
Point process models for sequence detection in high-dimensional neural spike trains

Alex H. Williams

Department of Statistics
Stanford University
Stanford, CA 94305
ahwillia@stanford.edu

Anthony Degleris

Department of Electrical Engineering
Stanford University
Stanford, CA 94305
degleris@stanford.edu

Yixin Wang

Department of Statistics
Columbia University
New York NY 10027
yixin.wang@columbia.edu

Scott W. Linderman

Department of Statistics
Stanford University
Stanford, CA 94305
scott.linderman@stanford.edu

Abstract

Sparse sequences of neural spikes are posited to underlie aspects of working memory [1], motor production [2], and learning [3, 4]. Discovering these sequences in an unsupervised manner is a longstanding problem in statistical neuroscience [5–7]. Promising recent work [4, 8] utilized a convolutive nonnegative matrix factorization model [9] to tackle this challenge. However, this model requires spike times to be discretized, utilizes a sub-optimal least-squares criterion, and does not provide uncertainty estimates for model predictions or estimated parameters. We address each of these shortcomings by developing a point process model that characterizes fine-scale sequences at the level of individual spikes and represents sequence occurrences as a small number of marked events in continuous time. This ultra-sparse representation of sequence events opens new possibilities for spike train modeling. For example, we introduce learnable time warping parameters to model sequences of varying duration, which have been experimentally observed in neural circuits [10]. We demonstrate these advantages on experimental recordings from songbird higher vocal center and rodent hippocampus.

1 Introduction

Identifying interpretable patterns in multi-electrode recordings is a longstanding and increasingly pressing challenge in neuroscience. Depending on the brain area and behavioral task, the activity of large neural populations may be low-dimensional as quantified by principal components analysis (PCA) or other latent variable models [11–23]. However, many datasets do not conform to these modeling assumptions and instead exhibit high-dimensional behaviors [24]. *Neural sequences* are an important example of high-dimensional structure: if N neurons fire sequentially with no overlap, the resulting dynamics are N -dimensional and cannot be efficiently summarized by PCA or other linear dimensionality reduction methods [4]. Such sequences underlie current theories of working memory [1, 25], motor production [2], and memory replay [10]. More generally, neural sequences are seen as flexible building blocks for learning and executing complex neural computations [26–28].

Prior Work In practice, neural sequences are usually identified in a supervised manner by correlating neural firing with salient sensory cues or behavioral actions. For example, hippocampal place cells fire sequentially when rodents travel down a narrow hallway, and these sequences can be found by averaging spikes times over multiple traversals of the hallway. After identifying this sequence

on a behavioral timescale lasting seconds, a template matching procedure can be used to show that these sequences reoccur on compressed timescales during wake [29] and sleep [30]. In other cases, sequences can be identified relative to salient features of the local field potential (LFP; [31, 32]).

Developing unsupervised alternatives that directly extract sequences from multineuronal spike trains would broaden the scope of this research and potentially uncover new sequences that are not linked to behaviors or sensations [33]. Several works have shown preliminary progress in this direction Maboudi et al. [34] proposed fitting a hidden Markov model (HMM) and then identifying sequences from the state transition matrix. Grossberger et al. [35] and van der Meij & Voytek [36] apply clustering to features computed over a sliding window. Others use statistics such as time-lagged correlations to detect sequences and other spatiotemporal patterns in a bottom-up fashion [6, 7].

In this paper, we develop a Bayesian point process modeling generalization of *convolutive nonnegative matrix factorization* (convNMF; [9]), which was recently used by Peter et al. [8] and Mackevicius et al. [4] to model neural sequences. Briefly, the convNMF model discretizes each neuron’s spike train into a vector of B time bins, $\mathbf{x}_n \in \mathbb{R}_+^B$ for neuron n , and approximates this collection of vectors as a sum of convolutions, $\mathbf{x}_n \approx \sum_{r=1}^R \mathbf{w}_{n,r} * \mathbf{h}_r$. The model parameters ($\mathbf{w}_{n,r} \in \mathbb{R}_+^L$ and $\mathbf{h}_r \in \mathbb{R}_+^B$) are optimized with respect to a least-squares criterion. Each component of the model, indexed by $r \in \{1, \dots, R\}$, consists of a *neural factor*, $\mathbf{W}_r \in \mathbb{R}_+^{N \times L}$ and a *temporal factor*, \mathbf{h}_r . The neural factor encodes a spatiotemporal pattern of neural activity over L time bins (which is hoped to be sequence), while the temporal factor indicates the times at which this pattern of neural activity occurs. A total of R motifs or sequence types, each corresponding to a different component, are extracted by this model.

There are several compelling features of this approach. Similar to classical nonnegative matrix factorization [37], and in contrast to clustering methods, convNMF captures sequences with overlapping groups of neurons by an intuitive “parts-based” representation. Indeed, convNMF uncovered overlapping sequences in experimental data from songbird higher vocal center (HVC) [4]. Further, if the same sequence is repeated with different peak firing rates, convNMF can capture this by varying the magnitude of the entries in \mathbf{h}_r , unlike many clustering methods. Finally, convNMF efficiently pools statistical power across large populations of neurons to identify sequences even when the correlations between temporally adjacent neurons are noisy—other methods, such as HMMs and bottom-up agglomerative clustering, require reliable pairwise correlations to string together a full sequence.

Our Contributions We propose a point process model for neural sequences (PP-Seq) which extends and generalizes convNMF to continuous time and uses a fully probabilistic Bayesian framework. This enables us to better quantify uncertainty in key parameters—e.g. the overall number of sequences the times at which they occur—and also characterize the data at finer timescales—e.g. whether individual spikes were evoked by a sequence. Most importantly, by achieving an extremely sparse representation of sequence event times, the PP-Seq model enables a variety of model extensions that are not easily incorporated into convNMF or other common methods. We explore one such extension that introduces time warping factors to model sequences of varying duration, as is often observed in neural data [10].

Though we focus on applications in neuroscience, our approach could be adapted to other temporal point processes, which are a natural framework to describe data that are collected at irregular intervals (e.g. social media posts, consumer behaviors, and medical records) [38, 39]. We draw a novel connection between Neyman-Scott processes [40], which encompass PP-Seq and other temporal point process models as special cases, and mixture of finite mixture models [41]. Exploiting this insight, we develop innovative Markov chain Monte Carlo (MCMC) methods for PP-Seq.

2 Model

2.1 Point Process Models and Neyman-Scott Processes

Point processes are probabilistic models that generate discrete sets $\{x_1, x_2, \dots\} \triangleq \{x_s\}_{s=1}^S$ over some continuous space \mathcal{X} . Each member of the set $x_s \in \mathcal{X}$ is called an “event.” A Poisson process [42] is a point process which satisfies three properties. First, the number of events falling within any region $V \subset \mathcal{X}$ is Poisson distributed. Second, there exists a function $\lambda(x) : \mathcal{X} \mapsto \mathbb{R}_+$, called the *intensity function*, for which $\int_V \lambda(x) dx$ equals the expected number of events in V . Third, the number of events in non-overlapping regions of \mathcal{X} are independent. A Poisson process is said to be *homogeneous* when $\lambda(x) = c$ for some constant c . Finally, a *marked* point process extends the

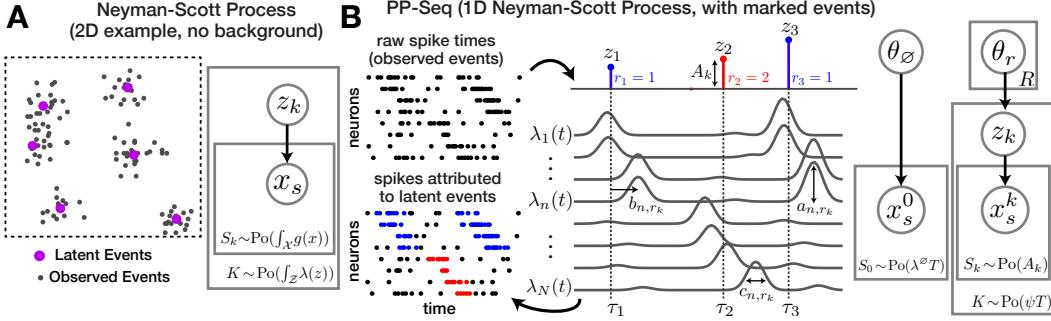


Figure 1: (A) Example of a Neyman-Scott process over a 2D region. Latent events (purple dots) are first sampled from a homogeneous Poisson process. Each latent event then spawns a number of nearby observed events (gray dots), according to an inhomogeneous Poisson process. (B) A spike train can be modeled as a Neyman-Scott process with *marked* events over a 1D interval representing time. Latent events (z_k ; sequences) evoke observed events (x_s ; spikes) ordered in a sequence. An example with $K = 3$ latent events evoking $R = 2$ different sequence types (blue & red) is shown.

usual definition of a point process to incorporate additional information into each event. A marked point process on \mathcal{X} generates random tuples $x_s = (\tilde{x}_s, m_s)$, where $\tilde{x}_s \in \mathcal{X}$ represents the random location and $m_s \in \mathcal{M}$ is the additional “mark” specifying metadata associated with event s . See Supplement A for further background details and references.

Neyman-Scott Processes [40] use Poisson processes as building blocks to model clustered data. To simulate a Neyman-Scott process, first sample a set of *latent events* $\{z_k\}_{k=1}^K \subset \mathcal{Z}$ from an initial Poisson process with intensity function $\lambda(z) : \mathcal{Z} \mapsto \mathbb{R}_+$. Thus, the number of latent events is a Poisson-distributed random variable, $K \sim \text{Poisson}(\int_{\mathcal{Z}} \lambda(z) dz)$. Given the latent events, the *observed events* $\{x_s\}_{s=1}^S$ are drawn from a second Poisson process with conditional intensity function $\lambda(x) = \lambda^\emptyset + \sum_{k=1}^K g(x; z_k)$. The nonnegative functions $g(x; z_k)$ can be thought of as impulse responses—each latent event adds to the rate of observed events, and stochastically generates some number of observed offspring events. Finally, the scalar parameter $\lambda^\emptyset > 0$ specifies a “background” rate; thus, if $K = 0$, the observations follow a homogeneous Poisson process with intensity λ^\emptyset .

A simple example of a Neyman-Scott Process in \mathbb{R}^2 is shown in fig. 1A. Latent events (purple dots) are first drawn from a homogenous Poisson process and specify the cluster centroids. Observed events (gray dots) are then sampled from a second Poisson process, whose intensity function is found by summing all impulse responses.

Importantly, we *do not* observe the number of latent events nor their locations. As described below, we use probabilistic inference to characterize this unobserved structure. A key idea is to attribute each observed event to a latent cause (either one of the latent events or the background process); these attributions are valid due to the additive nature of the latent intensity functions and the *superposition principle* of Poisson processes (see Supplement A). From this perspective, inferring the set of latent events is similar to inferring the number and location of clusters in a nonparametric mixture model.

2.2 Point Process Model of Neural Sequences (PP-Seq)

We model neural sequences as a Neyman-Scott process with *marked* events in a model we call PP-Seq. Consider a dataset with N neurons, emitting a total of S spikes over a time interval $[0, T]$. This can be encoded as a set of S marked spike times—the observed events are tuples $x_s = (t_s, n_s)$ specifying the time, $t_s \in [0, T]$, and neuron, $n_s \in \{1, \dots, N\}$, of each spike. Sequences correspond to latent events, which are also tuples $z_k = (\tau_k, r_k, A_k)$ specifying the time, $\tau_k \in [0, T]$, type, $r_k \in \{1, \dots, R\}$, and amplitude, $A_k > 0$ of the sequence. The hyperparameter R specifies the number of recurring sequence types, which is analogous to the number of components in convNMF.

To draw samples from the PP-Seq model, we first sample sequences (i.e., latent events) from a Poisson process with intensity $\lambda(z) \triangleq \lambda(\tau, r, A) = \psi \pi_r \text{Ga}(A; \alpha, \beta)$. Here, $\psi > 0$ sets the rate at which sequences occur within the data, $\pi \in \Delta_R$ sets the probability of the R sequence types, and α, β parameterize a gamma density which models the sequence amplitude, A . Note that the number

of sequence events is a Poisson-distributed random variable, $K \sim \text{Poisson}(\psi T)$, where the rate parameter ψT is found by integrating $\lambda(z)$ over all sequence types, amplitudes, and times.

