Analyzing second order stochasticity of neural spiking under stimuli-bundle exposure

Chris Glynn
University of New Hampshire

Surya T Tokdar
Duke University

Azeem Zaman
Harvard University

Valeria C Caruso
University of Michigan

Jeffrey T Mohl
Duke University

Shawn M Willett
Duke University

Jennifer M Groh
Duke University

Abstract

Conventional analysis of neuroscience data involves computing average neural activity over a group of trials and/or a period of time. This approach may be particularly problematic when assessing the response patterns of neurons to more than one simultaneously presented stimulus. In such cases, the brain must represent each individual component of the stimuli bundle, but trial-and-time-pooled averaging methods are fundamentally unequipped to address the means by which multi-item representation occurs. We introduce and investigate a novel statistical analysis framework that relates the firing pattern of a single cell, exposed to a stimuli bundle, to the ensemble of its firing patterns under each constituent stimulus. Existing statistical tools focus on what may be called "first order stochasticity" in trial-to-trial variation in the form of unstructured noise around a fixed firing rate curve associated with a given stimulus. Our analysis is based upon the theoretical premise that exposure to a stimuli bundle induces additional stochasticity in the cell's response pattern, in the form of a stochastically varying recombination of its single stimulus firing rate curves. We discuss challenges to statistical estimation of such "second order stochasticity" and address them with a novel dynamic admixture Poisson process (DAPP) model. DAPP is a hierarchical point process model that decomposes second order stochasticity into a Gaussian stochastic process and a random vector of interpretable features, and, facilitates borrowing of information on the latter across repeated trials through latent clustering. We present empirical evidence of the utility of the DAPP analysis with synthetic and real neural recordings.

1 Introduction

The brain is capable of encoding multiple objects presented simultaneously. But the neural computing behind this complex operation – of great relevance to computational and cognitive neuroscience – remains poorly understood. Presently lacking are statistical models and tools to quantify the relationship between an individual cell's response to a bundle of stimuli presented together, and the ensemble of its response patterns evoked when each stimulus is presented in isolation. We fill this gap with a novel statistical analysis framework, developed under the theory that a cell's response to a stimuli bundle is a stochastically varying, dynamic combination of its single stimulus response patterns. Such a theory allows the possibility that each item in the stimuli bundle dominates the cell's response pattern during distinct periods of time. We have recently presented evidence in favor of such an interpretation for auditory and visual stimuli (Caruso et al., 2018).

For simplicity, and also limited by available experimental data, we restrict this discussion to stimuli bundles consisting of two stimuli, each of which evokes detectable and separable response patterns from a neural cell. Neural activity in each experimental trial is measured as a spike train recorded over a common time horizon. We assume repeated trials are available from each of the following three experimental conditions, A: "exposure to a stimulus A alone", B: "exposure to a stimulus B alone", and, AB: "exposure to stimuli A and B together".

Statistical analysis of spike train data typically assumes an underlying, stimulus-driven response curve from which a stochastic point pattern of spiking times is generated on each experimental trial (Gerstein and Kiang, 1960); see Kass et al. (2005) and the references therein for a comprehensive overview. The response curve, taken as a function of time, is interpreted to give the potentially time-varying expected firing rates of the cell in response to the given stimulus. Variations of the spike train across multiple trials is considered "random noise" around this expected rate curve, realized in the form of a random point pattern. We refer to such variation as *first order stochasticity.* Statistical analyses under this framework usually proceed by aggregating spike trains across trials to improve accuracy in estimating the underlying response curve. We adopt this framework to estimate the expected firing rate curves $\lambda_{\mathrm{A}}(t)$ and $\lambda_{\mathrm{B}}(t)$ associated with, respectively, stimulus A and and stimulus B.

The same framework, however, may not apply to the case when both stimuli A and B are presented together, and the brain perceives them as distinct signals (perhaps revealed by behavioral response). To the brain, the stimuli are not fused together as a novel combined stimulus, but remain a stimuli bundle with each signal maintaining its individuality. It is conceivable that exposure to a stimuli bundle may induce a second type of stochasticity in the cell's response. Each trial under condition AB may involve its own distinct response curve that combines both λ_{A} and λ_{B}, with the combination depending on unmeasured upstream or contemporaneous representation of the stimuli bundle by other cells.

We refer to such random but structural variation across trials as *second order stochasticity.* We distinguish second order stochasticity from a broader umbrella term *trial-to-trial variation* often used in the literature (Kass et al., 2005; Ventura et al., 2005). Our focus is on quantifying variability that is information-encoding and intrinsic to the cell in a given experimental condition, as opposed to noisy fluctuations that may be caused by factors extrinsic to the primary stimuli. Quantifying the precise nature of this variability is the central focus of our analysis.

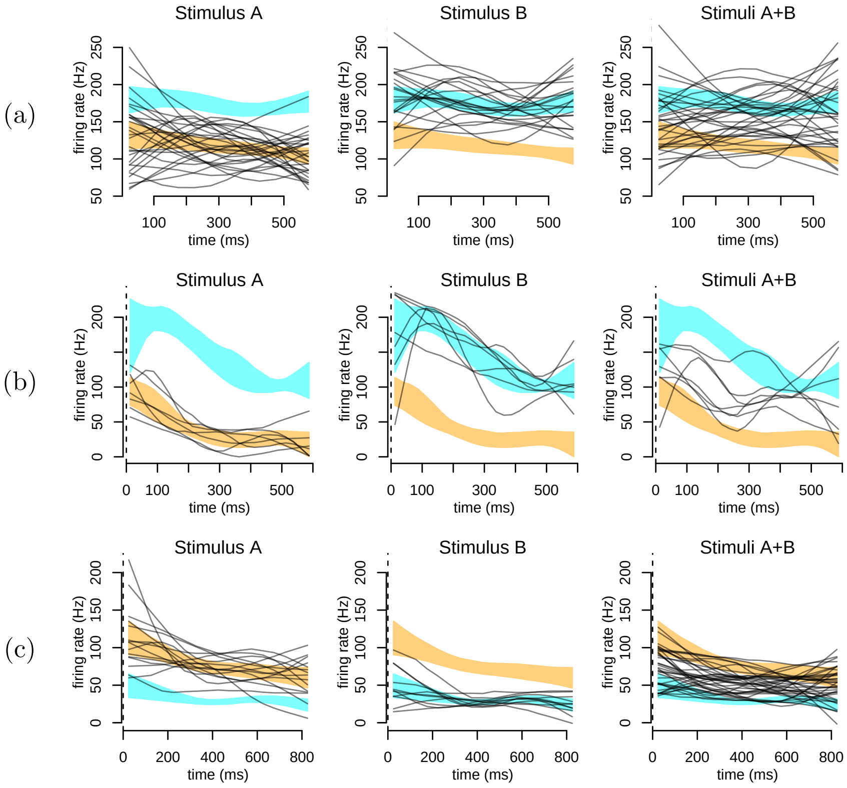


Figure 1: Multiple forms of stochasticity in inferior colliculus (IC). Each row corresponds to a distinct cell recorded from monkey IC and shows how the cell responds to the triple of experimental conditions A, B and AB, where A and B each corresponds to an auditory stimulus in the form of a bandpass filtered noise played from a certain angle. Each black curve represents one trial and shows the trial's spiking rate, which has been smoothed to aid visualization. The orange and cyan bands show estimates and uncertainty bands for λ_A and λ_B. (a) Cell 1: AB responses appear to be a superimposition of A and B responses. (b) Cell 2: AB responses appear to fluctuate more widely within each trial than A or B responses. (c) Cell 3: similar to Cell 1 in appearance, but here the AB responses appear more squeezed toward the middle than how a superimposition of A and B responses would appear.

2 Statistical analysis of second order stochasticity

2.1 The dynamic averaging model

Our general approach is to describe second order stochasticity as dynamic averaging, in which the relative contributions of A-like and B-like response patterns can vary across time on multiple scales. Specifically, we describe the rate curve behind any specific AB trial, as a convex combination $\alpha(t)\lambda_{\mathrm{A}}(t)+(1-\alpha(t))\lambda_{\mathrm{B}}(t)$, involving a possibly time varying weight curve $\alpha(t)$. Second order stochasticity manifests when the entire weight curve varies stochastically across AB trials, either stably within a trial but variably across trials or variably across both trials, and time within trials.

