatomical sites for three different clustering models (NB bi-clustering, Poisson bi-clustering and mixDPFA) are shown here for comparison. **D**. The maxPEAR estimates of the discrete states for NB and Poisson bi-clustering model. **E**. The largest maxPEAR cluster contains 9 neurons from CA1, 7 neurons from LP and 5 neurons from VISp. The upper panel shows the observed neural spikes. The lower panel shows the latent trajectories that most these neurons belong to, where the blue lines are posterior means and shaded light blue regions are 95% HPD regions.

assumption may be inappropriate for real data application, and the method can be more flexible to infer p at the same time. Previously, Wei et al. (2023) sample the latent dimension by a birth-and-death MCMC (BDMCMC) (Fokoué & Titterington, 2003; Stephens, 2000) with the marginalized likelihood, which requires very little mathmatical sophistication and is easy for interpretation. Some other methods, such as putting multiplicative Gamma process prior (Bhattacharya & Dunson, 2011), multiplicative exponential process prior (Wang et al., 2016) and Beta process priopr (Paisley & Carin, 2009; Chen et al., 2010) and Indian Buffet process prior (Knowles & Ghahramani, 2007; 2011; Ročková & George, 2016) may also be useful. Third, when clustering the temporal state, we tried to use the sticky-HDP-HMM to handle the rapid-switching issue. However, the method restrict to geometric state duration and doesn't allow for learning state-specific duration information. When applying to the Neuropixels data, the state looks change fast. This may suggest that we need to model the state duration more carefully, for example, by HDP-HSMM (Johnson & Willsky, 2013). Moreover, neither sticky-HDP-HMM nor HDP-HSMM allow the transition of discrete latent state ξ_t to depend on latent trajectories $\tilde{\boldsymbol{x}}_t$. Therefore, it may be possible to combine idea of recurrent HMM (Linderman et al., 2017) with HDP-HSMM, which may lead to some method like HDP-recurrent-HSMM, for instance. Finally, although the MCMC algorithm developed here is quite efficient, a deterministic approximation of MCMC, such as variational inference may be more computationally efficient and can be more attractive for scientific application.

To sum up, as the scale of neural spiking data becoming large both spatially and temporally, understanding the latent structures of multiple populations under different conditions can be a major statistical challenge. Here, we provide a way to extract spatio-temporal clustering structure, according to their low-dimensional latent trajectories. Although the proposed bi-clustering method is to resolve problems in neuroscience, this method can be potentially useful to extract insightful latent structures (bi-clustering and trajectories) from general large-scale (counting) time series.

REFERENCES

Mauricio A. Álvarez and Neil D. Lawrence. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12(41):1459–1500, 2011. URL http://jmlr.org/papers/v12/alvarez11a.html.

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291, 2011. ISSN 00063444. doi: 10.1093/BIOMET/ASR013.

Mihail Bota and Larry W Swanson. The neuron classification problem. *Brain Research Reviews*, 56(1):79–88, 2007. ISSN 0165-0173. doi: https://doi.org/10.1016/j.brainresrev.2007.05.005. URL https://www.sciencedirect.com/science/article/pii/S0165017307000768.

Phillip Boyle and Marcus Frean. Dependent Gaussian Processes. In L Saul, Y Weiss, and L Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/59eb5dd36914c29b299c84b7ddaf08ec-Paper.pdf.

Lars Buesing, Timothy A Machado, John P Cunningham, and Liam Paninski. Clustered factor analysis of multineuronal spike data. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Jiachen Cai, Robert J. B. Goudie, Colin Starr, and Brian D. M. Tom. Dynamic factor analysis with dependent gaussian processes for high-dimensional gene expression trajectories, 2023.

C. K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.3.541. URL https://doi.org/10.1093/biomet/81.3.541.

Minhua Chen, Jorge Silva, John Paisley, Chunping Wang, David Dunson, and Lawrence Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155, 2010. doi: 10.1109/TSP.2010.2070796.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012. ISSN 1476-4687. doi: 10.1038/nature11129. URL https://doi.org/10.1038/nature11129.

Benjamin R Cowley, Adam C Snyder, Katerina Acar, Ryan C Williamson, Byron M Yu, and Matthew A Smith. Slow Drift of Neural Activity as a Signature of Impulsivity in Macaque Visual and Prefrontal Cortex. *Neuron*, 108(3):551–567.e8, 2020. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2020.07.021. URL https://www.sciencedirect.com/science/article/pii/S0896627320305651.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. ISSN 1546-1726. doi: 10.1038/nn.3776. URL https://doi.org/10.1038/nn.3776.

Laura D'Angelo, Antonio Canale, Zhaoxia Yu, and Michele Guindani. Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. *Biometrics*, 79(2):1370–1382, jun 2023. ISSN 0006-341X. doi: https://doi.org/10.1111/biom.13626. URL https://doi.org/10.1111/biom.13626.

Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira. A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, 118(541):405–416, jan 2023. ISSN 0162-1459. doi: 10.1080/01621459.2021.1933499. URL https://doi.org/10.1080/01621459.2021.1933499.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130, 2010. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2010.01.003. URL https://www.sciencedirect.com/science/article/pii/S1364661310000045.