Conditioned on the set of K sequence events, the firing rate of neuron n is given by a sum of nonnegative impulse responses:

$$\lambda_n(t) = \lambda_n^\varnothing + \sum_{k=1}^K g_n(t; z_k). \quad (1)$$

We assume these impulse responses vary across neurons and follow a Gaussian form:

$$g_n(t; z_k) = A_k \cdot a_{nr_k} \cdot \mathcal{N}(t | \tau_k + b_{nr_k}, c_{nr_k}), \quad (2)$$

where $\mathcal{N}(t | \mu, \sigma^2)$ denotes a Gaussian density. The parameters $a_r = (a_{1r}, \dots, a_{Nr}) \in \Delta_N$, $b_{nr} \in \mathbb{R}$, and $c_{nr} \in \mathbb{R}_+$ correspond to the weight, latency, and width, respectively, of neurons' firing rates in sequences of type r . Since the firing rate is a sum of non-negative impulse responses, the superposition principle of Poisson processes (see Supplement A) implies that we can view the data as a union of "background" spikes and "induced" spikes from each sequence, justifying the connection to clustering. The expected number of spikes induced by sequence k is:

$$\sum_{n=1}^N \int_0^T g_n(t; z_k) dt \approx \sum_{n=1}^N \int_{-\infty}^{\infty} g_n(t; z_k) dt = A_k, \quad (3)$$

and thus we may view A_k as the amplitude of sequence event k .

Figure 1B schematizes a simple case containing $K = 3$ sequence events and $R = 2$ sequence types. A complete description of the model's generative process is provided in Supplement B, but it can be summarized by the graphical model in fig. 1B, where we have global parameters $\Theta = (\theta_\varnothing, \{\theta_r\}_{r=1}^R)$ with $\theta_r = (a_r, \{b_{nr}\}_{n=1}^N, \{c_{nr}\}_{n=1}^N)$ for each sequence type, and $\theta_\varnothing = \{\lambda_n^\varnothing\}_{n=1}^N$ for the background process. We place weak priors on each parameter: the neural response weights $\{a_{nr}\}_{n=1}^N$ follow a Dirichlet prior for each sequence type, and (b_{nr}, c_{nr}) follows a normal-inverse-gamma prior for every neuron and sequence type. The background rate, λ_n^\varnothing , follows a gamma prior. We set the sequence event rate, ψ , to be a fixed hyperparameter, though this assumption could be relaxed.

Time-warped sequences PP-Seq can be extended to model more diverse sequence patterns by using higher-dimensional marks on the latent sequences. For example, we can model variability in sequence duration by introducing a *time warping factor*, $\omega_k > 0$, to each sequence event and changing eq. (2) to,

$$g_n(t; z_k) = A_k \cdot a_{nr_k} \cdot \mathcal{N}(t | \tau_k + \omega_k b_{nr_k}, \omega_k^2 c_{nr_k}). \quad (4)$$

This has the effect of linearly compressing or stretching each sequence in time (when $\omega_k < 1$ or $\omega_k > 1$, respectively). Such time warping is commonly observed in neural data [43, 44], and indeed, hippocampal sequences unfold \sim 15–20 times faster during replay than during lived experiences [10]. We characterize this model in Supplement E and demonstrate its utility below.

In principle, it is equally possible to incorporate time warping into discrete time convNMF. However, since convNMF involves a dense temporal factor matrix $\mathbf{H} \in \mathbb{R}_+^{R \times B}$, the most straightforward extension would be to introduce a time warping factor for each component $r \in \{1, \dots, R\}$ and each time bin $b \in \{1, \dots, B\}$. This results in $O(RB)$ new trainable parameters, which poses non-trivial challenges in terms of computational efficiency, overfitting, and human interpretability. In contrast, PP-Seq represents sequence events as a set of K latent events in continuous time. This ultra-sparse representation of sequence events (since $K \ll RB$) naturally lends itself to modeling additional sequence features since this introduces only $O(K)$ new parameters.

3 Collapsed Gibbs Sampling for Neyman-Scott Processes

Developing efficient algorithms for parameter inference in Neyman-Scott process models is an area of active research [45–48]. To address this challenge, we developed a collapsed Gibbs sampling routine for Neyman-Scott processes, which encompasses the PP-Seq model as a special case. The method resembles "Algorithm 3" of Neal [49]—a well-known approach for sampling from a Dirichlet process mixture model—and the collapsed Gibbs sampling algorithm for "mixture of finite mixtures" models developed by Miller et al. [41]. The idea is to partition observed spikes into background spikes

and spikes induced by latent sequences, integrating over the sequence times, types, and amplitudes. Starting from an initial partition, the sampler iterates over individual spikes and probabilistically re-assigns them to (a) the background, (b) one of the remaining sequences, or (c) to a new sequence. The number of sequences in the partition, K^* , changes as spikes are removed and re-assigned; thus, the algorithm is able to explore the full trans-dimensional space of partitions.

The re-assignment probabilities are determined by the prior distribution of partitions under the Neyman-Scott process and by the likelihood of the induced spikes assigned to each sequence. We state the conditional probabilities below and provide a full derivation in Supplement D. Let K^* denote the number of sequences in the current partition after spike x_s has been removed from its current assignment. (Note that the number of latent sequences K may exceed K^* if some sequences produce zero spikes.) Likewise, let u_s denote the sequence assignment of the s -th spike, where $u_s = 0$ indicates assignment to the background process and $u_s \in \{1, \dots, K^*\}$ indicates assignment to one of the latent sequence events. Finally, let $X_k = \{x_{s'} : u_{s'} = k, s' \neq s\}$ denote the spikes in the k -th cluster, excluding x_s , and let $S_k = |X_k|$ denote its size. The conditional probability of the partition under the possible assignments of spike x_s are,

$$p(u_s = 0 | x_s, \{X_k\}_{k=1}^{K^*}, \Theta) \propto (1 + \beta) \lambda_{n_s}^\varnothing \quad (5)$$

$$p(u_s = k | x_s, \{X_k\}_{k=1}^{K^*}, \Theta) \propto (\alpha + S_k) \left[\sum_{r_k=1}^R p(r_k | X_k) a_{n_s r_k} p(t_s | X_k, r_k, n_s) \right] \quad (6)$$

$$p(u_s = K^* + 1 | x_s, \{X_k\}_{k=1}^{K^*}, \Theta) \propto \alpha \left(\frac{\beta}{1 + \beta} \right)^\alpha \psi \sum_{r=1}^R \pi_r a_{n_s r} \quad (7)$$

The sampling algorithm iterates over all spikes $x_s \in \{x_1, \dots, x_S\}$ and updates their assignments holding the other spikes' assignments fixed. The probability of assigning spike x_s to an existing cluster marginalizes the time, type, and amplitude of the sequence, resulting in a *collapsed* Gibbs sampler [49, 50]. The exact form of the posterior probability $p(r_k | X_k)$ and the parameters of the posterior predictive $p(t_s | X_k, r_k, n_s)$ in eq. (6) are given in Supplement C.

After attributing each spike to a latent cause, it is straightforward to draw samples over the remaining model parameters—the latent sequences $\{z_k\}_{k=1}^{K^*}$ and global parameters Θ . Given the spikes and assignments $\{x_s, u_s\}_{s=1}^S$, we sample the sequences (i.e. their time, amplitude, types, etc.) from the closed-form conditional $p(z_k | \{x_s : u_s = k\}, \Theta)$. Given the sequences and spikes, we sample the conditional distribution on global parameters $p(\Theta | \{z_k\}_{k=1}^{K^*}, \{x_s, u_s\}_{s=1}^S)$. Under conjugate formulations, these updates are straightforward. With these steps, the Markov chain targets the posterior distribution on model parameters and partitions. Complete derivations are in Supplement D.

Improving MCMC mixing times The intensity of sequence amplitudes A_k is proportional to the gamma density $\text{Ga}(A_k; \alpha, \beta)$, and these hyperparameters affect the mixing time of the Gibbs sampler. Intuitively, if there is little probability of low-amplitude sequences, the sampler is unlikely to create new sequences and is therefore slow to explore different partitions of spikes.¹ If, on the other hand, the variance of $\text{Ga}(\alpha, \beta)$ is large relative to the mean, then the probability of forming new clusters is non-negligible and the sampler tends to mix more effectively. Unfortunately, this latter regime is also probably of lesser scientific interest, since neural sequences are typically large in amplitude—they can involve many thousands of cells, each potentially contributing a small number of spikes [2, 28].

To address this issue, we propose an annealing procedure to initialize the Markov chain. We fix the mean of A_k and adjust α and β to slowly lower variance of amplitude distribution. Initially, the sampler produces many small clusters of spikes, and as we lower the variance of $\text{Ga}(\alpha, \beta)$ to its target value, the Markov chain typically combines these clusters into larger sequences. We further improve performance by interspersing “split-merge” Metropolis-Hastings updates [51, 52] between Gibbs sweeps (see Supplement D.6). Finally, though we have not found it necessary, one could use convNMF to initialize the MCMC algorithm.

Parallel MCMC Resampling the sequence assignments is the primary computational bottleneck for the Gibbs sampler. One pass over the data requires $O(SKR)$ operations, which quickly becomes costly when the operations are serially executed. While this computational cost is manageable for

¹This problem is common to other nonparametric Bayesian mixture models as well [41, e.g.].

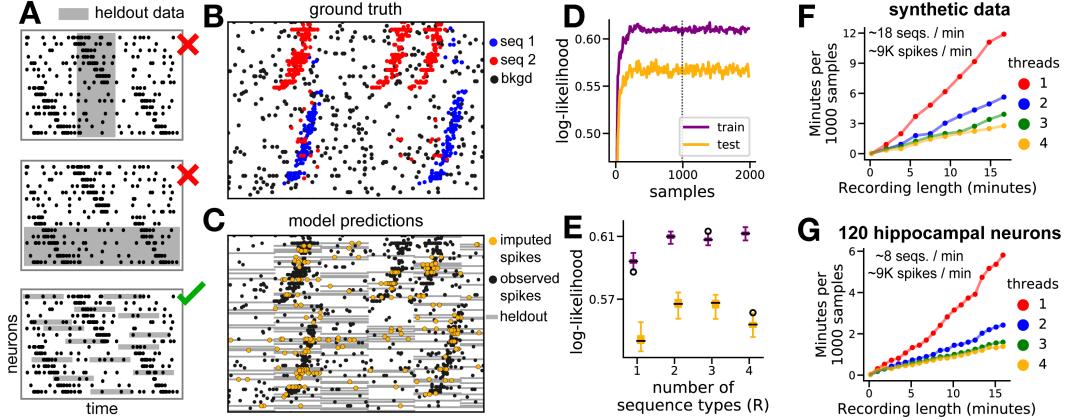


Figure 2: (A) Schematic of train/test partitions. We propose a speckled holdout pattern (bottom). (B) A subset of a synthetic spike train containing two sequences types. (C) Same data, but with grey regions showing the censored test set and yellow dots denoting imputed spikes. (D) Log-likelihood over Gibbs samples; positive values denote excess nats per unit time relative to a homogeneous Poisson process baseline. (E) Box plots showing range of log-likelihoods on the train and test sets for different choices of R ; cross-validation favors $R = 2$, in agreement with the ground truth shown in panel B. (F-G) Performance benefits of parallel MCMC on synthetic and experimental neural data.

many datasets, we can improve performance substantially by parallelizing the computation [53]. Given P processors and a spike train lasting T seconds, we divide the dataset into intervals lasting T/P seconds, and allocate one interval per processor. The current global parameters, Θ , are first broadcast to all processors. In parallel, the processors update the sequence assignments for their assigned spikes, and then send back sufficient statistics describing each sequence. After these sufficient statistics are collected on a single processor, the global parameters are re-sampled and then broadcast back to the processors to initiate another iteration. This algorithm introduces some error since clusters are not shared across processors. In essence, this introduces erroneous edge effects if a sequence of spikes is split across two processors. However, these errors are negligible when the sequence length is much less than T/P , which we expect is the practical regime of interest.

4 Experiments

4.1 Cross-Validation and Demonstration of Computational Efficiency

We evaluate model performance by computing the log-likelihood assigned to held-out data. Partitioning the data into training and testing sets must be done somewhat carefully—we cannot withhold time intervals completely (as in fig. 2A, *top*) or else the model will not accurately predict latent sequences occurring in these intervals; likewise, we cannot withhold individual neurons completely (as in fig. 2A, *middle*) or else the model will not accurately predict the response parameters of those held out cells. Thus, we adopt a “speckled” holdout strategy [54] as diagrammed at the bottom of fig. 2A. We treat held-out spikes as missing data and sample them as part of the MCMC algorithm. (Their conditional distribution is given by the PP-Seq generative model.) This approach involving a speckled holdout pattern and multiple imputation of missing data may be viewed as a continuous time extension of the methods proposed by Mackevicius et al. [27] for convNMF.