A weight curve $\alpha(t)$ that is stable across time within a trial, but clusters bimodally near

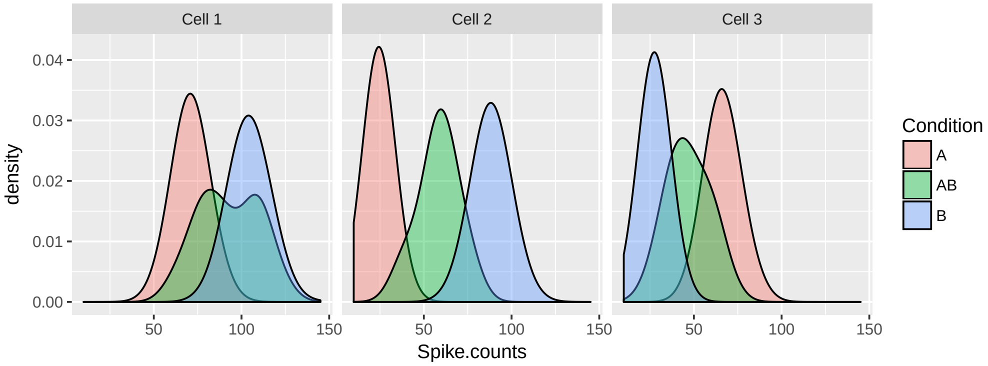


Figure 2: Smoothed histograms of whole-trial spike counts of the three IC cells, grouped by experimental condition A, B and AB. For each cell, the AB total spike count distribution sits between the distributions under conditions A and B. But the shape of the AB distribution varies across cells.

zero and one across trials, constitutes a special case, consistent with neurons encoding only one of the two stimuli per trial, and doing so in a fashion that is consistent with how they respond when only that stimulus is present. In our previous study, we referred to such cases as showing whole-trial fluctuations ("Mixtures"; Caruso et al., 2018). If the underlying firing rate dynamically alternates between those encoded by $\lambda_\mathrm{A}(t)$ and $\lambda_\mathrm{B}(t)$ within the course of a single trial, with $\alpha(t)$ approaching values of 0 and 1 for periods of time, the neuron may be encoding each stimulus separately during distinct temporal epochs of sub-trial durations.

In either of these two special cases, the neuron is imagined to encode for only a single stimulus at any given time point. Our dynamic averaging model goes beyond such *one signal at a time* view, and allows for cases where the neuron's firing rate at any time point on a AB trial is truly intermediary between its A-level and B-level firing rates at the same time point. Here, the weight curve $\alpha(t)$ is seen as undulating, within or across trials, between a range of values that are bounded away from the extremes of zero or one.

In Figures 1 and 2, we visualize response patterns of three example inferior colliculus cells from our experiments, where, the first two cells appear to exhibit, respectively, random selection of a particular stimulus at the whole trial level, and, interleaving of signals related to each stimulus across time within a trial. In contrast, for Cell 3 in Figures 1 and 2, the response pattern appears to match truly intermediary averaging that varies stochastically possibly across both time and trials.

2.2 Statistical estimation challenges

It is challenging to carry out statistical analysis of second order stochasticity under the dynamic averaging assumption. First, purely from a statistical accuracy perspective, estimation of the weight curves is difficult because one has access to only one spike train for each unknown weight curve. An ordinary aggregation across the AB trials no longer helps in combining information. Instead, one must rely on a hierarchical model that relates the weight curves to each other through a few meaningful features which are then estimated jointly from the pooled data.

Second, on a more conceptual level, simply estimating the weight curves is not enough to draw inference on the exact nature of the cell's second order stochasticity. What is more relevant is to be able to predict how the cell is going to respond if new trials were carried out under the AB condition. While the weight curves associated with the new trials cannot be predicted exactly, one should be able to predict what features these weight curves are likely to possess.

We address these challenges with a novel hierarchical point process model. We formulate the distribution of the weight curves as an unknown quantity to be estimated from the data. We reduce the estimation complexity of this problem by assuming the unknown stochasticity of the weight curves can be decomposed into a known Gaussian process distribution on smooth curves and an unknown probability distribution on a vector of a small number of meaningful summaries of the weight curves. The latter unknown probability distribution is conceived to be a discrete distribution, and, is assigned a Dirichlet (Ferguson, 1973) process prior to carry out a full Bayesian estimation. The discreteness assumption induces a (stochastic) clustering of the AB trials, facilitating borrowing of information across weight curves. Estimating the unknown distribution of the weight curves leads immediately to realistic prediction of the features of the weight curve in future trials.

3 Dynamic admixture point process model

3.1 Poisson process formulation

Let n_{A}, n_{B} and n_{AB} give the numbers of trials under, respectively, conditions A, B and AB. Each trial produces a distinct spike train measurement. We assume that any neural spike train recorded over a given response window $[0, T]$ is a realization of a stochastic point process $(N(t) : t \in [0, T])$ where $N(t)$ denotes the spike count between time zero and t, $0 \leq t \leq T$. For each condition $e \in \{\mathrm{A}, \mathrm{B}, \mathrm{AB}\}$ and each trial $j \in \{1, \ldots, n_e\}$, let $N_j^e(t)$ denote the corresponding point process.

For conceptual simplicity and analytical tractability we make a Poisson distributional assumption on these three sets of point processes.

1. N_j^{A}, $j = 1, \ldots, n_{\mathrm{A}}$, are independent realizations of an inhomogeneous Poisson process with intensity function $\lambda_{\mathrm{A}}(t)$, $t \in [0, T]$;

2. N_j^{B}, $j = 1, \ldots, n_{\mathrm{B}}$, are independent realizations of an inhomogeneous Poisson process with intensity function $\lambda_{\mathrm{B}}(t)$, $t \in [0, T]$;

3. N_j^{AB}, $j = 1, \ldots, n_{\mathrm{AB}}$ are independently distributed inhomogeneous Poisson processes but with distinct intensity functions. The intensity function of the j-th such process is given by
$$\lambda_j(t) = \alpha_j(t)\lambda_{\mathrm{A}}(t) + \{1 - \alpha_j(t)\}\lambda_{\mathrm{B}}(t),\ t \in [0, T]$$
where $\alpha_j(t) \in [0, 1]$, $t \in [0, T]$ is a possibly time varying weight curve.

To incorporate second order stochasticity in our model, we assume the weight curves $\alpha_j(t)$, $j = 1, \ldots, n_{\mathrm{AB}}$, are independently distributed according to some unknown probability law $\mathbb{P}$

on the space of weight curves. This probability law may be understood as a characteristic of the neural cell when subjected to condition AB. Estimation of $\mathbb{P}$ is the key goal of our statistical analysis. We call this model the *dynamic admixture of Poisson process* (DAPP) model.

3.2 Modeling the stochasticity of weight curves

The space of weight curves is large and complex, and statistical estimation of an unknown probability law on this space is next to impossible without strong structural assumptions. Below we introduce a model for $\mathbb{P}$ where the unknown stochasticity of the weight curve is reduced to an unknown stochasticity of only a limited number of its features, namely, the curve's long term average value, maximum deviation from the average, and, the extent of waviness around the average. The remaining stochasticity is assumed to be governed by a known probability law, namely a modified Gaussian measure.

3.2.1 A Gaussian probability law for curves on $[0, T]$

To be specific, for any $\ell > 0$, let $C_\ell^{\mathrm{SE}} : [0, T] \times [0, T] \to (0, \infty)$ denote the so called squared exponential kernel with characteristic length-scale ℓ, given by

$$C_\ell^{\mathrm{SE}}(s, t) = \sigma_0^2 \exp\left\{-\frac{(s-t)^2}{2\ell^2}\right\}, \quad s, t \in [0, T],$$

where σ_0^2 is a fixed scalar to be discussed later. For any scalars ϕ and $\psi > 0$, let $\mathrm{GP}(\phi, \psi C_\ell^{\mathrm{SE}})$ denote the probability law of a Gaussian process $(\eta(t) : t \in [0, T])$ with mean and covariance functions

$$\mathbb{E}[\eta(t)] \equiv \phi, \quad \mathbb{C}\mathrm{ov}[\eta(s), \eta(t)] = \psi C_\ell^{\mathrm{SE}}(s, t), \quad t, s \in [0, T]. \tag{1}$$

It is well known that $\mathrm{GP}(\phi, \psi C_\ell^{\mathrm{SE}})$ defines a Gaussian measure on the space of smooth curves on $[0, T]$.

Remark 1. Random curves generated from this measure are not exactly periodic, but are systematically wavy in the sense that the number of times such a curve crosses any fixed level is a random variable with a finite expectation. Indeed, the expected number of up-crossings[1] of level ϕ is precisely $T/(2\pi\ell) \approx 0.16 \cdot T/\ell$. Therefore, a $\mathrm{GP}(\phi, \psi C_\ell^{\mathrm{SE}})$ law favors flat or wavy curves depending upon whether ℓ is, respectively, large or small. With $\ell = 160\% T$, one expects little waviness since the expected number of up-crossing is only a tenth, whereas, with $\ell = 4\% T$ one expects four up-crossings and hence considerable waviness.