Ernest Fokoué and D. M. Titterington. Mixtures of factor analysers. Bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50(1-2):73–94, 2003. ISSN 08856125. doi: 10.1023/A:1020297828025.

Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. In D Koller, D Schuurmans, Y Bengio, and L Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008a. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/950a4152c2b4aa3ad78bdd6b366cc179-Paper.pdf.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for Systems with State Persistence. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 312–319, New York, NY, USA, 2008b. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390196. URL https://doi.org/10.1145/1390156.1390196.

Emily Beth Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

Arno Fritsch and Katja Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, 2009. doi: 10.1214/09-BA414. URL https://doi.org/10.1214/09-BA414.

S Frühwirth-Schnatter, G ; Malsiner-Walli, and B ; Grün. Generalized Mixtures of Finite Mixtures and Telescoping Sampling. *Bayesian Analysis*, 16(4):1279–1307, 2021. doi: 10.1214/21-ba1294. URL https://doi.org/10.1214/21-ba1294.

Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202, 1994.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, pp. 214262, jan 2017. doi: 10.1101/214262. URL http://biorxiv.org/content/early/2017/11/12/214262.abstract.

Z Ghahramani and G E Hinton. Variational Learning for Switching State-Space Models. *Neural Computation*, 12(4):831–864, 2000. ISSN 0899-7667 VO - 12. doi: 10.1162/089976600300015619.

Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14867–14878, 2020a. URL https://proceedings.neurips.cc/paper/2020/file/aa1f5f73327ba40d47ebce155e785aaf-Paper.pdf.

Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14867–14878, 2020b. URL https://proceedings.neurips.cc/paper/2020/file/aa1f5f73327ba40d47ebce155e785aaf-Paper.pdf.

Haochen Xu Xuening Zhu Guanhua Fang, Ganggang Xu and Yongtao Guan. Group network hawkes process. *Journal of the American Statistical Association*, 0(ja):1–78, 2023. doi: 10.1080/01621459.2023.2257889. URL https://doi.org/10.1080/01621459.2023.2257889.

Tal Harmelech and Rafael Malach. Neurocognitive biases and the patterns of spontaneous correlations in the human cortex. *Trends in Cognitive Sciences*, 17(12):606–615, dec 2013. ISSN 1364-6613. doi: 10.1016/j.tics.2013.09.014. URL https://doi.org/10.1016/j.tics.2013.09.014.

Matthew J Johnson and Alan S Willsky. Bayesian Nonparametric Hidden Semi-Markov Models. *J. Mach. Learn. Res.*, 14(1):673–701, feb 2013. ISSN 1532-4435.

Eric Jonas and Konrad Kording. Automatic discovery of cell types and microcircuitry from neural connectomics. *eLife*, 4:e04250, 2015. ISSN 2050-084X. doi: 10.7554/eLife.04250. URL https://doi.org/10.7554/eLife.04250.

James J. Jun, Nicholas A. Steinmetz, Joshua H. Siegle, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature 2017 551:7679*, 551(7679):232–236, 2017. ISSN 1476-4687. doi: 10.1038/nature24636. URL https://www.nature.com/articles/nature24636.

Matthew T Kaufman, Mark M Churchland, Stephen I Ryu, and Krishna V Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3):440–448, 2014. ISSN 1546-1726. doi: 10.1038/nn.3643. URL https://doi.org/10.1038/nn.3643.

David Knowles and Zoubin Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation*, pp. 381–388, Berlin, Heidelberg, 2007. ISBN 978-3-540-74494-8.

David Knowles and Zoubin Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534 – 1552, 2011. doi: 10.1214/10-AOAS435. URL https://doi.org/10.1214/10-AOAS435.

Scott Linderman, Ryan P Adams, and Jonathan W Pillow. Bayesian latent structure discovery from multi-neuron recordings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 914–922, 2017. URL https://proceedings.mlr.press/v54/linderman17a.html.

Scott Linderman, Annika Nichols, David Blei, Manuel Zimmer, and Liam Paninski. Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in ¡em¿C. elegans¡/em¿. *bioRxiv*, pp. 621540, jan 2019. doi: 10.1101/621540. URL http://biorxiv.org/content/early/2019/04/29/621540.abstract.

Jakob H. Macke, Lars Buesing, John P. Cunningham, Byron M. Yu, Krishna V. Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in Neural Information Processing Systems*, 24, 2011.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013. ISSN 1476-4687. doi: 10.1038/nature12742. URL https://doi.org/10.1038/nature12742.

Najja J Marshall, Joshua I Glaser, Eric M Trautmann, Elom A Amematsro, Sean M Perkins, Michael N Shadlen, L F Abbott, John P Cunningham, and Mark M Churchland. Flexible neural control of motor units. *Nature Neuroscience*, 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01165-8. URL https://doi.org/10.1038/s41593-022-01165-8.

Jeffrey W. Miller and Scott L. Carter. Inference in generalized bilinear models, 2020. URL https://arxiv.org/abs/2010.04896.

Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340, 2018. doi: 10.1080/01621459.2016.1255636.

Kevin Patrick Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22(10):1677–1686, 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0502-4. URL https://doi.org/10.1038/s41593-019-0502-4.