Panels B-E in fig. 2 show the results of this cross-validation scheme on a synthetic dataset with $R = 2$ sequence types. The predictions of the model in held-out test regions closely match the ground truth—missing spikes are reliably imputed when they are part of a sequence (fig. 2C). Further, the likelihood of the train and test sets improves over the course of MCMC sampling (fig. 2D), and can be used as a metric for model comparison—in agreement with the ground truth, test performance plateaus for models containing greater than $R = 2$ sequence types (fig. 2E).

Finally, to be of practical utility, the algorithm needs to run in a reasonable amount of time. Figure 2G shows that our Julia [55] implementation can fit a recording of 120 hippocampal neurons with hundreds of thousands of spikes in a matter of minutes, on a 2017 MacBook Pro (3.1 GHz Intel Core i7, 4 cores, 16 GB RAM). Run-time grows linearly with the number of spikes, as expected, but even

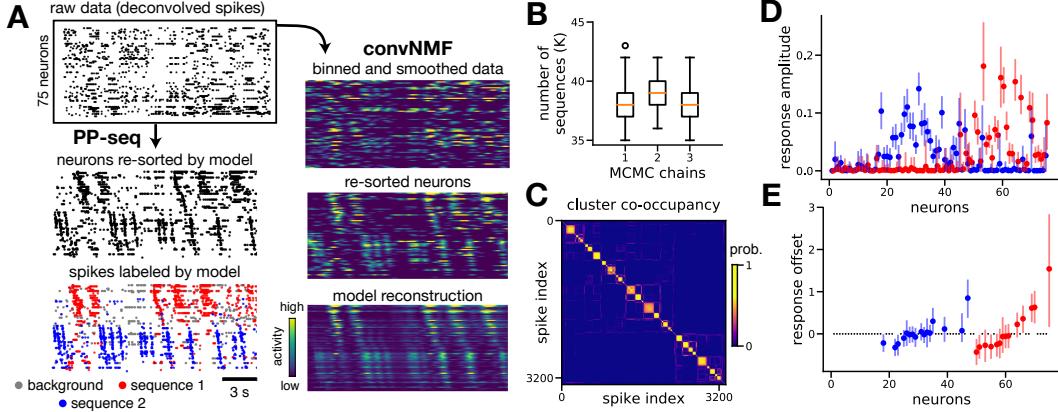


Figure 3: Zebra Finch HVC data. (A) Raw spike train (top) and sequences revealed by PP-Seq (left) and convNMF (right). (B) Box plots summarizing samples from the posterior on number of sequences, K , derived from three independent MCMC chains. (C) Co-occupancy matrix summarizing probabilities of spike pairs belonging to the same sequence. (D) Credible intervals for evoked amplitudes for sequence type 1 (red) and 2 (blue). (E) Credible intervals for response offsets (same order and coloring as D). Estimates are suppressed for small-amplitude responses (gray dots).

with a single thread it only takes six minutes to perform 1000 Gibbs sweeps on a 15-minute recording with $\sim 1.3 \times 10^5$ spikes. With parallel MCMC, this laptop performs the same number of sweeps in under two minutes. Our open-source implementation is available at:

<https://github.com/lindermanlab/PPSeq.jl>.

4.2 Zebra Finch Higher Vocal Center (HVC)

We first applied PP-Seq to a recording of HVC premotor neurons in a zebra finch,² which generate sequences that are time-locked to syllables in the bird’s courtship song. Figure 3A qualitatively compares the performance of convNMF and PP-Seq. The raw data (top panel) shows no visible spike patterns; however, clear sequences are revealed by sorting the neurons lexicographically by preferred sequence type and the temporal offset parameter inferred by PP-Seq. While both models extract similar sequences, PP-Seq provides a finer scale annotation of the final result, providing, for example, attributions at the level of individual spikes to sequences (bottom left of fig. 3A).

Further, PP-Seq can quantify uncertainty in key parameters by considering the full sequence of MCMC samples. Figure 3B summarizes uncertainty in the total number of sequence events, i.e. K , over three independent MCMC chains with different random seeds—all chains converge to similar estimates; the uncertainty is largely due to the rapid sequences (in blue) shown in panel A. Figure 3C displays a symmetric matrix where element (i, j) corresponds to the probability that spike i and spike j are attributed to same sequence. Finally, fig. 3D-E shows the amplitude and offset for each neuron’s sequence-evoked response with 95% posterior credible intervals. These results naturally fall out of the probabilistic construction of the PP-Seq model, but have no obvious analogue in convNMF.

4.3 Time Warping Extension and Robustness to Noise

Songbird HVC is a specialized circuit that generates unusually clean and easy-to-detect sequences. To compare the robustness of PP-Seq and convNMF under more challenging circumstances, we created a simple synthetic dataset with $R = 1$ sequence type and $N = 100$ neurons. We varied four parameters to manipulate the difficulty of sequence extraction: the rate of background spikes, λ_n^\otimes (“additive noise,” fig. 4A), the expected value of sequence amplitudes, A_k (“participation noise”, fig. 4B), the expected variance of the Gaussian impulse responses, c_{nr} (“jitter noise”, fig. 4C), and, finally, the maximal time warping coefficient (see eq. (4); fig. 4D). All simulated datasets involved sequences with low spike counts ($\mathbb{E}[A_k] < 100$ spikes). In this regime, the Poisson likelihood criterion used by PP-Seq is better matched to the statistics of the spike train. Since convNMF optimizes an alternative loss function (squared error instead of Poisson likelihood) we compared the models by their ability

²These data are available at <http://github.com/FeeLab/seqNMF>; originally published in [4].

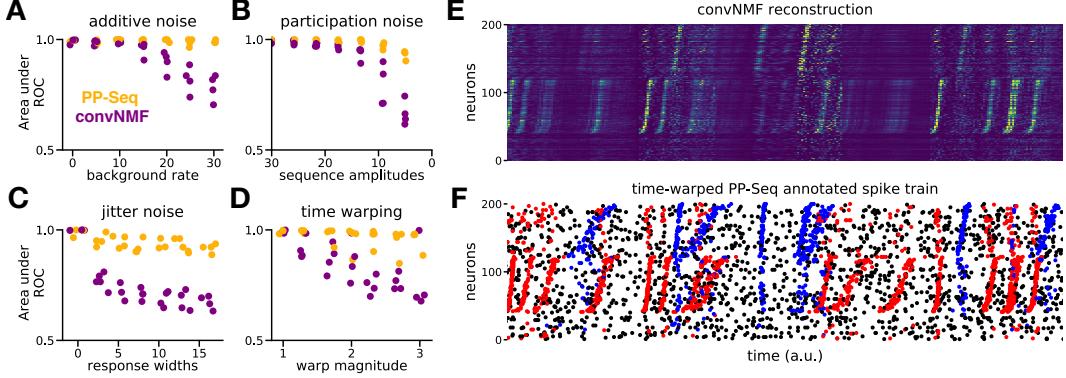


Figure 4: PP-Seq is more robust to various forms of noise than convNMF. (A-D) Comparison of PP-Seq and convNMF to detect sequence times in synthetic data with varying amounts of noise. Panel D shows the performance of the time-warped variant of PP-Seq (see eq. (4)). (E) convNMF reconstruction of a spike train containing sequences with 9-fold time warping. (F) Performance of time-warped PP-Seq (see eq. (4)) on the same data as panel E.

to extract the ground truth sequence event times. Using area under receiver operating characteristic (ROC) curves as a performance metric (see Supplement F.1), we see favorable results for PP-Seq as noise levels are increased.

We demonstrate the abilities of time-warped PP-Seq further in fig. 4E-F. Here, we show a synthetic dataset containing sequences with 9-fold variability in their duration, which is similar to levels observed in some experimental systems [10]. While convNMF fails to reconstruct many of these warped sequences, the PP-Seq model identifies sequences that are closely matched to ground truth.

4.4 Rodent Hippocampal Sequences

Finally, we tested PP-Seq and its time warping variant on a hippocampal recording in a rat making repeated runs down a linear track.³ This dataset is larger ($T \approx 16$ minutes, $S = 137,482$) and contains less stereotyped sequences than the songbird data. From prior work [56], we expect to see two sequences with overlapping populations of neurons, corresponding to the two running directions on the track. PP-Seq reveals these expected sequences in an unsupervised manner—i.e. without reference to the rat’s position—as shown in fig. 5A-C.

We performed a large cross-validation sweep over 2,000 random hyperparameter settings for this dataset (see Supplement F.2). This confirmed that models with $R = 2$ sequence performed well in terms of heldout performance (fig. 5D). Interestingly, despite variability in running speeds, this same analysis did not show a consistent benefit to including larger time warping factors into the model (fig. 5E). Higher performing models were characterized by larger sequence amplitudes, i.e. larger values of $\mathbb{E}[A_k] = \alpha/\beta$, and smaller background rates, i.e. smaller values of λ^ϕ . Other parameters had less pronounced effects on performance. Overall, these results demonstrate that PP-Seq can be fruitfully applied to large-scale and “messy” neural datasets, that hyperparameters can be tuned by cross-validation, and that the unsupervised learning of neural sequences conforms to existing scientific understanding gained via supervised methods.

5 Conclusion

We proposed a point process model (PP-Seq) inspired by convolutive NMF [4, 8, 9] to identify neural sequences. Both models approximate neural activity as the sum of nonnegative features, drawn from a fixed number of spatiotemporal motif types. Unlike convNMF, PP-Seq restricts each motif to be a true sequence—the impulse responses are Gaussian and hence unimodal. Further, PP-Seq is formulated in a probabilistic framework that better quantifies uncertainty (see fig. 3) and handles low firing rate regimes (see fig. 4). Finally, PP-Seq produces an extremely sparse representation of sequence events in continuous time, opening the door to a variety of model extensions including the

³These data are available at <http://crcns.org/data-sets/hc/hc-11>; originally published in [56].

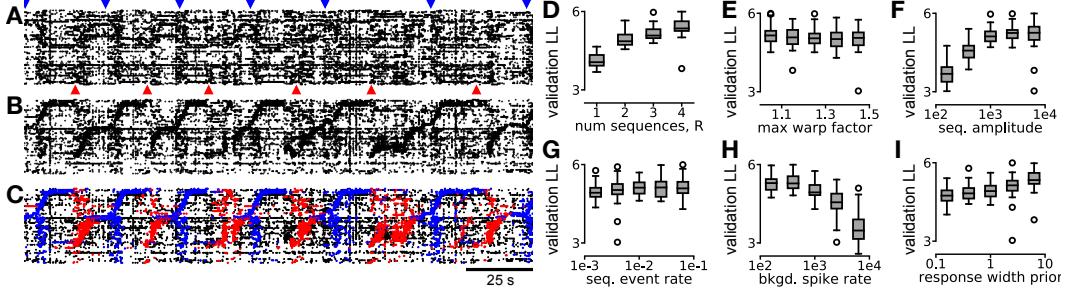


Figure 5: Automatic detection of place coding sequences in rat hippocampus. (A) Raw spike train (~20% of the full dataset is shown). Blue and red arrowheads indicate when the mouse reaches the end of the track and reverses direction. (B) Neurons re-sorted by PP-Seq (with $R = 2$). (C) Sequences annotated by PP-Seq. (D-I) Validation log-likelihoods as a function of hyperparameter value. Each boxplot summarizes the 50 highest scoring models, randomly sampling the other hyperparameters.

introduction of time warping (see fig. 4F), as well as other possibilities like truncated sequences and “clusterless” observations [57], which could be explored in future work.

Despite these benefits, fitting PP-Seq involves a tackling a challenging trans-dimensional inference problem inherent to Neyman-Scott point processes. We took several important steps towards overcoming this challenge by connecting these models to a more general class of Bayesian mixture models [41], developing and parallelizing a collapsed Gibbs sampler, and devising an annealed sampling approach to promote fast mixing. These innovations are sufficient to fit PP-Seq on datasets containing hundreds of thousands of spikes in just a few minutes on a modern laptop.

Acknowledgements

A.H.W. received funding support from the National Institutes of Health BRAIN initiative (1F32MH122998-01), and the Wu Tsai Stanford Neurosciences Institute Interdisciplinary Scholar Program. S.W.L. was supported by grants from the Simons Collaboration on the Global Brain (SCGB 697092) and the NIH BRAIN Initiative (U19NS113201 and R01NS113119). We thank the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

Broader Impact

Understanding neural computations in biological systems and ultimately the human brain is a grand and long-term challenge with broad implications for human health and society. The field of neuroscience is still taking early and incremental steps towards this goal. Our work develops a general-purpose, unsupervised method for identifying an important structure—neural sequences—which have been observed in a variety of experimental datasets and have been studied extensively by theorists. This work will serve to advance this growing understanding by providing new analytical tools for neuroscientists. We foresee no immediate impacts, positive or negative, concerning the general public.