Remark 2. Furthermore, for any random curve η generated from $\mathrm{GP}(\phi, \psi C_\ell^{\mathrm{SE}})$, the scalar ϕ gives the expected value of the curve at any time point t as well as the expected value of its long term average $\bar{\eta} := (1/T)\int_0^T \eta(t)dt$. If η' were another curve generated from the same law and independently of η, then $\mathbb{E}\{\eta(t) - \eta'(t)\}^2 = 2\psi\sigma_0^2$ at every $t \in [0, T]$, and, hence ψ represents the range of the curve across repeated random generations. Both ψ and ℓ play

[1] crossing from below; see Adler and Taylor (2009), Chapter 11.

a role in controlling the within-trial deviation of $\eta(t)$ around its long term average $\bar{\eta}$. This deviation can be quantified as

$$\mathbb{E}\left[\frac{1}{T}\int_0^T (\eta(t)-\bar{\eta})^2 dt\right] = \psi\sigma_0^2\left\{1 - T^{-2}\iint_{[0,T]^2} e^{-\frac{(s-t)^2}{2\ell^2}} ds dt\right\}. \tag{2}$$

The right hand side of (2) is a monotonically decreasing function[2] of ℓ/T, going from a maximum value of $\psi\sigma_0^2$ at $\ell = 0$ to 0 as $\ell/T \to \infty$. For $\ell = 4\%T$, the right hand side of (2) equals $0.9 \cdot \psi\sigma_0^2$, which means, the within-trial deviation is expected to be 90% of what across-trial variance of the curve at any single time point. On the other hand, for $\ell = 160\%T$, the within-trial deviation equals $0.03 \cdot \psi\sigma_0^2$, i.e, only 3% of the across-trial variance.

3.2.2 A hierarchical Gaussian measure model for $\mathbb{P}$

We model $\mathbb{P}$ as the probability law of a random weight curve $\alpha(t)$ generated by the following sequence of operations:

$$\text{draw} \;\; (\phi, \psi, \ell) \sim \mathbb{Q}, \tag{3}$$

$$\text{draw} \;\; \eta \sim \text{GP}(\phi, \psi C_\ell^{\text{SE}}), \tag{4}$$

$$\text{set} \;\; \alpha(t) = \frac{1}{1+e^{-\eta(t)}}, \quad t \in [0,T], \tag{5}$$

where $\mathbb{Q}$ is an unknown probability law on $(-\infty,\infty)\times(0,\infty)\times(0,\infty)$, to be estimated from data. Even without (3), one could simply take (4)-(5) as a model for $\mathbb{P}$ where the only unknown quantities are the three scalars ϕ, ψ and ℓ, which would render parameter estimation far easier. Therefore it is important to justify why we must include (3) in our model for $\mathbb{P}$.

Consider the case where a cell's second order stochasticity is close to 50-50 random selection; in nearly half of the AB trials $\alpha(t) \approx 0.9, t \in [0,T]$, while in the other half, $\alpha(t) \approx 0.1, t \in [0,T]$. Suppose our model for $\mathbb{P}$ were based of only (4)-(5) with the vector (ϕ, ψ, ℓ) being the only unknown quantity. In light of the remarks in Section 3.2.1, one would estimate $\phi \approx 0$ and both ψ and ℓ large. Hence, the estimated $\mathbb{P}$ will produce $\alpha(t)$ curves that are nearly flat across time, but, with no discernible concentration around either the 0.1 mark or the 0.9 mark. Therefore, while the estimated model will provide great fit to the observed data, it will completely fail to learn the true nature of the second order stochasticity.

Inclusion of (3) in modeling $\mathbb{P}$ offers a much richer framework to learn various kinds of second order stochasticity. The vector (ϕ, ψ, ℓ) in (4) exerts direct control on several broad features of the random weight curve α in (5), e.g., its waviness, range, long term average and deviation around the long term average. The unknown probability measure $\mathbb{Q}$ of (ϕ, ψ, ℓ) represents the unknown nature of stochasticity of these broad features.

4 Bayesian inference: prior specification

Although (3)-(5) offers a great reduction of complexity in statistical estimation of $\mathbb{P}$, estimating the remaining unknown quantity, the probability measure $\mathbb{Q}$, still remains a challenging

[2] given by $\psi\sigma_0^2\{1 - f(\ell/T)\}$ where $f(r) = 2[\sqrt{2\pi}r\{\Phi(r^{-1}) - 0.5\} - r^2\{1 - \exp(-0.5r^{-2})\}]$, $r \geq 0$.

problem. We adopt a Bayesian inference technique to estimate $\mathbb{Q}$ from data where a well chosen prior distribution on $\mathbb{Q}$ offers further structural simplification and regularization through latent clustering.

Remark 3 (Notation). Below we use the generic expression $p(x|y)$ to understand the conditional distribution and/or the conditional probability density function (pdf) of one variable x given another variable y. We use $Poi(\mu)$ to denote the Poisson distribution with mean μ; $Bin(n, p)$ to denote the binomial distribution with size n and success probability p; $N(m, v)$ to denote the (possibly multivariate) normal distribution with mean (vector) m and variance (matrix) v; $Be(a, b)$ to denote the beta distribution with shape parameters a and b; $Dir(a_1, \ldots, a_k)$ to denote the k-dimensional Dirichlet distributions with shape parameters a_i, $i = 1, \ldots, k$; $Ga(r, s)$ to denote the gamma distribution with shape r and rate s (so that mean is r/s); $IG(r, s)$ to denote the inverse-gamma distribution with shape r and rate s; δ_x to denote the Dirac probability measure that assigns probability one to a single atom x; and, for an ordered, finite set $A = \{a_1, \ldots, a_k\}$ of size k and a probability vector $\mathbf{p} = (p_1, \ldots, p_k)$, $P_{A,\mathbf{p}}$ to denote the discrete distribution $\sum_{i=1}^{k} p_i \delta_{a_i}$ supported on A that assigns probability p_i to atom a_i, $i = 1, \ldots, k$.

4.1 A structural simplification of characteristic length-scale

We restrict the characteristic length-scale ℓ to realize values in a known finite set $\mathcal{L} = \{\ell_1^*, \ldots, \ell_L^*\} \subset (0, \infty)$. Such a choice offers great computational speed and can be justified by Remarks 1 and 2. In particular, Remark 1 implies that ℓ is intimately related to the number of stochastic oscillations of α, with the expected number of up-crossing of its long-term average being $\approx 0.16T/\ell$. Since this number can be limited to a finite range that is scientifically relevant, one could find a suitable finite set $\mathcal{L}$ that offers a good coverage of plausible oscillatory behavior of the weight curves. For example, to represent between zero and four oscillations, one could work with $\mathcal{L} = \{0.16T/N : N \in \{0.1, 0.5, 1, 2, 3, 4\}\}$. In our experiments we typically have a response horizon of $T = 1000$ (measured in milliseconds), for which the corresponding grid, reordered from the smallest to the largest, is $\mathcal{L} = \{40, 53.3, 80, 160, 320, 1600\}$.

We model $\mathbb{Q}$ hierarchically as the distribution of (ϕ, ψ, ℓ) from the specification

$$(\phi, \psi, \boldsymbol{\pi}) \sim \mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}, \quad \ell \sim P_{\mathcal{L},\boldsymbol{\pi}} \tag{6}$$

where $\boldsymbol{\pi}$ is a random element of the L-dimensional probability simplex $\Delta_L = \{(\pi_1, \ldots, \pi_L) \in \mathbb{R}^L : \pi_i \geq 0, \sum_i \pi_i = 1\}$, and, $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$ is an unknown probability measure on $(-\infty, \infty) \times (0, \infty) \times \Delta_L$. A prior distribution on $\mathbb{Q}$ is specified by assigning a prior distribution to $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$.

4.2 Dirichlet process prior

We assign $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$ a Dirichlet process prior $\text{DP}(\kappa G_\kappa)$ with precision $\kappa > 0$ and base probability measure G_κ on $(-\infty, \infty) \times (0, \infty) \times \Delta_L$ that depends on the precision, to be specified below. This prior specification restricts $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$ to be a (random) discrete probability measure with infinitely many atoms

$$\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}} = \sum_{h=1}^{\infty} \omega_h \delta_{(\phi_h^*,\psi_h^*,\boldsymbol{\pi}_h^*)} \tag{7}$$

where the atoms $(\phi_h^*, \psi_h^*, \boldsymbol{\pi}_h^*)$, $h = 1, 2, \ldots$, are drawn independently from the base measure G_κ and the weights ω_h, $h = 1, 2, \ldots$, admit the stick-breaking representation $\omega_h = \beta_h \prod_{j=1}^{h-1}(1 - \beta_j)$ with β_h, $h = 1, 2, \ldots$, drawn independently from a $Be(1, \kappa)$ distribution (Sethuraman, 1994).

The discreteness of $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$ implies that repeated independent draws of $(\phi, \psi, \boldsymbol{\pi})$ from this probability law will produce duplication. Consequently, the AB trials can be grouped into clusters where within each cluster all trials have weight curves arising as in (4)-(5) with a single underlying (ϕ, ψ) and distinct realizations of ℓ arising from a shared probability vector $\boldsymbol{\pi}$. Therefore these weight curves would have broad features such as the long term average and the range roughly matched. If $\boldsymbol{\pi}$ was peaked in one coordinate, i.e, $\pi_i \approx 1$ for some i while other $\pi_{i'}$ are close to zero, then the weight curves from the cluster would also have similar oscillatory behavior. However, in spite of sharing these broad features, the exact forms of these curves will be different.