Radford M Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Sacha B Nelson, Ken Sugino, and Chris M Hempel. The problem of neuronal cell types: a physiological genomics approach. *Trends in Neurosciences*, 29(6):339–345, 2006. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2006.05.004. URL https://www.sciencedirect.com/science/article/pii/S0166223606000920.

Mariann Oemisch, Stephanie Westendorff, Stefan Everling, and Thilo Womelsdorf. Interareal Spike-Train Correlations of Anterior Cingulate and Dorsal Prefrontal Cortex during Attention Shifts. *The Journal of Neuroscience*, 35(38):13076 LP – 13089, sep 2015. doi: 10.1523/JNEUROSCI.1262-15.2015. URL http://www.jneurosci.org/content/35/38/13076.abstract.

John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 777–784, New York, NY, USA, 2009. ISBN 9781605585161. doi: 10.1145/1553374.1553474. URL https://doi.org/10.1145/1553374.1553474.

Panagiotis Papastamoulis and George Iliopoulos. An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, jan 2010. ISSN 1061-8600. doi: 10.1198/jcgs.2010.09008. URL https://doi.org/10.1198/jcgs.2010.09008.

Panagiotis Papastamoulis and Ioannis Ntzoufras. On the identifiability of Bayesian factor analytic models. *Statistics and Computing*, 32(2):23, 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10084-4. URL https://doi.org/10.1007/s11222-022-10084-4.

Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013. doi: 10.1080/01621459.2013.829001. URL https://doi.org/10.1080/01621459.2013.829001.

Arezoo Pooresmaeili, Jasper Poort, and Pieter R Roelfsema. Simultaneous selection by object-based attention in visual and frontal cortex. *Proceedings of the National Academy of Sciences*, 111(17):6467–6472, apr 2014. doi: 10.1073/pnas.1316181111. URL https://doi.org/10.1073/pnas.1316181111.

Raquel Prado and Mike West. *Time Series: Modeling, Computation, and Inference*. Chapman & Hall/CRC, 1st edition, 2010. ISBN 1420093363.

Carlos E Rodríguez and Stephen G Walker. Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, jan 2014. ISSN 1061-8600. doi: 10.1080/10618600.2012.735624. URL https://doi.org/10.1080/10618600.2012.735624.

Veronika Ročková and Edward I. George. Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622, 2016. doi: 10.1080/01621459.2015.1100620. URL https://doi.org/10.1080/01621459.2015.1100620.

Douglas A Ruff and Marlene R Cohen. Attention Increases Spike Count Correlations between Visual Cortical Areas. *The Journal of Neuroscience*, 36(28):7523 LP – 7534, jul 2016. doi: 10.1523/JNEUROSCI.0610-16.2016. URL http://www.jneurosci.org/content/36/28/7523.abstract.

Sepideh Sadaghiani and Andreas Kleinschmidt. Functional interactions between intrinsic brain activity and behavior. *NeuroImage*, 80:379–386, 2013. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.04.100. URL https://www.sciencedirect.com/science/article/pii/S105381191300459X.

John Sansom and Peter Thomson. Fitting Hidden Semi-Markov Models to Breakpoint Rainfall Data. *Journal of Applied Probability*, 38:142–157, sep 2001. ISSN 00219002. URL http://www.jstor.org/stable/3215875.

Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55:103–111, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.02.002. URL https://www.sciencedirect.com/science/article/pii/S0959438818300990.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, 2019. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2019.01.026. URL https://www.sciencedirect.com/science/article/pii/S0896627319300534.

Jayaram Sethuraman. A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS. *Statistica Sinica*, 4(2):639–650, nov 1994. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24305538.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature 2021 592:7852*, 592(7852):86–92, 2021. ISSN 1476-4687. doi: 10.1038/s41586-020-03171-x. URL https://www.nature.com/articles/s41586-020-03171-x.

A'yunin Sofro, Jian Qing Shi, and Chunzheng Cao. Regression analysis for multivariate dependent count data using convolved gaussian processes, 2017.

Sergey D Stavisky, Jonathan C Kao, Stephen I Ryu, and Krishna V Shenoy. Motor Cortical Visuomotor Feedback Activity Is Initially Isolated from Downstream Targets in Output-Null Neural State Space Dimensions. *Neuron*, 95(1):195–208.e9, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.05.023. URL https://www.sciencedirect.com/science/article/pii/S0896627317304610.

Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1787-x. URL https://doi.org/10.1038/s41586-019-1787-x.

Nicholas A. Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, Susu Chen, Jennifer Colonell, Richard J. Gardner, Bill Karsh, Fabian Kloosterman, Dimitar Kostadinov, Carolina Mora-Lopez, John O'Callaghan, Junchol Park, Jan Putzeys, Britton Sauerbrei, Rik J. J. van Daal, Abraham Z. Vollan, Shiwei Wang, Marleen Welkenhuysen, Zhiwen Ye, Joshua T. Dudman, Barundeb Dutta, Adam W. Hantman, Kenneth D. Harris, Albert K. Lee, Edvard I. Moser, John O'Keefe, Alfonso Renart, Karel Svoboda, Michael Häusser, Sebastian Haesler, Matteo Carandini, and Timothy D. Harris. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021. doi: 10.1126/science.abf4588. URL https://www.science.org/doi/abs/10.1126/science.abf4588.