References

- [1] Mark S Goldman. “Memory without feedback in a neural network”. *Neuron* 61.4 (2009), pp. 621–634.
- [2] Richard H R Hahnloser, Alexay A Kozhevnikov, and Michale S Fee. “An ultra-sparse code underlies the generation of neural sequences in a songbird”. *Nature* 419.6902 (2002), pp. 65–70.
- [3] Howard Eichenbaum. “Time cells in the hippocampus: a new dimension for mapping memories”. *Nat. Rev. Neurosci.* 15.11 (2014), pp. 732–744.
- [4] Emily L Mackevicius, Andrew H Bahle, Alex H Williams, Shijie Gu, Natalia I Denisenko, Mark S Goldman, and Michale S Fee. “Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience”. *Elife* 8 (2019).
- [5] Moshe Abeles and Itay Gat. “Detecting precise firing sequences in experimental data”. *J. Neurosci. Methods* 107.1-2 (2001), pp. 141–154.

- [6] Eleonora Russo and Daniel Durstewitz. “Cell assemblies at multiple time scales with arbitrary lag constellations”. *Elife* 6 (2017).
- [7] Pietro Quaglio, Vahid Rostami, Emiliano Torre, and Sonja Grün. “Methods for identification of spike patterns in massively parallel spike trains”. *Biol. Cybern.* 112.1-2 (2018), pp. 57–80.
- [8] Sven Peter, Elke Kirschbaum, Martin Both, Lee Campbell, Brandon Harvey, Conor Heins, Daniel Durstewitz, Ferran Diego, and Fred A Hamprecht. “Sparse convolutional coding for neuronal assembly detection”. *Advances in Neural Information Processing Systems 30*. Ed. by I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett. Curran Associates, Inc., 2017, pp. 3675–3685.
- [9] Paris Smaragdis. “Convulsive speech bases and their application to supervised speech separation”. *IEEE Trans. Audio Speech Lang. Processing* (2006).
- [10] Thomas J Davidson, Fabian Kloosterman, and Matthew A Wilson. “Hippocampal replay of extended experience”. *Neuron* 63.4 (2009), pp. 497–507.
- [11] Anne C Smith and Emery N Brown. “Estimating a state-space model from point process observations”. *Neural Computation* 15.5 (2003), pp. 965–991.
- [12] K L Briggman, H D I Abarbanel, and W B Kristan Jr. “Optical imaging of neuronal populations during decision-making”. *Science* 307.5711 (2005), pp. 896–901.
- [13] 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”. *J. Neurophysiol.* 102.1 (2009), pp. 614–635.
- [14] Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. “A new look at state-space models for neural data”. *Journal of Computational Neuroscience* 29.1-2 (2010), pp. 107–126.
- [15] Jakob H Macke, Lars Buesing, John P Cunningham, M Yu Byron, Krishna V Shenoy, and Maneesh Sahani. “Empirical models of spiking in neural populations”. *Advances in Neural Information Processing Systems*. 2011, pp. 1350–1358.
- [16] David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. “Robust learning of low-dimensional dynamics from large neural ensembles”. *Advances in Neural Information Processing Systems*. 2013, pp. 2391–2399.
- [17] Peiran Gao and Surya Ganguli. “On simplicity and complexity in the brave new world of large-scale neuroscience”. *Curr. Opin. Neurobiol.* 32 (2015), pp. 148–155.
- [18] Yuanjun Gao, Evan Archer, Liam Paninski, and John P Cunningham. “Linear dynamical neural population models through nonlinear embeddings” (2016). arXiv: 1605.08454 [q-bio.NC].
- [19] Yuan Zhao and II Memming Park. “Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains”. *Neural Comput.* 29.5 (2017), pp. 1293–1316.
- [20] Anqi Wu, Stan Pashkovski, Sandeep R Datta, and Jonathan W Pillow. “Learning a latent manifold of odor representations from neural responses in piriform cortex”. *Advances in Neural Information Processing Systems 31*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 5378–5388.
- [21] Alex H Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I Ryu, Krishna V Shenoy, Mark Schnitzer, Tamara G Kolda, and Surya Ganguli. “Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis”. *Neuron* 98.6 (2018), 1099–1115.e8.
- [22] 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 *C. elegans*”. 2019.
- [23] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. “Learning interpretable continuous-time models of latent stochastic dynamical systems”. *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 1726–1734.
- [24] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. “High-dimensional geometry of population responses in visual cortex”. *Nature* 571.7765 (2019), pp. 361–365.
- [25] Christopher D Harvey, Philip Coen, and David W Tank. “Choice-specific sequences in parietal cortex during a virtual-navigation decision task”. *Nature* 484.7392 (2012), pp. 62–68.

- [26] Mikhail I Rabinovich, Ramón Huerta, Pablo Varona, and Valentin S Afraimovich. “Generation and reshaping of sequences in neural systems”. *Biol. Cybern.* 95.6 (2006), pp. 519–536.
- [27] Emily Lambert Mackevicius and Michale Sean Fee. “Building a state space for song learning”. *Curr. Opin. Neurobiol.* 49 (2018), pp. 59–68.
- [28] György Buzsáki and David Tingley. “Space and time: the hippocampus as a sequence generator”. *Trends Cogn. Sci.* 22.10 (2018), pp. 853–869.
- [29] Eva Pastalkova, Vladimir Itskov, Asohan Amarasingham, and György Buzsáki. “Internally generated cell assembly sequences in the rat hippocampus”. *Science* 321.5894 (2008), pp. 1322–1327.
- [30] Daoyun Ji and Matthew A Wilson. “Coordinated memory replay in the visual cortex and hippocampus during sleep”. *Nat. Neurosci.* 10.1 (2007), pp. 100–107.
- [31] Hannah R Joo and Loren M Frank. “The hippocampal sharp wave-ripple in memory retrieval for immediate use and consolidation”. *Nat. Rev. Neurosci.* 19.12 (2018), pp. 744–757.
- [32] Wei Xu, Felipe de Carvalho, and Andrew Jackson. “Sequential neural activity in primary motor cortex during sleep”. *J. Neurosci.* 39.19 (2019), pp. 3698–3712.
- [33] David Tingley and Adrien Peyrache. “On the methods for reactivation and replay analysis”. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375.1799 (2020), p. 20190231.
- [34] Kourosh Maboudi, Etienne Ackermann, Laurel Watkins de Jong, Brad E Pfeiffer, David Foster, Kamran Diba, and Caleb Kemere. “Uncovering temporal structure in hippocampal output patterns”. *Elife* 7 (2018).
- [35] Lukas Grossberger, Francesco P Battaglia, and Martin Vinck. “Unsupervised clustering of temporal patterns in high-dimensional neuronal ensembles using a novel dissimilarity measure”. *PLoS Comput. Biol.* 14.7 (2018), e1006283.
- [36] Roemer van der Meij and Bradley Voytek. “Uncovering neuronal networks defined by consistent between-neuron spike timing from neuronal spike recordings”. *eNeuro* 5.3 (2018).
- [37] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. *Nature* 401.6755 (1999), pp. 788–791.
- [38] Jesper Moller and Rasmus Plenge Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Taylor & Francis, 2003.
- [39] Isabel Valera Manuel Gomez Rodriguez. *Learning with Temporal Point Processes*. Tutorial at ICML, 2018.
- [40] Jerzy Neyman and Elizabeth L Scott. “Statistical approach to problems of cosmology”. *J. R. Stat. Soc. Series B Stat. Methodol.* 20.1 (1958), pp. 1–29.
- [41] Jeffrey W Miller and Matthew T Harrison. “Mixture models with a prior on the number of components”. *J. Am. Stat. Assoc.* 113.521 (2018), pp. 340–356.
- [42] John Frank Charles Kingman. *Poisson processes*. Clarendon Press, 2002.
- [43] Lea Duncker and Maneesh Sahani. “Temporal alignment and latent Gaussian process factor inference in population spike trains”. *Advances in Neural Information Processing Systems 31*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 10445–10455.
- [44] Alex H Williams, Ben Poole, Niru Maheswaranathan, Ashesh K Dhawale, Tucker Fisher, Christopher D Wilson, David H Brann, Eric M Trautmann, Stephen Ryu, Roman Shusterman, Dmitry Rinberg, Bence P Ölveczky, Krishna V Shenoy, and Surya Ganguli. “Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping”. *Neuron* 105.2 (2020), 246–259.e8.
- [45] Ushio Tanaka, Yoshihiko Ogata, and Dietrich Stoyan. “Parameter estimation and model selection for Neyman–Scott point processes”. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.1 (2008), pp. 43–57.
- [46] Ushio Tanaka and Yoshihiko Ogata. “Identification and estimation of superposed Neyman–Scott spatial cluster processes”. *Ann. Inst. Stat. Math.* 66.4 (2014), pp. 687–702.
- [47] Jiří Kopecký and Tomáš Mrkvíčka. “On the Bayesian estimation for the stationary Neyman–Scott point processes”. *Appl. Math.* 61.4 (2016), pp. 503–514.
- [48] Yoshihiko Ogata. “Cluster analysis of spatial point patterns: posterior distribution of parents inferred from offspring”. *Japanese Journal of Statistics and Data Science* (2019).
- [49] Radford M Neal. “Markov chain sampling methods for Dirichlet process mixture models”. *J. Comput. Graph. Stat.* 9.2 (2000), pp. 249–265.

- [50] Jun S Liu, Wing Hung Wong, and Augustine Kong. “Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes”. *Biometrika* 81.1 (1994), pp. 27–40.
- [51] Sonia Jain and Radford M Neal. “A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model”. *J. Comput. Graph. Stat.* 13.1 (2004), pp. 158–182.
- [52] Sonia Jain and Radford M Neal. “Splitting and merging components of a nonconjugate Dirichlet process mixture model”. *Bayesian Anal.* 2.3 (2007), pp. 445–472.
- [53] Elaine Angelino, Matthew James Johnson, and Ryan P Adams. “Patterns of scalable Bayesian inference”. *Foundations and Trends® in Machine Learning* 9.2-3 (2016), pp. 119–247.
- [54] Svante Wold. “Cross-validatory estimation of the number of components in factor and principal components models”. *Technometrics* 20.4 (1978), pp. 397–405.
- [55] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. “Julia: A fresh approach to numerical computing”. *SIAM review* 59.1 (2017), pp. 65–98.
- [56] Andres D Grosmark and György Buzsáki. “Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences”. *Science* 351.6280 (2016), pp. 1440–1443.
- [57] Xinyi Deng, Daniel F Liu, Kenneth Kay, Loren M Frank, and Uri T Eden. “Clusterless decoding of position from multiunit activity using a marked point process filter”. *Neural computation* 27.7 (2015), pp. 1438–1460.

Supplementary Material: Point process models for sequence detection in high-dimensional spike trains

A Background

This section describes some basic properties and characteristics of Poisson processes that are relevant to understand the PP-Seq model. See Kingman [1] for a targeted introduction to Poisson processes and Daley et al. [2] for a broader introduction to point processes.

Definition of a Poisson Process — A point process model satisfying the following three properties is called a *Poisson process*.

- (i) The number of events falling within any region of the sample space \mathcal{X} is a random variable following a Poisson distribution. For a region $R \subset \mathcal{X}$ we denote the number of events occurring in R as $N(R)$.
- (ii) There exists a function $\lambda(x) : \mathcal{X} \mapsto \mathbb{R}_+$, called the *intensity function*, which satisfies $N(R) \sim \text{Poisson}(\int_R \lambda(x) dx)$ for every region $R \subset \mathcal{X}$.
- (iii) For any two non-overlapping regions R and R' , $N(R)$ and $N(R')$ are independent random variables.

We write $\{x_i\}_{i=1}^n \sim PP(\lambda)$ to denote that a set of events $\{x_i\}_{i=1}^n = \{x_1, \dots, x_n\}$ is distributed according to a Poisson process with intensity function λ .

Marked Poisson Processes — A *marked* Poisson process extends the above definition to sets of events marked with additional information. For example, we write $\{(x_i, m_i)\}_{i=1}^n \sim PP(\lambda(x, m))$ to denote a marked Poisson process where $m_i \in \mathcal{M}$ is the “mark” associated with each event. Note that the domain of the intensity function is extended to $\mathcal{X} \times \mathcal{M}$. To pick a concrete example of interest, the events could represent spike times, in which case \mathcal{X} would be an interval of the real line, e.g. $[0, T]$, representing the recorded period. Further, \mathcal{M} would be a discrete set of neuron labels, e.g. $\{1, \dots, N\}$ for a recording of N cells.