The precision parameter κ determines the extent of this clustering with larger values of κ leading to more distinct clusters. Following common practice (Escobar and West, 1995) we assign the precision parameter a further $Ga(1, 1)$ prior which makes the learning of this parameter relatively straightforward.

Remark 4. One could specify a Dirichlet process prior directly on $\mathbb{Q}$, without the introduction of the intermediary quantity $\boldsymbol{\pi}$ as in (6). But our choice of decoupling ℓ from (ϕ, ψ) via the introduction of $\boldsymbol{\pi}$ leads to much improved posterior computation. See Appendix B for more details.

4.3 An unconventional choice of the base distribution

We deviate from common practice in choosing the base measure G_κ which we take to depend on the precision κ and equal the law of $(\phi^*, \psi^*, \boldsymbol{\pi}^*)$ where

$$\boldsymbol{\pi}^* \sim Dir(a_1, \ldots, a_L), \quad \psi^* | \kappa \sim Be(1, \kappa), \quad \phi^* | (\psi^*, \kappa) \sim N(0, \sigma_0^2(1 - \psi^*)). \tag{8}$$

This choice of G_κ ensures that under (3)-(4), $\eta(t) \sim N(0, \sigma_0^2)$ at each $t \in [0, T]$. Consequently, with $\sigma_0 = 1.87$, our *a priori* belief is that $\alpha(t)$ is nearly uniformly distributed over the range $(0, 1)$ at each single time point t. In contrast, the more conventional choice of a normal-inverse gamma base measure (Escobar and West, 1995) would lead to a heavy tailed Student-t prior on $\eta(t)$, and consequently, the prior on $\alpha(t)$ would place more mass than a uniform prior near the extremes of $\alpha(t) = 0$ and $\alpha(t) = 1$.

The particular form of dependence of G_κ on κ in (8) offers additional structural control on the clusters formed by repeated draws of η from (3)-(4). Let η_j, $j = 1, \ldots, n$, denote such repeated draws and focus on the behavior of $Y_j = \eta_j(t)$, $j = 1, \ldots, n$, at any arbitrary $t \in [0, T]$. We know that the marginal distribution of Y_j's is $N(0, \sigma_0^2)$. This marginal distribution decomposes into a weighted average of the cluster specific distributions of Y_j's with the weights being proportional to cluster sizes. When κ is small, there is likely to be one dominating cluster of η_j's sharing a common atom (ϕ^*, ψ^*). It is desirable that the Y_j's in this dominating cluster should have a marginal distribution close to $N(0, \sigma_0^2)$. This is indeed the case under (8), which, for a small value of κ, makes ψ^* likely to be close to 1, and, hence ϕ^* likely to be close to 0.

On the other hand, for a large value of κ, there will be many clusters and it is desirable that the cluster specific distributions of Y_j's should be distinct from each other. In this case for any cluster ψ^* is likely to be small and ϕ^* is possibly away from zero, and, thus the corresponding marginal distribution of Y_j's would be a normal distribution with a small variance and a center that is drawn randomly from a normal distribution with variance close to σ_0^2.

Although our nonconventional choice of G_κ introduces some new computing challenges, they can be met with a fairly straightforward application of the widely used Algorithm 8 of Neal (2000). The additional dependence of G_κ on κ in (8) poses no serious obstacle to the learning of this precision parameter.

The hyperparameters $a_1, \ldots, a_L \in (0, \infty)$, of the Dirichlet distribution in (8) determine the prior expectation for $\boldsymbol{\pi}^*$ in the form of the probability vector $(a_1, \ldots, a_L)/\sum_i a_i$, with $\sum_i a_i$, called precision, serving as a measure of tightness of the prior around the prior expectation. For the default choice of $\mathcal{L}$ as given before, and arranged from the smallest to the largest, we choose $a_i \propto i$ and adjust them so that $\sum_i a_i = 2$. With this choice, larger length-scales and hence flatter weight curves are slightly favored *a priori*. The precision value 2 ensures the prior belief to be at par with the information content of two observations drawn from the multinomial distribution $P_{\mathcal{L},\boldsymbol{\pi}}$.

5 Posterior computing

5.1 Time discretized model approximation

For any step function $f(t)$ on $[0, T]$ that is continuous from the right, let $J(f) = \{t \in [0, T] : f(t) \neq f(t-)\}$ denote the set of its jump points. If $(N(t) : t \in [0, T])$ is an inhomogeneous Poisson process with intensity $\lambda(t)$, then with probability one, N is a step function that is continuous from the right and $J(N)$ is finite. In fact, the likelihood of observing N can be expressed as

$$p(N|\lambda) = e^{-\int_0^T \lambda(t)dt} \prod_{t \in J(N)} \lambda(t) \tag{9}$$

and may be used in a Bayesian update of a prior measure Π on λ to the posterior measure $\Pi(d\lambda|N) \propto p(N|\lambda)\Pi(d\lambda)$.

However, since no closed form analytical expression is typically available for the posterior measure, one needs to employ numerical algorithms, e.g., Markov chain Monte Carlo (MCMC), to carry out posterior inference on λ. For such numerical algorithms, a direct use of this exact likelihood function creates serious computational challenges. The evaluation of the integral $\int_0^T \lambda(t)dt$ involves the entire curve $\lambda(t)$, $t \in [0, T]$. Consequently, the numerical algorithm needs to run on the infinite dimensional space of curves, presenting nearly insurmountable computational difficulties. Rao and Teh (2011) circumvent this problem by augmenting additional latent variables which allow them to run a Gibbs sampler for MCMC computation. While this technique could be directly implemented to draw posterior inference on λ_{A} (or λ_{B}) based on only the A (B) trials data, its use in drawing inference on the α_j curves from AB trials data remains extremely challenging.

A less elegant but pragmatic alternative is to use time discretization. Fix an integer M and partition the response window $[0, T]$ into M contiguous subintervals $(0, w]$, $(w, 2w], \ldots, (T - w, T]$ of length $w = T/M$ each. Let $t_m^* = (m - 0.5)w$ be the midpoint of the m-th subinterval. When M is relatively large, one can appeal to the Riemann approximation of $\int_0^T \lambda(t)dt$ and express (9) as

$$p(N|\lambda) \approx \exp\left\{-\sum_{m=1}^{M} w\lambda(t_m^*)\right\} \prod_{m=1}^{M} \lambda(t_m^*)^{X_m} \propto \prod_{m=1}^{M} Poi(X_m|w\lambda(t_m^*)) \tag{10}$$

where $X_m = N(mw) - N((m-1)w)$ denotes the number of jumps in the m-th subinterval, and, the second and third terms are proportional as functions of λ.

By using (10), an MCMC now needs to be run only on the M-dimensional vector $(\lambda(t_1^*), \ldots, \lambda(t_M^*))$. Although one could obtain more accurate, M-term numerical approximation to $\int_0^T \lambda(t)dt$ by using Gaussian quadrature or Romberg's method, the equivalence of the second and third terms in (10) is a real advantage of using the Riemann approximation as it allows us to develop an extremely efficient Gibbs sampler based MCMC algorithm for joint posterior inference on all model parameters.

5.2 Reduced data and two-stage analysis

Following the notation of the above subsection, let X_{jm}^e, denote the spike count in the m-th subinterval for the j-th trial under experimental condition e, where, $m = 1, \ldots, M$, $j = 1, \ldots, n_e$, $e \in \{\mathrm{A}, \mathrm{B}, \mathrm{AB}\}$. Under the approximation given by (10), our data model now looks as follows:

1. $X_{jm}^{\mathrm{A}} \sim Poi(w\lambda_{\mathrm{A}}(t_m^*))$, $m = 1, \ldots, M$, $j = 1, \ldots, n_{\mathrm{A}}$,
2. $X_{jm}^{\mathrm{B}} \sim Poi(w\lambda_{\mathrm{B}}(t_m^*))$, $m = 1, \ldots, M$, $j = 1, \ldots, n_{\mathrm{B}}$,
3. $X_{jm}^{\mathrm{AB}} \sim Poi(w\{\alpha_j(t_m^*)\lambda_{\mathrm{A}}(t_m^*) + (1 - \alpha_j(t_m^*))\lambda_{\mathrm{B}}(t_m^*)\})$, $m = 1, \ldots, M$, $j = 1, \ldots, n_{\mathrm{AB}}$,

and all these random variables are independent of each other given λ_{A}, λ_{B} and α_j, $j = 1, \ldots, n_{\mathrm{AB}}$. Let $\mathbf{X}^e = (X_{jm}^e : 1 \le j \le n_e, 1 \le m \le M)$ denote the $n_e \times M$ dimensional data matrix of bin counts from experiment $e \in \{\mathrm{A}, \mathrm{B}, \mathrm{AB}\}$.