Matthew Stephens. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40 – 74, 2000. doi: 10.1214/aos/1016120364. URL https://doi.org/10.1214/aos/1016120364.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, apr 2019. doi: 10.1126/science.aav7893. URL https://doi.org/10.1126/science.aav7893.

Valérie Ventura. Traditional waveform based spike sorting yields biased rate code estimates. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):6921–6926, 2009. ISSN 00278424. doi: 10.1073/PNAS.0901771106/SUPPL-FILE/SI-APPENDIX.PDF. URL www.pnas.org/cgi/content/full/.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, 43(1):249–275, 2020. ISSN 0147-006X. doi: 10.1146/annurev-neuro-092619-094115. URL https://doi.org/10.1146/annurev-neuro-092619-094115.

Ye Wang, Antonio Canale, and David Dunson. Scalable geometric density estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 857–865, Cadiz, Spain, 2016. URL https://proceedings.mlr.press/v51/wang16e.html.

Ganchao Wei, Ian H Stevenson, and Xiaojing Wang. Bayesian clustering of neural spiking activity using a mixture of dynamic poisson factor analyzers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19499–19510. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b39f4512a2e3899edcc59c7501f3cd4-Paper-Conference.pdf.

Ganchao Wei, Ian H. Stevenson, and Xiaojing Wang. A flexible bayesian clustering of dynamic subpopulations in neural spiking activity, 2023.

Meytal Wilf, Francesca Strappini, Tal Golan, Avital Hahamy, Michal Harel, and Rafael Malach. Spontaneously Emerging Patterns in Human Visual Cortex Reflect Responses to Naturalistic Sensory Stimuli. *Cerebral Cortex*, 27(1):750–763, jan 2017. ISSN 1047-3211. doi: 10.1093/cercor/bhv275. URL https://doi.org/10.1093/cercor/bhv275.

Jesse Windle, Carlos M. Carvalho, James G. Scott, and Liang Sun. Efficient Data Augmentation in Dynamic Models for Binary and Count Data. 2013. URL http://arxiv.org/abs/1308.0774.

Hongteng Xu and Hongyuan Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dd8eb9f23fbd362da0e3f4e70b878c16-Paper.pdf.

Lihao Yin, Ganggang Xu, Huiyan Sang, and Yongtao Guan. Row-clustering of a point process-valued matrix. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20028–20039. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a6a38989dc7e433f1f42388e7afca318-Paper.pdf.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. ISSN 15221598. doi: 10.1152/JN.90941.2008/ASSET/IMAGES/LARGE/Z9K007099508A002.JPEG. URL https://journals.physiology.org/doi/abs/10.1152/jn.90941.2008.

Shun-Zheng Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, 2010. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2009.11.011. URL https://www.sciencedirect.com/science/article/pii/S0004370209001416.

Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.06.031. URL https://doi.org/10.1016/j.cell.2022.06.031.

Mingyuan Zhou, Lingbo Li, David Dunson, and Lawrence Carin. Lognormal and Gamma Mixed Negative Binomial Regression. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 859–866, 2012. ISBN 9781450312851.

David M Zoltowski, Jonathan W Pillow, and Scott W Linderman. A General Recurrent State Space Framework for Modeling Neural Dynamics during Decision-Making. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

A APPENDIX

A.1 MCMC UPDATES

The posteriors of the model parameters are sampled by a MCMC algorithm. To illustrate the sampling steps for each iteration, we expand them into six steps and provide the details as follows.

A.1.1 UPDATE LATENT FACTORS

Denote the p.m.f. for NB(r, μ) as

$$f_{NB}(y \mid r, \mu) = \frac{\Gamma(r+y)}{y!\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y$$

, then the full conditional distribution of $\tilde{\boldsymbol{X}}_{1:T}$ is as follows:

$$\begin{aligned} & P(\tilde{\boldsymbol{X}}_{1:T} \mid \{y_{i,1:T}, d_i.\boldsymbol{c}_i\}_{i=1}^N, \{\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l\}, \xi_t) \\ \propto \quad & \left(\prod_{i=1}^N \prod_{t=1}^T f_{NB}(y_{it} \mid d_i, \mu_{it})\right) \exp\left(-\frac{1}{2}\|\tilde{\boldsymbol{X}}_1\|_2^2\right) \prod_{t=2}^T \exp\left(-\frac{1}{2}\boldsymbol{s}_t'|\boldsymbol{Q}_{\xi_t}|^{-1}\boldsymbol{s}_t\right), \end{aligned}$$

, where $\mu_{it} = \exp\left(d_i + \tilde{\boldsymbol{c}}_i'\tilde{\boldsymbol{x}}_t^{(z_i)}\right)$ and $\boldsymbol{s}_t = \tilde{\boldsymbol{X}}_t - \boldsymbol{A}_{\xi_t}\tilde{\boldsymbol{X}}_{t-1} - \boldsymbol{b}_{\xi_t}$. The full conditional distribution has no closed form, and we sample it via PG augmentation technique, i.e., by introducing the auxiliary PG variable, the transformed "effective" observation $\hat{y}_{it}$ performs like Gaussian, and hence the regular sampling algorithm like forward-filtering-backward-sampling (FFBS) can be implemented to sample $\tilde{X}_{1:T}$. Specifically, we can sample $\tilde{\boldsymbol{X}}_{1:T}$ from full conditionals in two steps:

- Step 1: sample the auxiliary PG parameter ω_{it} and calculate the transformed response $\hat{y}_{it}$. For $i = 1, \ldots, N$ and $t = 1, \ldots T$, sample the ω_{it} from a Pólya-Gamma distribution as $\omega_{it} \sim \text{PG}\left(r_i + y_{it},\, d_i + \tilde{\boldsymbol{c}}_i'\tilde{\boldsymbol{x}}_t^{(z_i)} - \log r_{it}\right)$. Then we can calculate the transformed response as $\hat{y}_{it} = \omega_{it}^{-1}\kappa_{it}$, where $\kappa_{it} = (y_{it} - r_i)/2 + \omega_{it}(\log r_i - d_i)$
- Step 2: sample the $\tilde{\boldsymbol{X}}_{1:T}$ by FFBS. Since the transformed response $\hat{y}_{it} \sim N(\tilde{\boldsymbol{c}}_i'\tilde{\boldsymbol{x}}_t, \omega_{it}^{-1})$, we can use the regular FFBS algorithm for Gaussian state-space model, the detailed algorithm can be found in Chapter 4 in Prado & West (2010).

For data doesn't have severe over-dispersion, it can be useful to assume Poisson distributed observations (i.e. assume $y_{it} \sim \text{Poisson}(\mu_{it})$), for ease of the model inference. The Poisson distribution doesn't fall within the PG augmentation scheme, and we cannot use the method described here. However, motivated by the fact that NB distribution approximates to Poisson distribution as r goes to infinity, we can use the sample from NB model as the proposal, and add one more Metropolis-Hasting step to accept or reject it as in (Wei et al., 2023).

To ensure the model identifiability, we project the posterior samples to the constraint space, such that 1) $\tilde{\boldsymbol{x}}^{(j)}$ is zero-centered, i.e. $\sum_{t=1}^T \tilde{\boldsymbol{x}}_t^{(j)} = \mathbf{0}$ and 2) $\tilde{\boldsymbol{x}}_{1:T}^{(j)}\tilde{\boldsymbol{x}}_{1:T}^{\prime(j)}$ is diagonal, where $\boldsymbol{x}_{1:T}^{(j)} = (\tilde{\boldsymbol{x}}_1^{(j)}, \ldots, \tilde{\boldsymbol{x}}_T^{(j)}) \in \mathbb{R}^{(p+1)\times T}$ (Fokoué & Titterington, 2003). However, the model is still not identifiable in terms of the signed-permutation. For deterministic algorithm, we can put constraints based on singular value decomposition (SVD) as in Miller & Carter (2020), but this is not appropriate for the sampling algorithm. To resolve this, we simply search for the signed-permutation that has the closest Euclidean distance to reference trajectories (in this implementation, we use sample in the 100-th iteration. Before that, align current step to previous). Instead of fixing the reference trajectories manually, we may find the optimized value recursively using the Varimax Rotation-Sign-Permutation (Varimax-RSP) algorithm developed by (Papastamoulis & Ntzoufras, 2022).

A.1.2 UPDATE SUBJECT BASELINE AND FACTOR LOADING

The sampling of subject baseline d_i and factor loading $\boldsymbol{c}_i$ from full conditional distribution is regular NB regression problem for each neuron, which is again updated by PG augmentation technique. The idea is the same as in sampling latent factors, i.e., transform the spikes y_{it} to be Gaussian like

by introducing the augmented parameters $\omega_{it} \sim \text{PG}\left(r_i + y_{it}, \mu_t^{(z_i)} + (1, \boldsymbol{x}_t'^{(z_i)})(d_i, \boldsymbol{c}_i')' - \log r_i\right)$. Therefore, $\hat{y}_{it} = \omega_{it}^{-1}\kappa_{it} \sim \mathcal{N}((1, \boldsymbol{x}_t'^{(z_i)})(d_i, \boldsymbol{c}_i')', \omega_{it}^{-1})$, where $\kappa_{it} = (y_{it} - r_i)/2 + \omega_{it}(\log r_i - \mu_t^{(z_i)})$.

A.1.3 UPDATE DISPERSION

The dispersion for each neuron r_i is updated via a Gibbs sampler, since the gamma distribution is the conjugate prior to it, under the compound Poisson representation. Specifically, let $p_{it} = \mu_{it}/(\mu_{it} + r_i)$, then the conditional posterior of r_i is Gamma $\left(a_0 + \sum_{t=1}^T L_t, 1/(h - \sum_{t=1}^T \log(1 - p_{it}))\right)$, where $L_t \sim \text{Poisson}(-r_i \log(1 - p_{it}))$. Refer to Zhou et al. (2012) for more technical details such as how to sample L_t.