Likelihood function — Given a set of events $\{x_i\}_{i=1}^n$ in \mathcal{X} the likelihood of this set, with respect to Poisson process $PP(\lambda)$ is:

$$p(\{x_i\}_{i=1}^n | \lambda) = \prod_{i=1}^n \lambda(x_i) \cdot \exp \left\{ - \int_{\mathcal{X}} \lambda(x) dx \right\} \quad (8)$$

Sampling — Given an intensity function $\lambda(x)$ and a sample space \mathcal{X} , we can simulate data from a Poisson process by first sampling the number of events:

$$N \sim \text{Poisson} \left(\int_{\mathcal{X}} \lambda(x) dx \right) \quad (9)$$

and then sampling the location of each event according to the normalized intensity function. That is, for $i = \{1, \dots, N\}$, we independently sample the events according to:

$$x_i \sim F \quad \text{where the p.d.f. of } F \text{ is } \frac{\lambda(x)}{\int_{\mathcal{X}} \lambda(x) dx} \quad (10)$$

Superposition Principle — The union of K independent Poisson processes with nonnegative intensity functions $\lambda_1(x), \dots, \lambda_K(x)$ is a Poisson process with intensity function $\Lambda(x) = \sum_{k=1}^K \lambda_k(x)$.

As a consequence, we can sample from a complicated Poisson process by first breaking it down into a series of simpler Poisson processes, drawing independent samples from these simpler processes, and taking the union of all events to achieve a sample from the original Poisson process. We exploit this property in section B.1 to draw samples from the PP-Seq model.

B The PP-Seq Model

B.1 Generative Process

Recall the generative model described in the main text. Let $z_k = (\tau_k, r_k, A_k)$ denote the k -th latent sequence, which consists of a time τ_k , type r_k , and amplitude A_k . Extensions of the model introduce further properties like time warping

factors ω_k for each latent sequence. The latent events (i.e. instantiations of a sequence) are sampled from a Poisson process,

$$\{(\tau_k, r_k, A_k)\}_{k=1}^K \sim \text{PP}(\lambda(\tau, r, A)) \quad (11)$$

$$\lambda(\tau, r, A) = \psi \pi_r \text{Ga}(A; \alpha, \beta). \quad (12)$$

Note that this is a *homogeneous* Poisson process since the intensity function does not depend on time. Conditioned on the latent events, the observed spikes are then sampled from a second, inhomogeneous, Poisson process,

$$\{(t_s, n_s)\}_{s=1}^S \sim \text{PP}(\lambda(t, n | \{(\tau_k, r_k, A_k)\}_{k=1}^K)) \quad (13)$$

$$\lambda(t, n | \{(\tau_k, r_k, A_k)\}_{k=1}^K) = \lambda^\varnothing(t, n) + \sum_{k=1}^K g(t, n | \tau_k, r_k, A_k) \quad (14)$$

$$g(t, n | \tau_k, r_k, A_k) = A_k \cdot a_{nr_k} \cdot \mathcal{N}(t | \tau_k + b_{nr_k}, c_{nr_k}). \quad (15)$$

In the main text we abbreviate $\lambda_n(t) \triangleq \lambda(t, n | \{(\tau_k, r_k, A_k)\}_{k=1}^K)$, $\lambda^\varnothing(t, n) \triangleq \lambda_n^\varnothing$, and $g_n(t | z_k) \triangleq g(t, n | \tau_k, r_k, A_k)$.

Observe that the firing rates are sums of non-negative intensity functions. We can use the superposition principle of Poisson processes (see section A) to write the observed spikes as the union of spikes generated by the background and by each of the latent events.

$$\{(t_s^{(0)}, n_s^{(0)})\}_{s=1}^{S_0} \sim \text{PP}(\lambda^\varnothing(t, n)) \quad (16)$$

$$\{(t_s^{(k)}, n_s^{(k)})\}_{s=1}^{S_k} \sim \text{PP}(A_k \cdot a_{nr_k} \cdot \mathcal{N}(t | \tau_k + b_{nr_k}, c_{nr_k})) \quad \text{for } k = 1, \dots, K \quad (17)$$

$$\{(t_s, n_s)\}_{s=1}^S = \bigcup_{k=0}^K \{(t_s^{(k)}, n_s^{(k)})\}_{s=1}^{S_k}, \quad (18)$$

where S_k denotes the number of spikes generated by the k -th component, with $k = 0$ denoting the background spikes.

Next, note that each of the constituent Poisson processes intensity functions are simple, in that the normalized intensities correspond to simple densities. In the case of the background, the normalized intensity is categorical in the neuron indices and uniform in time. The impulse responses, once normalized, are categorical in the neuron indices and conditionally Gaussian in time. We use these facts to sample the Poisson processes as described in Section A.

The final sampling procedure for PP-Seq is as follows. First sample the global parameters,

$$\lambda^\varnothing \sim \text{Gamma}(\alpha_\varnothing, \beta_\varnothing) \quad (19)$$

$$\pi^\varnothing \sim \text{Dirichlet}(\gamma_\varnothing \cdot \mathbf{1}_N) \quad (20)$$

$$\pi \sim \text{Dirichlet}(\gamma \cdot \mathbf{1}_R) \quad (21)$$

$$\mathbf{a}_r \sim \text{Dirichlet}(\varphi \cdot \mathbf{1}_N) \quad \text{for } r \in \{1, \dots, R\} \quad (22)$$

$$c_{nr} \sim \text{Inv-}\chi^2(\nu, \sigma^2) \quad \text{for } (n, r) \in \{1, \dots, N\} \times \{1, \dots, R\} \quad (23)$$

$$b_{nr} \sim \text{Normal}(0, c_{nr}/\kappa) \quad \text{for } (n, r) \in \{1, \dots, N\} \times \{1, \dots, R\}. \quad (24)$$

Then sample the latent events,

$$K \sim \text{Poisson}(\psi \cdot T) \quad (25)$$

$$r_k \sim \text{Categorical}(\pi) \quad \text{for } k \in \{1, \dots, K\} \quad (26)$$

$$\tau_k \sim \text{Uniform}([0, T]) \quad \text{for } k \in \{1, \dots, K\} \quad (27)$$

$$A_k \sim \text{Gamma}(\alpha, \beta) \quad \text{for } k \in \{1, \dots, K\}. \quad (28)$$

Next sample the background spikes,

$$S_\varnothing \sim \text{Poisson}(\lambda^\varnothing \cdot T) \quad (29)$$

$$n_0^{(s)} \sim \text{Categorical}(\pi^\varnothing) \quad \text{for } s \in \{1, \dots, S_\varnothing\} \quad (30)$$

$$t_0^{(s)} \sim \text{Uniform}([0, T]) \quad \text{for } s \in \{1, \dots, S_\varnothing\} \quad (31)$$

Note that we have decomposed each neuron's background firing rate as the product: $\lambda_n^\varnothing = \lambda^\varnothing \pi_n^\varnothing$. Finally, sample the induced spikes in each sequence,

$$S_k \sim \text{Poisson}(A_k) \quad \text{for } k \in \{1, \dots, K\} \quad (32)$$

$$n_k^{(s)} \sim \text{Categorical}(\mathbf{a}_{r_k}) \quad \text{for } s \in \{1, \dots, S_k\}; \text{ for } k \in \{1, \dots, K\} \quad (33)$$

$$t_k^{(s)} \sim \text{Normal}(\tau_k + b_{nsr_k}, c_{nsr_k}) \quad \text{for } s \in \{1, \dots, S_k\}; \text{ for } k \in \{1, \dots, K\} \quad (34)$$

The hyperparameters of the model are $\xi = \{\gamma, \gamma_\emptyset, \alpha_\emptyset, \beta_\emptyset, \varphi, \nu, \sigma^2, \kappa, \psi, \alpha, \beta\}$, and the global parameters consist of $\Theta = \{\theta_\emptyset, \{\theta_r\}_{r=1}^R\}$, where $\theta_\emptyset = \{\lambda^\emptyset, \{\pi_n^\emptyset\}_{n=1}^N\}$ denotes parameters related to the background spiking process and $\theta_r = \{\pi_r, \mathbf{a}_r, \{b_{nr}, c_{nr}\}_{n=1}^N\}$ denotes parameters associated with each sequence type for $r \in \{1, \dots, R\}$. The resulting dataset is the union of background and induced spikes, $\bigcup_{k=0}^K \bigcup_{s=1}^{S_k} \{(n_k^{(s)}, t_k^{(s)})\}$.

B.2 Expected Sequence Sizes

As described in the main text, each sequence type induces a different pattern of activation across the neural population. In particular, each neuron's response is modeled by a scaled normal distribution described by three variables—the amplitude (a_{nr}), the mean (b_{nr}), and the variance (c_{nr}).

We can interpret b_{nr} as the temporal offset or delay for neuron n , while c_{nr} sets the width of the response. They are drawn from a *normal-inverse-chi-squared* distribution, which is a standard choice for the conjugate prior when inferring the mean and variance of a univariate normal distribution. Refer to [3] (section 3.3) and [4] (section 4.6.3.7) for appropriate background material.

We can interpret a_{nr} as the expected number of spikes emitted by neuron n for a sequence of type r with unit amplitude (i.e., when $A_k = 1$). By sampling $\mathbf{a}_r \in \mathbb{R}_+^N$ from a symmetric Dirichlet prior, we introduce the constraint that $\sum_n a_{nr} = 1$, which avoids a degeneracy in the amplitude parameters over neurons (a_{nr}) and sequences (A_k). As a consequence, we can interpret A_k as the expected number of spikes evoked by latent event k over the full neural population:

$$\mathbb{E}[S_k] = \sum_{n=1}^N \int_0^T A_k \cdot a_{nr} \cdot \mathcal{N}(t | \tau_k + b_{nr}, c_{nr}) dt = A_k \cdot \underbrace{\sum_{n=1}^N a_{nr}}_{=1} \cdot \underbrace{\int_0^T \mathcal{N}(t | \tau_k + b_{nr}, c_{nr}) dt}_{\approx 1} \approx A_k. \quad (35)$$

This approximation ignores boundary effects, but it is justified when $0 \ll \tau_k \ll T$ and $b_{nr}, \sqrt{c_{nr}} \ll T$; i.e. when the sequences are short relative to the interval $[0, T]$. This condition will generally be satisfied.

B.3 Log Likelihood

Given a set of latent events $\{z_k\} = \{z_k\}_{k=1}^K$ and sequence type parameters $\{\theta_r\} = \{\theta_r\}_{r=1}^R$, we can evaluate the likelihood of the dataset $\{x_s\} = \{x_s\}_{s=1}^S$ as:

$$\log p(\{x_s\} | \{z_k\}, \{\theta_r\}) = \underbrace{\sum_{s=1}^S \log \left[\lambda^\emptyset \pi_{n_s}^\emptyset + \sum_{k=1}^K A_k a_{n_s r_k} \mathcal{N}(t_s | \tau_k + b_{n_s r_k}, c_{n_s r_k}) \right]}_{(*)} - \underbrace{\lambda^\emptyset T - \sum_{k=1}^K A_k}_{(**)} \quad (36)$$

where terms $(*)$ and $(**)$ respectively correspond to the logarithms of $[\prod_i \lambda(x_i)]$ and $[\exp[-\int_R \lambda(x) dx]]$ appearing in Equation (8).

C Complete Gibbs Sampling Algorithm

In this section we describe the collapsed Gibbs sampling routine for PP-Seq in detail. It adapts “Algorithm 3” of Neal [5] and the extension by Miller et al. [6] to draw approximate samples from the model posterior. To do this, we introduce auxiliary parent variables $u_s \in \{0\} \cup \mathbb{Z}_+$ for every spike. These variables denote assignments of each spike to one of the latent sequence events, $u_s = k$ for some $k \in \{0, \dots, K\}$, or to the background process, $u_s = 0$. Let $X_k = \{x_s : u_s = k\}$ denote the set of spikes assigned to sequence $k \in \{1, \dots, K\}$ under parent assignments $\{u_s\}_{s=1}^S$. Similar assignment variables are used in standard Gibbs sampling routines for finite mixture models (see, e.g., section 24.2.4 in [4]) and for Dirichlet Process Mixture (DPM) models (see [5]; section 25.2.4 in [4]). In these contexts, each u_s may be thought of as a sequence assignment or indicator variable. In the present context of the PP-Seq model, the attributions of spikes to latent events follows from the superposition principle of Poisson processes (see section A), which allows us to decompose each neuron's firing rate function into a sum of simple functions (in this case, univariate Gaussians) and thus derive efficient Gibbs sampler updates.