Notice that only the AB trial data $\mathbf{X}^{\mathrm{AB}}$ is relevant to second order stochasticity analysis as it provides information on the α_j curves and their unknown feature generating distribution $\mathbb{Q}$. Below, we first describe how posterior inference can be drawn on these quantities from $\mathbf{X}^{\mathrm{AB}}$ alone under the working assumption that λ_{A} and λ_{B} have already been estimated. Then, in Section 5.2.2 we describe how the estimates of λ_{A} and λ_{B} may be obtained by analyzing $\mathbf{X}^{\mathrm{A}}$ and $\mathbf{X}^{\mathrm{B}}$ in a preprocessing step. We also discuss how the uncertainty in these estimates may be incorporated in the the second stage analysis of $\mathbf{X}^{\mathrm{AB}}$.

5.2.1 MCMC inference for $\mathbb{Q}$ and α_j curves

Recall that underlying each α_j curve are a vector $(\phi_j, \psi_j, \boldsymbol{\pi}_j) \sim \mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$, a scalar $\ell_j \sim \boldsymbol{\pi}_j$ and a curve $\eta_j \sim \mathrm{GP}(\phi_j, \psi_j C^{\mathrm{SE}}_{\ell_j})$ such that $\alpha_j(t) = [1+\exp\{-\eta_j(t)\}]^{-1}$, $t \in [0, T]$, $j = 1, \ldots, n_{\mathrm{AB}}$. Clearly we can focus on the posterior distribution of these η_j curves instead of the original α_j's. The other model parameters to be estimated are $\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}$ and the precision parameter κ. However,

$$\begin{aligned}
p\big(\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}, \kappa, \, \{\eta_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} \mid \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}\big) \\
\propto p\big(\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} \mid \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}\big) \\
\times \prod_{j=1}^{n_{\mathrm{AB}}} p(\eta_j | \boldsymbol{\eta}_j, \phi_j, \psi_j, \ell_j) \\
\times p(\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}} | \kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}})
\end{aligned}$$

where $\boldsymbol{\eta}_j = (\eta_j(t_1^*), \ldots, \eta_j(t_M^*))$, $j = 1, \ldots, n_{\mathrm{AB}}$. Notice that each of the conditional probability distributions appearing in the last two lines above is available in closed form. Therefore, to obtain MCMC inference on all model parameters it suffices to focus on building a Markov chain sampler with target stationary distribution $p(\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} | \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}})$.

Toward this goal, first rewrite our Poisson observational model $X^{\mathrm{AB}}_{jm} \sim \mathit{Poi}(w\{\alpha_j(t_m^*)\lambda_{\mathrm{A}}(t_m^*) + (1 - \alpha_j(t_m^*))\lambda_{\mathrm{B}}(t_m^*)\})$ as

$$\begin{aligned}
(Z^{\mathrm{A}}_{jm}, Z^{\mathrm{B}}_{jm}) &\sim \mathit{Poi}(w\lambda_{\mathrm{A}}(t_m^*)) \times \mathit{Poi}(w\lambda_{\mathrm{B}}(t_m^*)) \\
(Y^{\mathrm{A}}_{jm}, Y^{\mathrm{B}}_{jm}) &\sim \mathit{Bin}(Z^{\mathrm{A}}_{jm}, \alpha_j(t_m^*)) \times \mathit{Bin}(Z^{\mathrm{B}}_{jm}, 1 - \alpha_j(t_m^*)) \\
X^{\mathrm{AB}}_{jm} &= Y^{\mathrm{A}}_{jm} + Y^{\mathrm{B}}_{jm},
\end{aligned}$$

with independence assumed across $j = 1, \ldots, n_{\mathrm{AB}}$, $m = 1, \ldots, M$. This representation is valid since

$$Z \sim \mathit{Poi}(\mu), Y|Z \sim \mathit{Bin}(Z, p) \implies (Y, Z - Y) \sim \mathit{Poi}(p\mu) \times \mathit{Poi}((1 - p)\mu). \tag{11}$$

Consequently, it is sufficient to construct a Markov chain sampler for the augmented target distribution

$$p(\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}}, \kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} \mid \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda\mathrm{B})$$

where $\mathbf{Z}^e = (Z^e_{jm} : 1 \le j \le n_{\mathrm{AB}}, 1 \le m \le M)$, $\mathbf{Y}^e = (Y^e_{jm} : 1 \le j \le n_{\mathrm{AB}}, 1 \le m \le M)$, $e \in \{\mathrm{A}, \mathrm{B}\}$. Algorithm 1 gives a schematic representation of our Markov chain sampler. All technical details are provided in Appendix A.

5.2.2 Estimating λ_{A} and λ_{B}

One could use any existing aggregation and smoothing technique to estimate the average firing rate curves λ_{A} and λ_{B} from A and B trial data. Popular techniques include kernel and spline smoothing as well as more advanced nonparametric methods (Kass et al., 2003; Rao and Teh, 2011). However, when either or both of n_{A} and n_{B} are small, it is important to account for the uncertainty in estimating these curves in the second stage analysis AB trial

```
Algorithm 1: Schematic description of the Markov chain sampler
  Input: Binned spike counts $\mathbf{X}^{\mathrm{AB}}$ from AB trials, and, $\lambda_{\mathrm{A}}$ and $\lambda_{\mathrm{B}}$ curves (evaluated at
         the bin midpoints). Also, starting values for the model parameters
         $\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}$. These values maybe drawn from the prior.
  Output: $S$ Markov chain samples of model parameters $\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}$
  for $s \leftarrow 1$ to $S$ do
      1. Impute $(\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}})$ by a combination of Poisson and binomial
         draws leveraging upon (11).
      2. Carry out a parameter-expanded Gibbs update of $\{\boldsymbol{\eta}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}$ by using
         the Pólya-Gamma augmentation method of Polson et al. (2013).
      3. Carry out a parameter-expanded Gibbs update of $\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}$ by
         using Algorithm 8 of Neal (2000).
      4. Carry out a parameter-expanded Gibbs update of $\kappa$ along the lines
         of Escobar and West (1995).
      5. Given the current grouping of $\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}$, update the shared
         parameters $(\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*)$ of each cluster $c$. Of these, $\boldsymbol{\pi}_c^*$ is updated by
         a Gibbs step by utilizing the multinomial-Dirichlet conjugacy, and,
         $(\phi_c^*, \psi_c^*)$ is updated by a combination of an independent proposal
         Metropolis-Hastings update for $\psi_c^*$, followed by a draw of $\phi_c^*$ from a
         normal distribution.
      6. Save current parameter values as the $s$-th sample draw.
  end
```

data. For a full Bayesian analysis, suppose these two unknown curves were assigned prior measures Π_{A} and Π_{B} respectively. Then posterior computation can proceed by first updating these priors to posteriors $\Pi_{\mathrm{A}}(\lambda_{\mathrm{A}}|\mathbf{X}^{\mathrm{A}})$ and $\Pi_{\mathrm{B}}(\lambda_{\mathrm{B}}|\mathbf{X}^{\mathrm{B}})$ by using data from only, respectively, the A and the B trials, and, then using these posteriors as new priors for λ_{A} and λ_{B} in the second stage analysis of $\mathbf{X}^{\mathrm{AB}}$ detailed above.

From a practicality perspective, it is most convenient to have the second-stage priors for λ_{A} and λ_{B} in the following form:

$$\Pi_e(\lambda_e(t_1^*), \ldots, \lambda_e(t_M^*)|\mathbf{X}^e) = \prod_{j=1}^{M} Ga(\lambda_e(t_m^*)|a_m^e, b_m^e), \quad e \in \{\mathrm{A}, \mathrm{B}\} \tag{12}$$

for some a_m^e, b_m^e, $m = 1, \ldots, M$, which depend only on $\mathbf{X}^e$, i.e., data from the condition $e \in \{\mathrm{A}, \mathrm{B}\}$. Such a structure allows us to fully exploit the conjugacy between the Poisson and the gamma families of distributions. One only needs to extend the MCMC updates detailed in Section 5.2.1 by making an additional set of draws of $\lambda_e(t_m^*) \sim Ga(a_m^e + \sum_j Z_{jm}^e, b_m^e + n_{\mathrm{AB}})$, independently across $e \in \{\mathrm{A}, \mathrm{B}\}$ and $m = 1, \ldots, M$. These draws could be made right after

Step 1 of 5.2.1.

We fix the parameters a^e_m, b^e_m by first smoothing the bin counts of the corresponding single stimulus spike trains. Each spike train is smoothed by using Friedman's super smoother (Friedman, 1984). The average and the variance of the smoothed spike trains are then taken to give the bin specific prior mean (a^e_m/b^e_m) and variance $(a^e_m/(b^e_m)^2)$ for the second stage analysis.

Remark 5. The product nature of the second stage prior in (12) is at best a working hypothesis. It may appear less than satisfactory because it introduces additional random variation across bins, even when prior mean and variances are smoothed. One could overcome this deficiency by using importance sampling correction. Suppose $\Pi^*_e(\lambda_e(t^*_1), \dots, \lambda_e(t^*_M)|\mathbf{X}^e)$, $e \in \{a, \mathrm{B}\}$ were the actual prior distributions one had intended to use for the second stage, but the MCMC was run with the product prior given in (12) with a^e_m, b^e_m properly chosen so as to match the first two moments under Π^*_e. One could then obtain Monte Carlo estimates under the intended prior by simply using weighted averages of the saved MCMC draws with the weights being given by the ratio of Π^*_e to Π_e evaluated at the drawn values of $(\lambda_e(t^*_1), \dots, \lambda_e(t^*_M))$.