A.1.4 UPDATE DISCRETE STATES

To update the discrete states ξ_t, we use the weak-limit Gibbs sampler for sticky HDP-HMM as in (Fox et al., 2008a), by constructing a finite approximation to the HDP transitions with finite Dirichlet distribution. This is motivated by the fact the infinite limit of hierarchical mixture model converges in distribution to a true HDP as $M \to \infty$, such that

$$\beta \sim \text{Dir}(\eta/M, \ldots, \eta/M),$$
$$\psi_l \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_L).$$

By using this weak limit approximation, we can update the states by an efficient, blocked sampling algorithms. Refer to Fox et al. (2008b) for more details.

A.1.5 UPDATE LINEAR DYNAMICS

Because of the Gaussian assumption in the model, updating the linear dynamics for each state is a Bayesian multivariate linear regression problem. Specifically, let $\tilde{\boldsymbol{X}}_{1:T}^* = (\boldsymbol{1}_T, \tilde{\boldsymbol{X}}_{1:T})$, then for state l assume the conjugate priors for $\{\boldsymbol{b}_l, \boldsymbol{A}_l, \boldsymbol{Q}_l\}$ as

$$\boldsymbol{Q}_l \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \gamma_0),$$
$$\text{vec}((\boldsymbol{b}_l, \boldsymbol{A}_l')) \sim \mathcal{N}(\text{vec}(\boldsymbol{B})_0, \boldsymbol{Q}_l \otimes \Gamma_0^{-1})$$

, where $\mathcal{W}^{-1}$ denotes the inverse-Wishart distribution. The priors are set as $\Psi_0 = 0.01\boldsymbol{I}_{p+1}$, $\gamma_0 = (p+1) + 2$ and $\boldsymbol{B}_0 = (\boldsymbol{0}, \boldsymbol{I})'$. If there are $Q-$chunks of latent factors having $\xi_t = l$, and the denote the time steps for $q-$th chunk as $\tau_q - k_q : \tau_q$.Then, the full conditional distribution are:

$$\boldsymbol{Q}_l \mid \tilde{\boldsymbol{X}}_{1:T} \sim \mathcal{W}^{-1}(\Psi_n, \gamma_n),$$
$$\text{vec}((\boldsymbol{b}_l, \boldsymbol{A}_l')) \mid \tilde{\boldsymbol{X}}_{1:T} \sim \mathcal{N}(\boldsymbol{B}_n, \boldsymbol{Q}_l \otimes \Gamma_n^{-1})$$

, where

$$\Psi_n = \Psi_0 + \sum_{q=1}^Q S_q' S_q + (\boldsymbol{B}_n - \boldsymbol{B}_0)'\Gamma_0(\boldsymbol{B}_n - \boldsymbol{B}_0),$$
$$S_q = \tilde{\boldsymbol{X}}_{\tau_q - k_q + 1:\tau_q} - \tilde{\boldsymbol{X}}_{\tau_q - k_q:\tau_q - 1}^* \boldsymbol{B}_n$$
$$\gamma_n = \gamma_0 + \sum_{q=1}^Q k_q$$
$$\boldsymbol{B}_n = \Gamma_n^{-1}(\sum_{q=1}^Q (\tilde{\boldsymbol{X}}_{\tau_q - k_q:\tau_q - 1}'^* - \tilde{\boldsymbol{X}}_{\tau_q - k_q + 1:\tau_q}) + \Gamma_0 \boldsymbol{B}_0)$$
$$\Gamma_n = \sum_{q=1}^Q (\tilde{\boldsymbol{X}}_{\tau_q - k_q:\tau_q - 1}'^* \tilde{\boldsymbol{X}}_{\tau_q - k_q:\tau_q - 1}^*) + \Gamma_0$$

A.1.6 UPDATE SUBJECT CLUSTER ASSIGNMENTS

To update the subject cluster labels z_is, we use a partition based algorithm, similarly to Miller & Harrison (2018). Let $\mathcal{C}$ denote a partition of neurons, and $\mathcal{C}\backslash i$ denote the partition obtained by removing neuron i from $\mathcal{C}$.

1. Initialize $\mathcal{C}$ and $\tilde{\boldsymbol{X}}_{1:T}^{(c)} : c \in \mathcal{C}\}$ (e.g., one cluster for all neurons in our simulation).
2. In each iteration, for $i = 1, \ldots, N$: remove neuron i from $\mathcal{C}$ and place it:
   (a) in $c \in \mathcal{C}\backslash i$ with probability $\propto (|c| + \gamma)M_c(y_{i,1:T})$, where γ is the hyperparameter of the Dirichlet distribution in equation 3 and $)M_c(y_{i,1:T})$ denotes the marginalized likelihood of neuron i in cluster c, when integrating the loading $\tilde{\boldsymbol{c}}_i$ out.
   (b) in a new cluster c^* with probability $\propto \gamma \frac{V_n(s+1)}{V_n(s)} M_{\boldsymbol{\theta}^{(c^*)}}(\boldsymbol{y}_i)$, where s is the number of partitions by removing the neuron i and $V_n(s) = \sum_{l=1}^{\infty} \frac{l_{(s)}}{(\gamma l)^{(n)}} f_k(l)$, with $x^{(m)} = x(x+1)\cdots(x+m-1)$, $x_{(m)} = x(x-1)\cdots(x-m+1)$, $x^{(0)} = 1$ and $x_{(0)} = 1$.