The Gibbs sampling routine is summarized in Algorithm 1. While the approach closely mirrors existing work by Neal [5] and Miller et al. [6] there are a couple notable differences. First, the background spikes in the PP-Seq model do not have an obvious counterpart in these prior algorithms, and this feature of the model renders the partition of datapoints only *partially* exchangeable (intuitively, the set of datapoints defined by $u_s = 0$ is treated differently from all other groups). Second, the factors $S_k + \alpha$ and $\alpha \left(\frac{\beta}{1+\beta} \right)^\alpha \psi$ in the collapsed Gibbs updates are specific to the special form of the PP-Seq model as a Neyman-Scott process; for example, in the analogous Gibbs sampling routine for DPMs,

Algorithm 1: Collapsed Gibbs sampling for PPSeq model

Input: Spikes $\{x_1, \dots, x_S\}$ and hyperparameters ξ .

Initialize by assigning all spikes to the background: $u_s = 0$ for $s \in \{1 \dots S\}$.

repeat M times to draw M samples

1. Resample parent assignments, integrating over latent events:

for $s = 1, \dots, S$. Remove x_s its current assignment and place it...

a. in the background, $u_s = 0$, with probability $\propto (1 + \beta) \lambda_{n_s}^\varnothing$

b. in cluster $u_s = k$, with probability

$$\propto (\alpha + |X_k|) \left[\sum_{r_k=1}^R p(r_k | X_k) a_{n_s r_k} p(t_s | X_k, r_k, n_s) \right]$$

c. in a new cluster, $u_s = K + 1$, with probability $\propto \alpha \left(\frac{\beta}{1+\beta} \right)^\alpha \psi \sum_{r=1}^R \pi_r a_{n_s r}$

end

2. Resample latent events

for $k = 1, \dots, K$ sample $r_k, \tau_k, A_k \sim p(r_k, \tau_k, A_k | X_k, \Theta)$ with the following steps:

a. Sample $r_k \sim p(r_k | X_k, \Theta, \xi)$ where

$$p(r_k = r | X_k, \Theta, \xi) \propto \pi_{r_k} \left(\prod_{x_s \in X_k} a_{n_s r_k} \right) \frac{Z(\sum_{x_s \in X_k} J_{sk}, \sum_{x_s \in X_k} h_{sk})}{\prod_{x_s \in X_k} Z(J_{sk}, h_{sk})}$$

and where $J_{sk} = 1/c_{n_s r_k}$, $h_{sk} = (t_s - b_{n_s r_k})/c_{n_s r_k}$, and
 $Z(J, h) = (2\pi)^{1/2} J^{-1/2} \exp\left\{-\frac{1}{2}h^2 J^{-1}\right\}$

b. Sample $\tau_k \sim p(\tau_k | r_k, X_k, \Theta)$ where,

$$p(\tau_k | r_k, X_k, \Theta) = \mathcal{N}(\tau_k | \mu_k, \sigma_k^2)$$

and where $\sigma_k^2 = (\sum_{x_s \in X_k} J_{sk})^{-1}$ and $\mu_k = \sigma_k^2 (\sum_{x_s \in X_k} h_{sk})$

c. Sample $A_k \sim \text{Ga}(\alpha + |X_k|, \beta + 1)$.

end

3. Resample global parameters

a. Sample $\lambda_n^\varnothing \sim \text{Ga}(\alpha_\varnothing + \sum_{x_s \in X_0} \mathbb{I}[n_s = n], \beta_\varnothing + T)$ for $n = 1, \dots, N$.

b. Sample $\pi \sim \text{Dir}([\gamma + \sum_k \mathbb{I}[r_k = 1], \dots, \gamma + \sum_k \mathbb{I}[r_k = R]])$

c. Sample $\mathbf{a}_r \sim \text{Dir}([\varphi + \sum_k \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = 1], \dots, \varphi + \sum_k \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = N]])$

d. Sample $b_{nr}, c_{nr} \sim \text{Inv}-\chi^2(c_{nr}; \nu_{nr}, \sigma_{nr}^2) \mathcal{N}(b_{nr} | \mu_{nr}, c_{nr}/\kappa_{nr})$ for $n = 1 \dots, N$,
 $r = 1 \dots, R$ where

$$S_{nr} = \sum_{k=1}^{K^*} \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = n]$$

$$\nu_{nr} = \nu + S_{nr} \quad \sigma_{nr}^2 = \frac{\nu \sigma^2 + \sum_{k=1}^{K^*} \sum_{x_s \in X_k} (t_s - \tau_k)^2 \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\nu + S_{nr}},$$

$$\kappa_{nr} = \kappa + S_{nr} \quad \mu_{nr} = \frac{\sum_{k=1}^{K^*} \sum_{x_s \in X_k} (t_s - \tau_k) \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\kappa + S_{nr}}.$$

end

$(\alpha + |X_k|)$ is replaced with $|X_k|$, and $\alpha \psi T \left(\frac{\beta}{1+\beta}\right)^\alpha$ is replaced with the concentration parameter of the Dirichlet process.

D Derivations

D.1 Prior Distribution Over Partitions

First we derive the prior distribution on the partition of spikes under a Neyman-Scott process. Technically, we derive the prior distribution on spike *indices*, $\{1, \dots, S\}$, not yet including the spike times and neurons. A partition of indices is a set of disjoint, non-empty sets whose union is $\{1, \dots, S\}$. We represent the partition as $\mathcal{I} = \{\mathcal{I}_k\}_{k=0}^{K^*}$, where $\mathcal{I}_k \subset \{1, \dots, S\}$ contains the indices of spikes assigned to sequence k , with $k = 0$ denoting the background spikes. Here, K^* denotes the number of non-empty sequences in the partition, and $|\mathcal{I}_k|$ denotes the number of spikes assigned to the k -th sequence. Likewise, $|\mathcal{I}_0|$ denotes the number of spikes assigned to the background.

Proposition 1. *The prior probability of the partition, integrating over the latent event parameters, is,*

$$p(\mathcal{I} | \Theta, \xi) = V(K^*; \xi) \frac{\text{Po}(|\mathcal{I}_0|; \lambda^\varnothing T) |\mathcal{I}_0|!}{S! (1+\beta)^{S-|\mathcal{I}_0|}} \prod_{k=1}^{K^*} \frac{\Gamma(|\mathcal{I}_k| + \alpha)}{\Gamma(\alpha)}, \quad (37)$$

where α and β are the hyperparameters of the gamma intensity on amplitudes, $\lambda^\varnothing T$ is the expected number of background events, and

$$V(K^*; \xi) = \sum_{K=K^*}^{\infty} \text{Po}(K; \psi T) \frac{K!}{(K - K^*)!} \left(\frac{\beta}{1+\beta}\right)^{\alpha K}. \quad (38)$$

Proof. Our derivation roughly follows that of Miller et al. [6] for mixture of finite mixture models, but we adapt it to Neyman-Scott processes. The major difference is that the distribution above is over partitions of random size. First, return to the generative process described in Section B.1 and integrate over the latent event amplitudes to obtain the marginal distribution on sequence sizes given the total number of sequences K . We have,

$$p(S_0, \dots, S_K | K, \Theta, \xi) = \text{Po}(S_0; \lambda^\varnothing T) \prod_{k=1}^K \int \text{Po}(S_k; \gamma_k) \text{Ga}(A_k; \alpha, \beta) dA_k \quad (39)$$

$$= \text{Po}(S_0; \lambda^\varnothing T) \prod_{k=1}^K \text{NB}(S_k; \alpha, (1+\beta)^{-1}) \quad (40)$$

$$= \frac{1}{S_0!} (\lambda^\varnothing T)^{S_0} e^{-\lambda^\varnothing T} \prod_{k=1}^K \frac{\Gamma(S_k + \alpha)}{N_k! \Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^\alpha \left(\frac{1}{1+\beta}\right)^{S_k} \quad (41)$$

$$= \frac{1}{S_0!} (\lambda^\varnothing T)^{S_0} e^{-\lambda^\varnothing T} \left(\frac{\beta}{1+\beta}\right)^{K\alpha} \left(\frac{1}{1+\beta}\right)^{S-S_0} \prod_{k=1}^K \frac{\Gamma(S_k + \alpha)}{S_k! \Gamma(\alpha)}. \quad (42)$$

Let $\{u_s\}_{s=1}^S$ denote a set of sequence assignments. There are $\binom{S}{S_0, \dots, S_K}$ parent assignments consistent with the sequence sizes S_0, \dots, S_K , and they are all equally likely under the prior, so the conditional probability of the parent assignments is,

$$p(\{u_s\}_{s=1}^S | K, \Theta, \xi) = \frac{1}{S_0!} (\lambda^\varnothing T)^{S_0} e^{-\lambda^\varnothing T} \prod_{k=1}^K \frac{\Gamma(S_k + \alpha)}{S_k! \Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^\alpha \left(\frac{1}{1+\beta}\right)^{N_k} \binom{S}{S_0, \dots, S_K}^{-1} \quad (43)$$

$$= \frac{1}{S!} (\lambda^\varnothing T)^{S_0} e^{-\lambda^\varnothing T} \left(\frac{\beta}{1+\beta}\right)^{K\alpha} \left(\frac{1}{1+\beta}\right)^{S-S_0} \prod_{k=1}^K \frac{\Gamma(S_k + \alpha)}{\Gamma(\alpha)}. \quad (44)$$

The parent assignments above induce a partition, but technically they assume a particular ordering of the latent events. Moreover, if some of the latent events fail to produce any observed events, they will not be included in the partition. In performing a change of variables from parent assignments to partitions, we need to sum over latent event assignments that produce the same partition. There are $\binom{K}{K^*} K^*! = \frac{K!}{(K-K^*)!}$ such assignments if there are K latent events but only K^* partitions. Thus,

$$p(\mathcal{I} | K, \Theta, \xi) = \frac{K!}{(K-K^*)!} \frac{1}{S!} (\lambda^\varnothing T)^{|\mathcal{I}_0|} e^{-\lambda^\varnothing T} \left(\frac{\beta}{1+\beta}\right)^{K\alpha} \left(\frac{1}{1+\beta}\right)^{S-|\mathcal{I}_0|} \prod_{k=1}^{K^*} \frac{\Gamma(|\mathcal{I}_k| + \alpha)}{\Gamma(\alpha)}. \quad (45)$$

Clearly, K must be at least K^* in order to produce the partition.

Finally, we sum over the number of latent events K to obtain the marginal probability of the partition,

$$p(\mathcal{I} \mid \Theta, \xi) = \sum_{K=K^*}^{\infty} \text{Po}(K; \psi T) p(\mathcal{I} \mid K, \Theta, \xi) \quad (46)$$

$$= V(K^*; \xi) \frac{(\lambda^\varnothing T)^{|\mathcal{I}_0|} e^{-\lambda^\varnothing T}}{S! (1+\beta)^{S-|\mathcal{I}_0|}} \prod_{k=1}^{K^*} \frac{\Gamma(|\mathcal{I}_k| + \alpha)}{\Gamma(\alpha)}, \quad (47)$$

where

$$V(K^*; \xi) = \sum_{K=K^*}^{\infty} \text{Po}(K; \psi T) \frac{K!}{(K-K^*)!} \left(\frac{\beta}{1+\beta} \right)^{K\alpha}. \quad (48)$$

□

D.2 Marginal Likelihood of Spikes in a Partition

Now we combine the prior on partitions of spike indices with a likelihood of spikes under that partition. To match the notation in the main text and Section C, let $X = \{X_k\}_{k=0}^{K^*}$ denote a partition of the spikes where $X_k = \{x_s : s \in \mathcal{I}_k\}$ denote the set of spikes in sequence k . In this section we derive the marginal probability of spikes in a sequence, integrating out the sequence time and type. For the background spikes, there are no parameters to integrate and the marginal probability is simply,

$$p(X_0 \mid \Theta) = \prod_{x_s \in X_0} \text{Unif}(t_s; [0, T]) \text{Cat}(n_s \mid [\lambda_1^\varnothing / \lambda^\varnothing, \dots, \lambda_N^\varnothing / \lambda^\varnothing]) \quad (49)$$

$$= T^{-|X_0|} \prod_{x_s \in X_0} \frac{\lambda_{n_s}^\varnothing}{\lambda^\varnothing}. \quad (50)$$

For the spikes attributed to a latent event, however, we have to marginalize over the type and time of the event. We have,

$$p(X_k \mid \Theta) = \sum_{r_k=1}^R p(r_k) \int_0^T p(\tau_k) \prod_{s \in \mathcal{I}_k} p(t_s, n_s \mid r_k, \tau_k, \Theta) d\tau_k \quad (51)$$

$$= \sum_{r_k=1}^R \pi_{r_k} \int_0^T \text{Unif}(\tau_k; [0, T]) \prod_{x_s \in X_k} \text{Cat}(n_s \mid r_k, \Theta) \mathcal{N}(t_s \mid \tau_k + b_{n_s r_k}, c_{n_s r_k}) d\tau_k \quad (52)$$

$$= \frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} \int_0^T \prod_{x_s \in X_k} \frac{a_{n_s r_k}}{\sqrt{2\pi c_{n_s r_k}}} \exp \left\{ -\frac{1}{2c_{n_s r_k}} (t_s - \tau_k - b_{n_s r_k})^2 \right\} d\tau_k. \quad (53)$$