5.3 Prediction

Inference on $\mathbb{Q}$ is best quantified and visualized through the weight curves α^* it is likely to produce in future AB trials. Such α^* may be simulated by making draws from the posterior predictive distribution

$$p(\alpha^*|\mathbf{X}^{\mathrm{AB}}, \mathbf{X}^{\mathrm{A}}, \mathbf{X}^{\mathrm{B}}) = \int p(\alpha^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}^{\mathrm{AB}}, \mathbf{X}^{\mathrm{A}}, \mathbf{X}^{b}) d\boldsymbol{\theta} \tag{13}$$

where $\boldsymbol{\theta} = (\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}})$ denotes the ensemble of all model parameters that are included in the MCMC sampling of Section 5.2.1. Draws of $\boldsymbol{\eta}^*$ from (13) may be made by drawing one α^* from $p(\alpha^*|\boldsymbol{\theta})$ for each saved draw of $\boldsymbol{\theta}$ from the Markov chain sampler. Let ϕ^*, ψ^*, $\boldsymbol{\pi}^*$, ℓ^* and $\boldsymbol{\eta}^*$ denote the latent quantities associated with α^* as in (3)-(5). Notice that,

$$p(\alpha^*|\boldsymbol{\theta}) = p(\alpha^*|\boldsymbol{\eta}^*, \phi^*, \psi^*, \ell^*) p(\boldsymbol{\eta}^*|\phi^*, \psi^*, \ell^*) p(\ell^*|\boldsymbol{\pi}^*) \tag{14}$$

$$\times\, p(\phi^*, \psi^*, \boldsymbol{\pi}^*|\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}) \tag{15}$$

and hence a draw of α^* from $p(\alpha^*|\theta)$ can be made by making draws from the four conditional distributions on the right hand side, proceeding sequentially from right to left. It is easy to make draws from the three posterior distributions appearing on (14) as they are governed purely by the relationships in (3)-(5). The conditional distribution in (15), again by the Pólya urn scheme representation of the Dirichlet process, is given by:

$$p(\phi^*, \psi^*, \boldsymbol{\pi}^*|\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}) = \frac{\kappa}{\kappa + n_{\mathrm{AB}}} G_\kappa + \frac{1}{\kappa + n_{\mathrm{AB}}} \sum_{c=1}^{K} \delta_{(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c)},$$

where K denote the number of distinct elements $(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c)$, $c = 1, \dots, K$, among the collection $\{(\phi_j, \psi_j, \boldsymbol{\pi}_j) : j = 1, \dots, n_{\mathrm{AB}}\}$.

6 Numerical studies

6.1 Case studies with synthetic data

We report here three simulation experiments in which we assessed the scope of the DAPP model and the accuracy of the statistical method introduced in this work. Each experiment consisted of one cell with a distinct form of second order stochasticity:

Experiment 1 (random selection). The cell always produces flat weight curves $\alpha(t) \equiv \alpha$, with the magnitude α drawn either uniformly from $(0.05, 0.25)$ with probability 60% or uniformly from $(0.85, 0.95)$ with probability 40%.

Experiment 2 (dynamic averaging with random period). The cell always produces sinusoidal weight curves $\alpha(t) = 0.01 + 0.49\{1 + \sin(2\pi \frac{a+t}{b})\}$ which oscillate between 0.01 and 0.99, where the random period b (in ms) is drawn uniformly from the range (400, 1000) and the random shift (also in ms) is drawn uniformly from $(0, b)$.

Experiment 3 (mixture of flat averaging and dynamic averaging). The cell produces a 50-50 mixture of flat and sinusoidal weight curves. For the flat curves, the time invariant magnitude is drawn uniformly from (0.4, 0.7). The sinusoidal curves are again of the form $\alpha(t) = 0.01 + 0.49\{1 + \sin(2\pi \frac{a+t}{b})\}$, but with the random period now drawn uniformly from the range (320, 340) and the random shift is drawn uniformly from (0, b).

For each experiment, we assumed $\lambda_{\mathrm{A}} \equiv 400$ (in Hz) and $\lambda_{\mathrm{B}} = 100$ and simulated 20 A trials, 20 B trials and 20 AB trials with a common response horizon of $T = 1000$ (in ms). The resulting 60 spike trains were analyzed under the DAPP model with a 50 ms bin-width used for time discretization. To assess what the DAPP model learned about the nature of the second order stochasticity, we focus on the posterior predictive summaries of three broad features of the weight curves: 1) the range of α defined as $\mathrm{range}(\alpha) = \max_{t\in[0,T]} \alpha(t) - \min_{t\in[0,T]} \alpha(t)$; 2) the long term average $\bar{\alpha}$; and, 3) the waviness as captured by the expected up-crossing count $0.16T/\ell$, with ℓ denoting the characteristic length-scale underlying α. Figures 3, 4 and 5 show the results of our data analysis for these three synthetic cases.

From the figures we conclude that the DAPP model is able to correctly recover the broad features of the second order stochasticity in each case. This is particularly evident from the second row plots. For Experiment 1, the posterior predictive histogram of range(α) is peaked near zero and the posterior predictive distribution of the expected up-crossing count is also peaked at the smallest allowable value of 0.1. Both pictures suggest the model correctly learned that cell 1 produces mostly flat curves. Furthermore, the posterior predictive histogram of $\bar{\alpha}$ is bimodal with peaks near zero and one, indicating that the model correctly learned nearly half of the flat weight curves are close to zero while the other half are close to one.

For Experiment 2, the posterior predictive histogram of range(α) is peaked near one and the distribution of the expected up-crossing count is peaked at 1 and 2, suggesting, quite accurately, that the cell mostly produces wavy weight curves that oscillate from one extreme to the other with a period that is mostly in the range of $T/2 = 500$ to $T/1 = 1000$ ms.

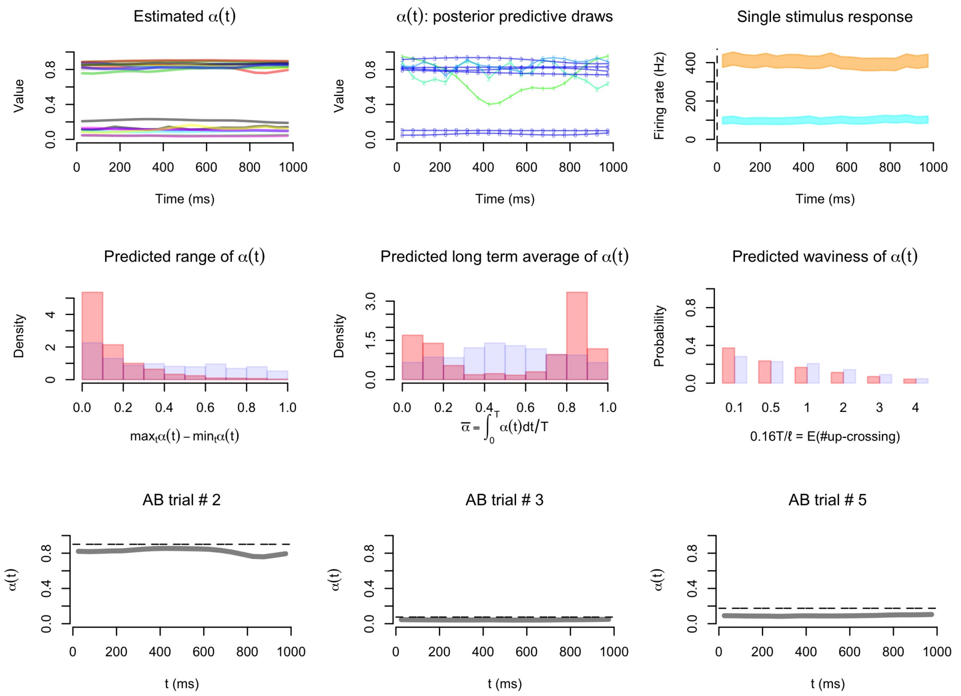


Figure 3: Analysis of Experiment 1 data consisting of 20 flat α curves. Eight of them had constant magnitudes in the range (0.05, 0.25) and the other 12 in the range (0.85, 0.95). Top row displays three posterior summaries: 1) estimates of the 20 α curves based on observed data; 2) 9 posterior predictive draws of α; and, 3) estimates and 95% credible intervals for $\lambda_{\mathrm{A}}(t)$ and $\lambda_{\mathrm{B}}(t)$. Second row shows (in pale red) the posterior predictive distributions of 1) $\mathrm{range}(\alpha) = \max_{t \in [0,T]} \alpha(t) - \min_{t \in [0,T]} \alpha(t)$; 2) the long term average $\bar{\alpha}$; and, 3) the waviness as captured by the expected up-crossing count. The histograms shown are created from 1000 posterior predictive draws of α and the underlying quantities. The pale blue histograms show the same but with draws of α made from the prior distribution. The third row compares the estimated α (solid line) against the true curve (dashed line) for 3 randomly chosen AB trials.