The update is an adaptation of partition-based algorithm for DPM (Neal, 2000), but with two substitutions: 1) replace $|c_i|$ by $|c_i| + \gamma$ and 2) replace α by $\gamma V_n(t+1)/V_n(t)$. See more details and discussions in (Miller & Harrison, 2018).

Instead of evaluating the full likelihood, we integrate the subject-specific factor loading $\tilde{\boldsymbol{c}}_i$ out to obtain the marginalized likelihood $M_c(y_{i,1:T})$ to achieve better mixing for high dimensional situation. The marginalized likelihood is evaluated by the Laplace approximation as follows:

$$M_c(y_{i,1:T}) \quad \underset{\sim}{\propto} \quad \prod_{t=1}^{T} (f_{NB}(y_{it} \mid r_i, \hat{\mu}_{it}^{(c)}))\pi(\hat{\boldsymbol{c}}_i)|\hat{\Sigma}_{c_i}|^{1/2}$$
$$\hat{\mu}_{it}^{(c)} = d_i + \mu_{it}^{(c)} + \hat{\boldsymbol{c}}_i' \boldsymbol{x}_t^{(c)}$$

, where $\underset{\sim}{\propto}$ means "approximately proportional to", $\hat{\boldsymbol{c}}_i$ and $\hat{\Sigma}_{c_i}$ are MLE estimates and corresponding variance estimates (inverse of the negative Hessian at $\hat{\boldsymbol{c}}_i$) from NB regression on $\boldsymbol{c}_i$.

Since we need to align the latent trajectories to reference value for each cluster, it's necessary to handle the label switching problem for cluster assignment. Here, we implement the Equivalence Casses Representative (ECR) alforithm (Papastamoulis & Iliopoulos, 2010), by setting the cluster assignment at the 100-th iteration (before 100-th step, align current allocation to previous step) as the pivot allocation. To relax the dependency on setting pivot manually, we can further consider iterative versions by Rodríguez & Walker (2014).

A.2 SUPPLEMENTARY RESULTS FOR SIMULATIONS

Here, we show trace plots (Figure 4) of several parameters for NB bi-clustering model in simulation (section 3 and Figure 2). Visual inspection and Geweke diagnostics don't show significant issues with convergence.

A.3 SUPPLEMENTARY RESULTS FOR NEURAL PIXELS

In this section, we first provide histograms and trace plots for some parameters in the MCMC chain for NB bi-clustering model (chain 1 of NB bi-clustering model) in shown in Figure 3. Again, visual inspection and Geweke diagnostics don't show significant issues with convergence.

We then provide results from two independent chains for both NB and Poisson bi-clustering model (i.e., four chains in total). Specifically, we show the similarity matrices of subject clusters sorted by maxPEAR estimates (Figure 6A) and by maxPEAR with anatomical sites (Figure 6B). These orders are obtained from first chain of NB model, which is the same order as used in Figure 3C. We further show the similarity matrices for state cluster for these four chains. Overall, these results show that NB model provides different clustering results compared to the Poisson one.