Let $J_{sk} = 1/c_{n_s r_k}$, and $h_{sk} = (t_s - b_{n_s r_k})/c_{n_s r_k}$. Then,

$$p(X_k \mid \Theta) = \frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} \int_0^T \prod_{x_s \in X_k} a_{n_s r_k} \sqrt{\frac{J_{sk}}{2\pi}} \exp \left\{ -\frac{1}{2} J_{sk} \tau_k^2 + h_{sk} \tau_k - \frac{1}{2} h_{sk}^2 J_{sk}^{-1} \right\} d\tau_k \quad (54)$$

$$= \frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} \prod_{x_s \in X_k} \frac{a_{n_s r_k}}{Z(J_{sk}, h_{sk})} \int_0^T \exp \left\{ -\frac{1}{2} (\sum_{x_s \in X_k} J_{sk}) \tau_k^2 + (\sum_{x_s \in X_k} h_{sk}) \tau_k \right\} d\tau_k \quad (55)$$

$$\approx \frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} \left(\prod_{x_s \in X_k} a_{n_s r_k} \right) \frac{Z(\sum_{x_s \in X_k} J_{sk}, \sum_{x_s \in X_k} h_{sk})}{\prod_{x_s \in X_k} Z(J_{sk}, h_{sk})}, \quad (56)$$

where

$$Z(J, h) = (2\pi)^{1/2} J^{-1/2} \exp \left\{ \frac{1}{2} h^2 J^{-1} \right\} \quad (57)$$

is the normalizing constant of a Gaussian density in information form with precision J and linear coefficient h . For a sequence with a single spike, this reduces to,

$$p(x_s \mid \Theta) \approx \frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} a_{n_s r_k}. \quad (58)$$

The approximations above are due to the truncated integral over $[0, T]$ rather than the whole real line. In practice, this truncation is negligible for sequences far from the boundaries of the interval.

D.3 Collapsed Gibbs Sampling the Partitions

The conditional distribution on partitions given data and model parameters is given by,

$$p(\mathcal{I} \mid \{x_s\}_{s=1}^S, \Theta, \xi) \propto p(\mathcal{I} \mid \Theta, \xi) p(\{x_s\}_{s=1}^S \mid \mathcal{I}, \Theta) \quad (59)$$

$$= p(\mathcal{I} \mid \Theta, \xi) p(X_0 \mid \Theta) \prod_{k=1}^{K^*} p(X_k \mid \Theta). \quad (60)$$

Let $X \setminus x_s$ denote the partition with spike x_s removed. Though it is slightly inconsistent with the proof in Section D.1, let u_s denote the assignment of spike x_s , as in the main text. Moreover, let S_k denote the size of the k -th sequence in the partition, *not including spike* x_s . We now consider the relative probability of the full partition X when x_s is added to each one of the possible parts: the background, an existing sequence, or a new sequence.

For the background assignment,

$$p(u_s = 0 \mid x_s, X \setminus x_s, \Theta, \xi) \propto \frac{p(u_s = 0, x_s, X \setminus x_s, \Theta, \xi)}{p(X \setminus x_s, \Theta, \xi)} \quad (61)$$

$$= \frac{V(K^*; \xi) \frac{(\lambda^\varnothing T)^{S_0+1} e^{-\lambda^\varnothing T}}{S! (1+\beta)^{S-S_0-1}} p(X_0 \cup x_s \mid \Theta) \prod_{k=1}^{K^*} \frac{\Gamma(S_k + \alpha)}{\Gamma(\alpha)} p(X_k \mid \Theta)}{V(K^*; \xi) \frac{(\lambda^\varnothing T)^{S_0} e^{-\lambda^\varnothing T}}{S! (1+\beta)^{S-S_0}} p(X_0 \mid \Theta) \prod_{k=1}^{K^*} \frac{\Gamma(S_k + \alpha)}{\Gamma(\alpha)} p(X_k \mid \Theta)} \quad (62)$$

$$= \lambda^\varnothing T (1 + \beta) \frac{1}{T} \frac{\lambda^\varnothing}{\lambda^\varnothing} \quad (63)$$

$$= (1 + \beta) \lambda^\varnothing. \quad (64)$$

By a similar process, we arrive at the conditional probabilities of adding a spike to an existing sequence,

$$p(u_s = k \mid x_s, X \setminus x_s, \Theta, \xi) \propto \frac{\Gamma(S_k + \alpha + 1)}{\Gamma(S_k + \alpha)} \frac{p(X_k \cup x_s \mid \Theta)}{p(X_k \mid \Theta)} \quad (65)$$

$$= (S_k + \alpha) p(x_s \mid X_k, \Theta). \quad (66)$$

The predictive likelihood can be obtained via the ratio of marginal likelihoods derived above, or by explicitly calculating the categorical posterior distribution on types r_k and then the categorical and Gaussian posterior predictive distributions of n_s and t_s , respectively, given the type and the other spikes. The latter approach is what is presented in the main text.

Finally, the conditional probability of assigning a spike to a new sequence simplifies as,

$$p(u_s = K^* + 1 \mid X \setminus x_s, \Theta, \xi) \propto \left(\frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \right) \left(\frac{V(K^* + 1; \xi)}{V(K^*; \xi)} \right) p(x_s \mid \Theta) \quad (67)$$

$$= \alpha \left(\frac{\sum_{J=K^*+1}^{\infty} \frac{1}{J!} e^{-\psi T} (\psi T)^J \frac{J!}{(J-K^*-1)!} \left(\frac{\beta}{1+\beta} \right)^{J\alpha}}{\sum_{K=K^*}^{\infty} \frac{1}{K!} e^{-\psi T} (\psi T)^K \frac{K!}{(K-K^*)!} \left(\frac{\beta}{1+\beta} \right)^{K\alpha}} \right) \left(\frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} a_{n_s r_k} \right) \quad (68)$$

$$= \alpha \left(\frac{\sum_{K=K^*}^{\infty} (\psi T)^{K+1} \frac{1}{(K-K^*)!} \left(\frac{\beta}{1+\beta} \right)^{(K+1)\alpha}}{\sum_{K=K^*}^{\infty} (\psi T)^K \frac{1}{(K-K^*)!} \left(\frac{\beta}{1+\beta} \right)^{K\alpha}} \right) \left(\frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} a_{n_s r_k} \right) \quad (69)$$

$$= \alpha \left(\frac{\beta}{1+\beta} \right)^\alpha \psi T \left(\frac{1}{T} \sum_{r_k=1}^R \pi_{r_k} a_{n_s r_k} \right) \quad (70)$$

$$= \alpha \left(\frac{\beta}{1+\beta} \right)^\alpha \psi \sum_{r_k=1}^R \pi_{r_k} a_{n_s r_k}. \quad (71)$$

D.4 Gibbs Sampling the Latent Event Parameters

Given the partition and global parameters, it is straightforward to sample the latent events. In fact, most of the calculations were derived in Section D.2. The conditional distribution of the event type is,

$$p(r_k | X_k, \Theta) \propto p(r_k) \int_0^T p(\tau_k) \prod_{x_s \in X_k} p(t_s, n_s | r_k, \tau_k, \Theta) d\tau_k \quad (72)$$

$$\approx \pi_{r_k} \left(\prod_{x_s \in X_k} a_{n_s r_k} \right) \frac{Z(\sum_{x_s \in X_k} J_{sk}, \sum_{x_s \in X_k} h_{sk})}{\prod_{x_s \in X_k} Z(J_{sk}, h_{sk})}, \quad (73)$$

where, again, the approximation comes from truncating the integral to the range $[0, T]$, and is negligible in our cases. Given the type, the latent event time is conditionally Gaussian,

$$p(\tau_k | r_k, X_k, \Theta) \propto p(\tau_k) \prod_{x_s \in X_k} p(t_s | r_k, n_s, \tau_k, \Theta) \quad (74)$$

$$= \text{Unif}(\tau_k; [0, T]) \prod_{x_s \in X_k} \mathcal{N}(t_s | \tau_k + b_{n_s r_k}, c_{n_s r_k}) \quad (75)$$

$$\propto \mathcal{N}(\tau_k | \mu_k, \sigma_k^2) \quad (76)$$

where

$$\sigma_k^2 = \left(\sum_{x_s \in X_k} J_{sk} \right)^{-1} \quad (77)$$

$$\mu_k^2 = \sigma_k^2 \left(\sum_{x_s \in X_k} h_{sk} \right). \quad (78)$$

Finally, the amplitude is independent of the type and time, and its conditional distribution is given by,

$$p(A_k | X_k, \xi) \propto \text{Ga}(A_k | \alpha, \beta) \text{Po}(S_k | A_k) \quad (79)$$

$$\propto \text{Ga}(A_k | \alpha + S_k, \beta_\emptyset + 1). \quad (80)$$

D.5 Gibbs Sampling the Global Parameters

Finally, the Gibbs updates of the global parameters are simple due to their conjugate priors.

Background rates The conditional distribution of the background rates is,

$$p(\lambda_n^\emptyset | X_0, \xi) \propto \text{Ga}(\lambda_n^\emptyset | \alpha_\emptyset, \beta_\emptyset) \text{Po}(|\{x_s : x_s \in X_0, n_s = n\}| | \lambda_n^\emptyset T) \quad (81)$$

$$\propto \text{Ga}(\lambda_n^\emptyset | \alpha_\emptyset + \sum_{x_s \in X_0} \mathbb{I}[n_s = n], \beta_\emptyset + T). \quad (82)$$

Sequence type probabilities The conditional distribution of the sequence type probability vector π is,

$$p(\pi | \{(r_k, \tau_k, A_k)\}_{k=1}^{K^*}, \xi) \propto \text{Dir}(\pi | \gamma \mathbf{1}_R) \prod_{k=1}^{K^*} \text{Cat}(r_k | \pi) \quad (83)$$

$$\propto \text{Dir}\left(\pi | \left[\gamma + \sum_k \mathbb{I}[r_k = 1], \dots, \gamma + \sum_k \mathbb{I}[r_k = 1] \right]\right). \quad (84)$$

Neuron weights for each sequence type The conditional distribution of the neuron weight vector \mathbf{a}_r is,

$$p(\mathbf{a}_r | X, \{(r_k, \tau_k, A_k)\}_{k=1}^{K^*}, \xi) \propto \text{Dir}(\mathbf{a}_r | \varphi \mathbf{1}_N) \prod_{k=1}^{K^*} \prod_{x_s \in X_k} \text{Cat}(n_s | \mathbf{a}_{r_k}), \quad (85)$$

$$\propto \text{Dir}(\mathbf{a}_r | \phi_r) \quad (86)$$

$$\phi_{rn} = \varphi + \sum_k \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = n]. \quad (87)$$

Neuron widths and delays The conditional distribution of the neuron delays b_{nr} and widths c_{nr} is,

$$p(b_{nr}, c_{nr} \mid X, \{(r_k, \tau_k, A_k)\}_{k=1}^{K^*}, \xi) \propto \text{Inv}-\chi^2(c_{nr} \mid \nu, \sigma^2) \mathcal{N}(b_{nr} \mid 0, c_{nr}/\kappa) \prod_{k=1}^{K^*} \prod_{x_s \in X_k} (\mathcal{N}(t_s \mid \tau_k + b_{nr}, c_{nr}))^{\mathbb{I}[r_k=r \wedge n_s=n]} \quad (88)$$

$$\propto \text{Inv}-\chi^2(c_{nr} \mid \nu_{nr}, \sigma_{nr}^2) \mathcal{N}(b_{nr} \mid \mu_{nr}, c_{nr}/\kappa_{nr}) \quad (89)$$

where

$$S_{nr} = \sum_{k=1}^{K^*} \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = n] \quad (90)$$

$$\nu_{nr} = \nu + S_{nr} \quad (91)$$

$$\sigma_{nr}^2 = \frac{\nu\sigma^2 + \sum_{k=1}^{K^*} \sum_{x_s \in X_k} (t_s - \tau_k)^2 \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\nu + S_{nr}} \quad (92)$$

$$\kappa_{nr} = \kappa + S_{nr} \quad (93)$$

$$\mu_{nr} = \frac{\sum_{k=1}^{K^*} \sum_{x_s \in X_k} (t_s - \tau_k) \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\nu + S_{nr}}. \quad (94)$$

D.6 Split-Merge Sampling Moves

As discussed in the main text, when sequences containing of a small number of spikes are unlikely under the prior, Gibbs sampling can be slow to mix over the posterior of spike partitions since the probability of forming new sequences (i.e. singleton clusters) is very low. In addition to the annealed sampling approach outlined in the main text, we adapted the split-merge sampling method proposed by Jain et al. [7] for Dirichlet process mixture models to PP-Seq.⁴ For simplicity, we implemented randomized split-merge proposals—i.e. without the additional restricted Gibbs sweeps described proposed in Jain et al. [7]. In practice, we found that these random proposals were sufficient for the sampler to accept both split and merge moves at a satisfactorily high rate.