For Experiment 3, the posterior predictive histogram of $\mathrm{range}(\alpha)$ is bimodal with peaks near zero and one, and, the posterior predictive distribution of the expected up-crossing count is also bimodal with peaks at 0.1 and 3. These collectively suggest, again quite accurately, that the cell produces a mix of flat and wavy curves, where the latter ones oscillate the entire range with a period of about $T/3 = 333$ ms.

6.2 Second order stochasticity of inferior colliculus neurons

We report here results of the DAPP model analysis of the spiking activity of the three inferior colliculus (IC) cells referred to in Figures 1 and 2. We focus on these three specic cells to assess whether the DAPP model picks up different modes of naturally occurring second order stochasticity.

The neural data reported here comes from the data set described in Caruso et al. (2018). Briefly, the activity of individual neurons in the inferior colliculus was recorded while two

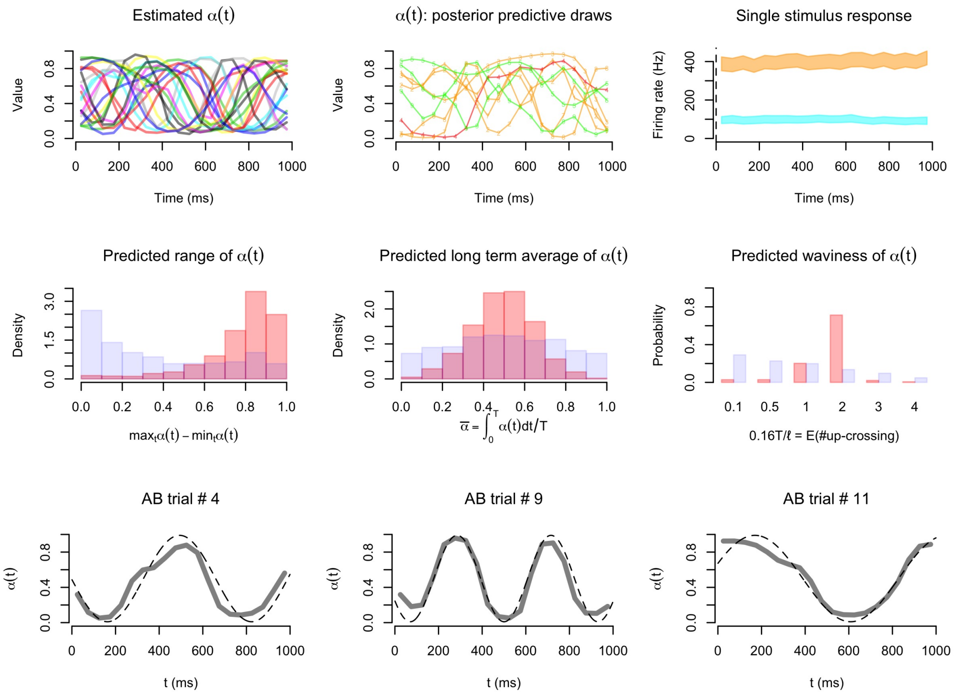


Figure 4: Analysis of Experiment 2 data consisting of 20 sinusoidal α curves, each swinging between 0.01 and 0.99 with a period (in ms) drawn randomly from the interval (400, 1000).

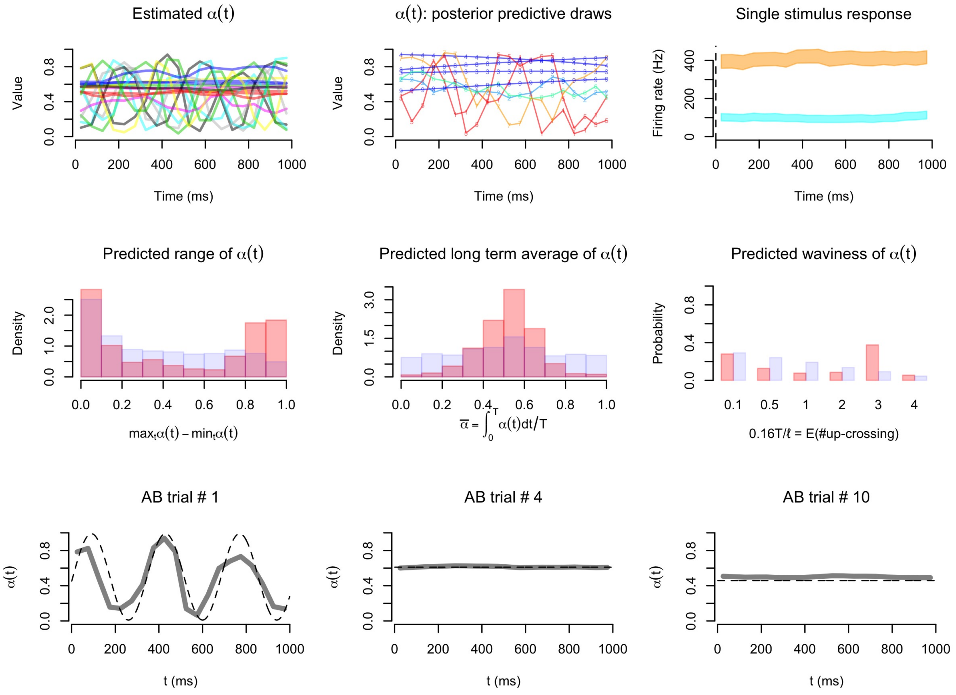


Figure 5: Analysis of Experiment 3 data consisting of 11 flat (magnitude between 0.4 and 0.7) and 9 sinusoidal α curves (swing between (0.01, 0.99), period between (320, 340) ms).

monkeys listened for and made eye movements to the locations of sounds. Each trial began with the onset of a visual target located straight ahead, which the monkey was required to fixate on before the trial could proceed. Then, either one or two sounds were presented. These sounds stayed on for 600-1000 ms, before the fixation light was extinguished, cuing the monkey to make eye movements to each sound (one if one sound, two if two sounds). The dual sounds were located at either (-24 deg, +6 deg) or (-6 deg, +24 deg) horizontally, and consisted of bandpass noise with different center frequencies (742 Hz and another frequency that differed by a ratio of 1.22 or an integer power of that ratio. Single sounds were drawn from the same set of locations and frequencies that were used on the dual sound trials. The neural activity was analyzed during the first 600-1000 ms of the epoch that the sounds were on but the monkey was maintaining fixation. All conditions were randomly interleaved.

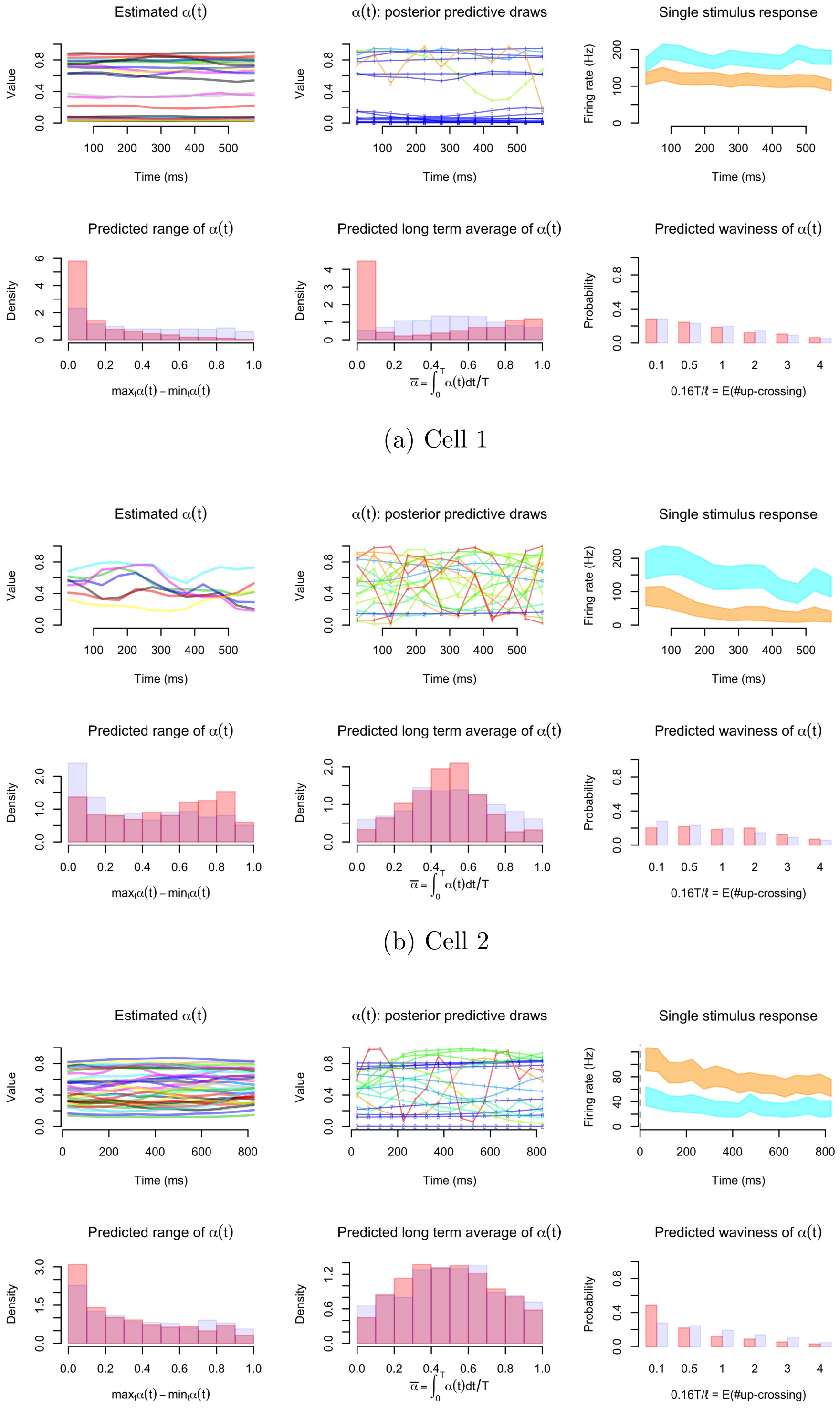
Of the three cells shown here, Cell 1 comes from monkey P and involved a 903 Hz sound (A) at 24 degrees to the left and 742 Hz sound (B) at 6 degrees to the right. There were 29 A trials, 21 B trials and 34 AB trials, and the analysis was conducted on a 600 ms response period. Cell 2 and Cell 3 recordings were made from monkey Y. For Cell 2, sound B was centered at 742 Hz and located at 6 degrees to the left and sound A was 609 Hz located at 24 degrees to the right. Cell 3 involved the same sound B frequency and location (742 Hz, 6 degrees left), but sound A was a 500 Hz sound (24 degrees right). Cell 2's response period was 600 ms whereas cell 3 had a response period of length 1000 ms. Cell 2 had 7 trials each in conditions A, B and AB. Cell 3 had 15 A trials, 11 B trials and 37 AB trials.