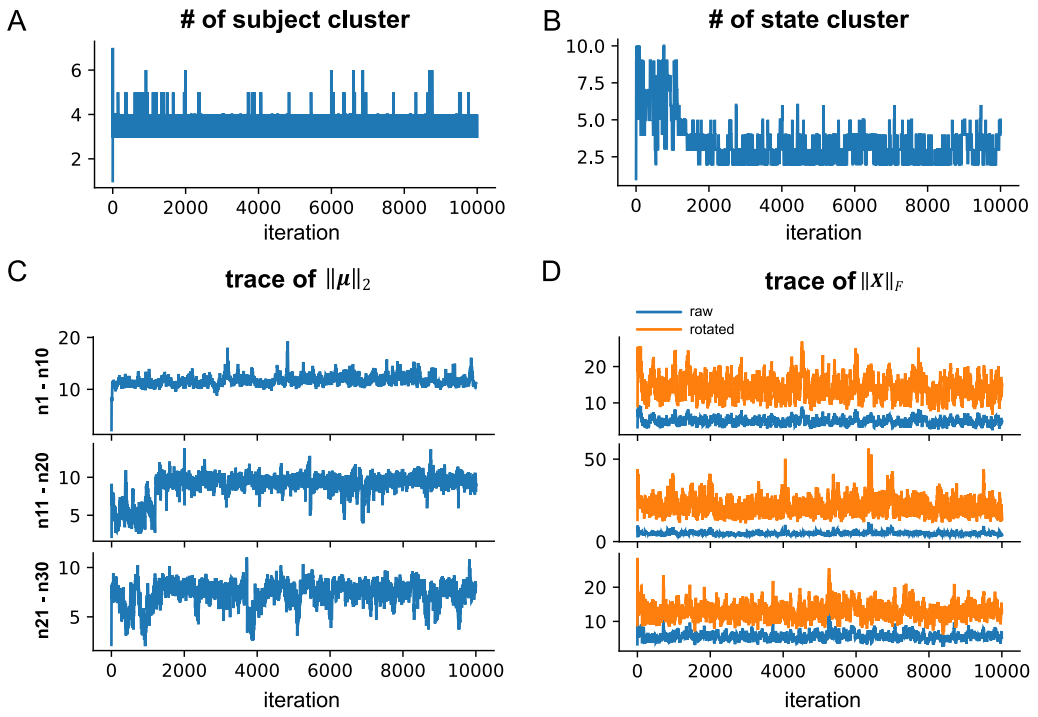

Figure 4: **Trace plots for simulations.** Here we show trace plots of several parameters for NB bi-clustering model in simulation (section 3 and Figure 2), including number of subject cluster (**A**) and number of state cluster (**B**). We further show the traces of $||\mu_{1:T}^{(j)}||_2$ (**C**) and $||\boldsymbol{x}_{1:T}^{(j)}||_F$ (**D**) for the 3 detected clusters (equivalent to true subject cluster). The way we define detected cluster is illustrated in section 3 and Figure 2E. For $||\boldsymbol{x}_{1:T}^{(j)}||_F$, the traces for both the raw (blue) and rotated (orange) samples are shown here

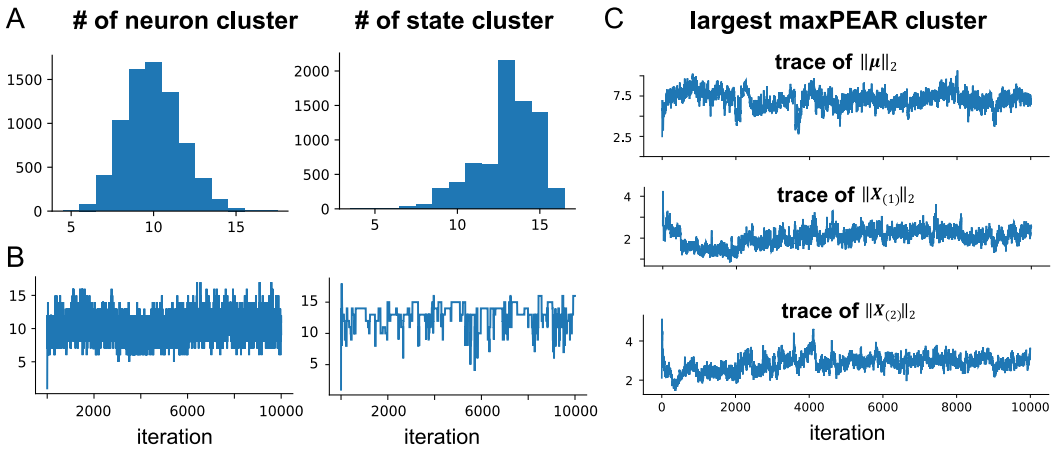

Figure 5: **Histograms and trace plots for chain 1 of NB-biclustering. A**. Histograms for number of neuron/ state cluster, using the samples from iteration 2500 to 10,000. The corresponding trace plots are shown in panel **B**. We then show the L2 norm for each latent trajectory, i.e. $\boldsymbol{\mu}, \boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$, for the largest maxPEAR cluster, corresponding to latent trajectories in Figure 3E. Here $\boldsymbol{\mu}$ denotes $\mu_{1:T}$, $\boldsymbol{X}_{(i)}$ denotes i-th row of $\boldsymbol{x}_{1:T}$.

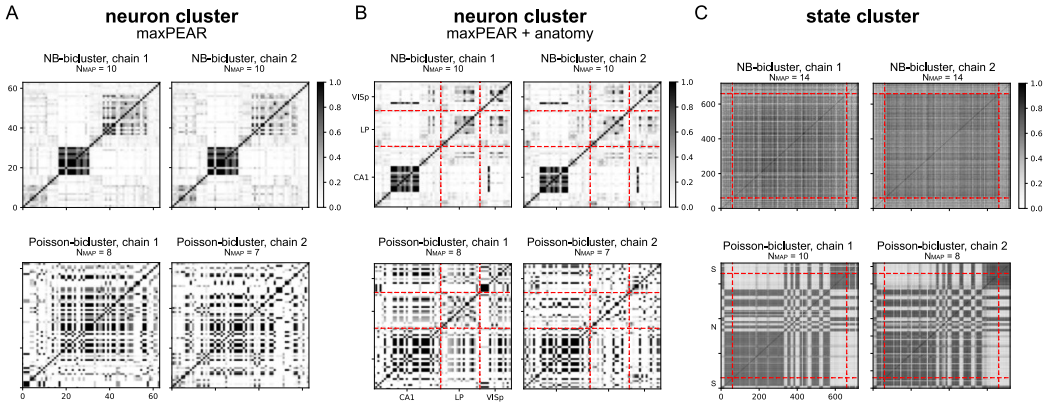

Figure 6: **Supplementary results for the application to Neuropixels data. A**. similarity matrices for neuron cluster, for 2 independent MCMC chains each on both NB and Poisson bi-clustering models. These matrices are sorted by maxPEAR estimates of chain 1 for NB model (also used in main text, Figure 3C i). **B**. We further sorted the matrices by anatomical sites, which is the same order used in the main text, Figure 3C ii-iv. C. The similarity matrices for state cluster, for 2 independent chains on each model.