Briefly, our method starts by randomly choosing two spikes (t_i, n_i) and (t_j, n_j) that are not assigned to the background partition and satisfy $|t_i - t_j| < W$ for some user-specified time window W . One could set $W = T$, in which case all pairs of spikes not assigned to the background are considered; however, we will see that this will generally result in proposals that are extremely likely to be rejected, so it is advisable to set W to be an upper bound on the expected sequence length to encourage faster mixing. After identifying the pair of spikes, we propose to either merge their sequences (if $u_i \neq u_j$), or, if they belong to the same sequence ($u_i = u_j = k$) we propose to form two new sequences, each containing one of the two spikes, the remaining $|X_k| - 2$ spikes are randomly assigned with equal probability.

Given the proposed split or merge move, $\mathcal{I} \rightarrow \mathcal{I}^*$, the Metropolis-Hastings acceptance probability is:

$$\min \left[1, \frac{q(\mathcal{I} \mid \mathcal{I}^*) p(\mathcal{I}^* \mid \{x_s\}_{s=1}^S, \Theta, \xi)}{q(\mathcal{I}^* \mid \mathcal{I}) p(\mathcal{I} \mid \{x_s\}_{s=1}^S, \Theta, \xi)} \right] \quad (95)$$

where $q(\mathcal{I}^* \mid \mathcal{I})$ is the proposal density, and $p(\mathcal{I} \mid \{x_s\}_{s=1}^S, \Theta, \xi)$ is given in eqs. (59) and (60). The sampling method then directly follows Jain et al. [7]. The ratio of proposal probabilities for split moves is:

$$\frac{q(\mathcal{I} \mid \mathcal{I}^{\text{split}})}{q(\mathcal{I}^{\text{split}} \mid \mathcal{I})} = \left(\frac{1}{2} \right)^{-(|X_k|-2)} \quad (96)$$

And the ratio for merge moves is:

$$\frac{q(\mathcal{I} \mid \mathcal{I}^{\text{merge}})}{q(\mathcal{I}^{\text{merge}} \mid \mathcal{I})} = \left(\frac{1}{2} \right)^{(|X_k|-2)} \quad (97)$$

E Time-Warped Sequences

To account for sequences that unfold at different speeds, we incorporate a time-warping component into the generative model. Each sequence is endowed with a *warp value*, which is drawn from a discrete distribution,

$$\omega_k \sim \sum_{f=1}^F \eta_f \delta_{w_f}(\omega_k), \quad (98)$$

⁴See also Jain et al. [8] for the case of non-conjugate priors.

where $\eta_f \geq 0$ and $\sum_{f=1}^F \eta_f = 1$ are the probabilities of the corresponding warp values $w_f > 0$. In this paper, we set

$$w_f = w_F^{-1+2(f-1)/(F-1)} \quad (99)$$

$$\eta_f \propto \mathcal{N}(f | \frac{F-1}{2}, \sigma_w^2). \quad (100)$$

The hyperparameters include the number of warp values F , the maximum warp value w_F , and the variance of the warp probabilities σ_w^2 . We use an odd number of warp values so that $w_{(F-1)/2} = 1$.

The warp value changes the distribution of spike times in the corresponding sequence by scaling the mean and standard deviation,

$$t_s^{(k)} \sim \mathcal{N}(\tau_k + \omega_k b_{n_s^{(k)} r_k}, \omega_k^2 c_{n_s^{(k)} r_k}). \quad (101)$$

Note that one could also choose to linearly scale the variance rather than the standard deviation, as in a warped random walk. Ultimately, this is a subjective modeling choice, and scaling the standard deviation leads to slightly easier Gibbs updates later on.

We can equivalently endow each sequence with a latent *warp index* $f_k \in \{1, \dots, F\}$ where $p(f_k) \propto \eta_{f_k}$ and,

$$t_s^{(k)} \sim \mathcal{N}\left(\tau_k + w_{f_k} b_{n_s^{(k)} r_k}, w_{f_k}^2 c_{n_s^{(k)} r_k}\right). \quad (102)$$

Since there are only a finite number of warp values, we can treat the warp index and the sequence type as a single discrete latent variable $(r_k, f_k) \in \{1, \dots, R\} \times \{1, \dots, F\}$. This formulation allows us to re-use all of the collapsed Gibbs updates derived above, replacing our inferences over r_k with inferences of (r_k, f_k) . Computationally, this increases the run-time by a factor of F .

Given the sequence types, times, and warp values, the conditional distribution of the neuron delays b_{nr} and widths c_{nr} is,

$$p(b_{nr}, c_{nr} | X, \{(r_k, \tau_k, f_k, A_k)\}_{k=1}^{K^*}, \xi) \quad (103)$$

$$\propto \text{Inv-}\chi^2(c_{nr} | \nu, \sigma^2) \mathcal{N}(b_{nr} | 0, c_{nr}/\kappa) \prod_{k=1}^{K^*} \prod_{x_s \in X_k} (\mathcal{N}(t_s | \tau_k + w_{f_k} b_{nr}, w_{f_k}^2 c_{nr}))^{\mathbb{I}[r_k=r \wedge n_s=n]} \quad (104)$$

$$\propto \text{Inv-}\chi^2(c_{nr} | \nu_{nr}, \sigma_{nr}^2) \mathcal{N}(b_{nr} | \mu_{nr}, c_{nr}/\kappa_{nr}) \quad (105)$$

where

$$S_{nr} = \sum_{k=1}^{K^*} \sum_{x_s \in X_k} \mathbb{I}[r_k = r \wedge n_s = n] \quad (106)$$

$$\Delta_{sk} = \frac{t_s - \tau_k}{w_{f_k}} \quad (107)$$

$$\nu_{nr} = \nu + S_{nr} \quad (108)$$

$$\sigma_{nr}^2 = \frac{\nu \sigma^2 + \sum_{k=1}^{K^*} \sum_{x_s \in X_k} \Delta_{sk}^2 \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\nu + S_{nr}} \quad (109)$$

$$\kappa_{nr} = \kappa + S_{nr} \quad (110)$$

$$\mu_{nr} = \frac{\sum_{k=1}^{K^*} \sum_{x_s \in X_k} \Delta_{sk} \cdot \mathbb{I}[r_k = r \wedge n_s = n]}{\kappa + S_{nr}}. \quad (111)$$

If instead you choose to linearly scale the variances, as mentioned above, the conditional distribution of the delays and widths is no longer inverse χ^2 . However, it is straightforward to sample the delays and widths one at a time from their univariate conditional distributions.

F Supplemental Details on Experiments

E.1 ROC curve comparison of convNMF and PP-Seq

Simulated datasets consisted of $N = 100$ neurons, $T = 2000$ time units, $R = 1$ sequence type, sequence event rate $\psi = 0.02$, a default average background rate of $\lambda_n^\emptyset = 0.03$, default settings of $\alpha = 225$ and $\beta = 7.5$. Further, we set

$\kappa_{nr} = c_{nr}$ and $c_{nr} = 0.04$ so that neuron offsets had unit standard deviation and small jitter by default. For the case of “jitter noise” we modified c_{nr} but continued to set $\kappa_{nr} = c_{nr}$ to preserve the expected distribution of neuron offsets.

For each synthetic dataset, we fit convNMF models with $R = 1$ component and a time bin size of 0.2 time units. We optimized using alternating projected gradient descent with a backtracking line search. Each convNMF model produced a temporal factor $\mathbf{h} \in \mathbb{R}_+^B$ (where $B = T/0.2 = 40000$ is the total number of time bins). We encoded the ground truth sequence times in a binary vector of length B and computed ROC curves by thresholding \mathbf{h} over a fine grid of values over the range $[0, \max(\mathbf{h})]$. We consider each timebin, indexed by b , a false positive if h_b exceeds the threshold, but no ground truth sequence was present in the timebin. Likewise, a timebin is a true positive when h_b exceeds the threshold and a sequence event did occur in this timebin.

We repeated a similar analysis with 100 (approximate) samples from the PP-Seq posterior distribution. Specifically, we discretized the sampled sequence event times, i.e. $\{\tau_k\}_{k=1}^K$ for every sample, into the same B time bins used by convNMF. For every timebin we computed the empirical probability that it contained a sampled event from the model, averaging over MCMC samples. This resulted in a discretized, nonnegative temporal factor that is directly analogous to the factor \mathbf{h} produced convNMF. We compute an ROC curve by the same method outlined above.

Unfortunately, both convNMF and PP-Seq parameters are only weakly identifiable. For example, in PP-Seq, one can add a constant to all latent event times, $\tau_k \mapsto \tau_k + C$, and subtract the same constant from all neuron offsets $b_{nr} \mapsto b_{nr} - C$. This manipulation does not modify the likelihood of the data (it does, however, affect the probability of the parameters under the prior, and so we say the model is “weakly identifiable”). The result of this is that both PP-Seq and convNMF may consistently misestimate the “true” sequence times by a small constant of time bins. To discount this nuisance source of model error we repeated the above analysis on shifted copies of the temporal factor (up to 20 time bins in each direction); we then selected the optimal ROC curve for each model and computed the area under the curve to quantify accuracy.

F.2 Hyperparameter sweep on hippocampal data

We performed MCMC inference on 2000 randomly sampled hyperparameter sets, masking out a random 7.5% of the data as a validation set on every run. We fixed the following hyperparameter values: $\gamma = 3$, $\varphi = 1$, $\mu_{nr} = 0$ for every neuron and sequence type, $\nu_{nr} = 4$ for every neuron and sequence type. We also fixed the following hyperparameters, pertaining to time warping: $F = 10$, $\sigma_w^2 = 100$. We randomized the number of sequence types $R \in \{1, 2, 3, 4\}$ uniformly. We randomized the maximum warp value w_F uniformly on $(1, 1.5)$. We randomized the mean sequence amplitude, α/β , log-uniformly over the interval $(10^2, 10^4)$, and set the variance of the sequence amplitude α/β^2 equal to the mean. We randomized the expected total background amplitude, λ^\varnothing , log-uniformly over the interval $(10^2, 10^4)$ and set the variance of this parameter equal to its mean. We randomized the sequence rate ψ log-uniformly over $(10^{-3}, 10^{-1})$. Finally, we randomized the neuron response widths c_{nr} log-uniformly over $(10^{-1}, 10^1)$.

We randomly initialized all spike assignments either with convNMF or the annealed MCMC procedure outlined in the main text (we observed successful outcomes in both cases). In annealed runs, we set an initial temperature of 500 and exponentially relaxed the temperature to one over 20 stages, each containing 100 Gibbs sweeps. (Recall that this temperature parameter scales the variance of $\text{Ga}(\alpha, \beta)$ while preserving the mean of this prior distribution.) After annealing was completed we performed 100 final Gibbs sweeps over the data, the final half of which were used as approximate samples from the posterior. In this final stage of sampling, we also performed 1000 randomized split-merge Metropolis Hastings moves after every Gibbs sweep.

All MCMC runs were performed using our Julia implementation of PP-Seq, run in parallel on the Sherlock cluster at Stanford University. Each job was allocated 1 CPU and 8 GB of RAM.

References

- [1] John Frank Charles Kingman. *Poisson processes*. Clarendon Press, 2002.
- [2] D J Daley and D Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, New York, NY, 2003.
- [3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. 3rd ed. CRC press, 2013.
- [4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [5] Radford M Neal. “Markov chain sampling methods for Dirichlet process mixture models”. *J. Comput. Graph. Stat.* 9.2 (2000), pp. 249–265.
- [6] Jeffrey W Miller and Matthew T Harrison. “Mixture models with a prior on the number of components”. *J. Am. Stat. Assoc.* 113.521 (2018), pp. 340–356.
- [7] Sonia Jain and Radford M Neal. “A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model”. *J. Comput. Graph. Stat.* 13.1 (2004), pp. 158–182.
- [8] Sonia Jain and Radford M Neal. “Splitting and merging components of a nonconjugate Dirichlet process mixture model”. *Bayesian Anal.* 2.3 (2007), pp. 445–472.