An underlying assumption of the DAPP model is that the AB firing rates lie within the range defined by the A and B firing rates. For the IC cells, we assess the validity of this assumption through whole-trial spike counts. Recall that Figure 2 shows smoothed histograms of whole trial spike counts grouped by conditions A, B and AB. Notice that for each cell, the AB distribution appears to sit between the distributions under conditions A and B, conforming with the DAPP assumption.

Figure 6 visualizes the DAPP analysis results for the 3 IC cells. It is immediately apparent that the three cells exhibit different patterns of second order stochasticity. Cell 1 exhibits a second order stochasticity pattern similar to random selection. It has nearly 50% chance of producing an α curve that is flat and very close to zero, whereas, with nearly 30% probability it would produce a flat α curve with magnitude in the range (0.6, 1). That is, when exposed to both A and B sounds, almost half of the time the cell would respond like it is responding only to sound B. But in about every third trial its response will be more resembling of its sound A spiking activity, although this latter resemblance is less exact.

In contrast, cell 2 appears to have a higher likelihood of producing a wavy α curve. The posterior predictive distribution of range(α) is bimodal with 40% mass concentrated in the interval (0.6, 0.9). Recall that range(α) ≥ 0.6 means that the weight $\alpha(t)$ attached to $\lambda_{\mathrm{A}}(t)$ dynamically swings from 0-20% to 80-100% (or the reverse) – a phenomenon consistent with within trial random interleaving. However, the posterior distribution of expected up-crossing count does not support many oscillations. The likely scenario is that a future AB trial firing rate curve would exhibit one swing from being nearly A like to nearly B like. Notice that the there is an overall lack of high concentration of these posterior distributions underlining that the DAPP model parameters were not learned with high confidence. This is not surprising given that the cell had only 7 trials in each condition.

Cell 3 appears to have a high propensity of producing flat α curves (50% chance of no



(a) Cell 1

(b) Cell 2

(c) Cell 3

Figure 6: Posterior inference for three example IC cells.

up-crossing) with the long term average distributed over the entire interval $(0, 1)$ with a mild concentration near the center. This pattern is entirely different from the concept of random selection or interleaving. Here, on any AB trial, the cell appears to exhibit a spiking activity with a firing rate that is a non-dynamic weighted average of λ_{A} and λ_{B}, but the weights assigned to the two pure sound average firing rates vary across trials with a concentration around the case where both sounds are given equal weights.

7 Discussion

In this paper we have introduced a novel concept of second order stochasticity in neuronal firing rates in response to a stimuli bundle. The very definition of second order stochasticity, rooted in the information-preserving, stochastic variation of the firing rate curve from one trial to the next, rules out the commonly used time-and-trial aggregated statistical methods for analyzing spike train data. We have developed a detailed point pattern model, namely the DAPP model, based on the assumption of stochastically varying, dynamic averaging of single stimulus firing rate curves. Our model is generative in nature. The fitted model can be used to draw inference on how the cell is likely to respond in future hypothetical trials under the stimuli bundle exposure.

Our treatment of second order stochasticity leaves room for many further developments. The overarching assumption of dynamic averaging can only explain special kinds of second order stochasticity where under the stimuli bundle exposure, the overall firing rate of the cell resides in between the rates it exhibits under each individual stimulus. Stimuli bundles that evoke either enhancement or suppression of activity, i.e., producing rates outside the range of single stimulus response rates, cannot be analyzed with the DAPP model.

Our model assumes spike counts are Poisson distributed with possibly time varying firing rate curves. It is known that the Poisson assumption does not always provide the best fit to inter-spiking interval distributions observed in reality. One issue with the Poisson assumption is its inability to account for the refractory period which is a short time gap immediately after a spike during which the neuron cannot fire again no matter what stimulus is presented to it. However, this is not a big issue in our applications where spiking activity is aggregated in 50 ms time bins, which is much longer time scale than the typical length of a refractory period which is usually no more than 2 ms. A second issue with the Poisson assumption is its inability to account for *over-dispersion* where the variance of the spiking activity is larger than its mean. This may be accounted for by extending our DAPP model where the Poisson assumption is replaced with a negative-binomial assumption. However, defining and analyzing negative-binomial point processes with smooth rate curves pose serious technical challenges that are beyond the scope of the current paper.

In applying the DAPP model and method developed here, one needs to choose the binning interval width to carry out the time discretization of spiking activity. While shorter bins allow more flexible estimation of the time varying dynamics of the α curves, the increased number of bins adds to computing cost. Shorter bins, with fewer spike counts in each, also pose some difficulty to our two-stage estimation process for the λ_{A} and λ_{B} curves. Specifically, the second stage prior in (12) which assumes conditional independence of the curve values across bins, do not offer adequate smoothing. There is potential to remedy this problem by replacing the product gamma prior in (12) with an autoregressive gamma prior (Wolpert and Ickstadt, 1998). Our choice of 50 ms for bin width was based on our understanding of the scale of IC firing rate. We also repeated the analyses in Section 6.2 with 25 ms bins and the results were robust to this change in bin width.

These challenges notwithstanding, the DAPP analysis framework presented in this paper offers an important first step toward understanding, modeling and estimating second order stochasticity. Caruso et al. (2018) give strong evidence of the prevalence of second order stochasticity in the primate brain, as well as, of the utility of the DAPP analysis in cataloging various modes of such stochastic variation. Clearly, it will take a system level understanding of neural computing to completely describe how the brain might represent multiple simultaneous signals. The cell level DAPP analysis promises to be an important building block toward such a goal.

Appendix A Details of Markov chain sampling

Step 1. Gibbs update of $(\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}})$. Since

$$\begin{aligned} p(\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, &\mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}} \mid \kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}, \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) \\ &= p(\mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}} \mid \mathbf{X}^{\mathrm{AB}}, \{\boldsymbol{\eta}_j\}_{j=1}^{n_{\mathrm{AB}}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) \\ &\qquad \times p(\mathbf{Z}^{\mathrm{A}} | \mathbf{Y}^{\mathrm{A}}, \{\boldsymbol{\eta}_j\}_{j=1}^{n_{\mathrm{AB}}}, \lambda_{\mathrm{A}}) \times p(\mathbf{Z}^{\mathrm{B}} | \mathbf{Y}^{\mathrm{B}}, \{\boldsymbol{\eta}_j\}_{j=1}^{n_{\mathrm{AB}}}, \lambda_{\mathrm{B}}), \end{aligned}$$

one can draw $(\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}})$ from their joint conditional posterior by drawing, in succession, from the three conditional distributions on the right of the above display. This proceeds in three steps for each $j = 1, \dots, n_{\mathrm{AB}}$, $m = 1, \dots, M$:

1. Draw $Y_{jm}^{\mathrm{A}} \sim Bin(X_{jm}^{\mathrm{AB}}, \frac{\alpha_{jm}\lambda_{\mathrm{A}}(t_m^*)}{\alpha_{jm}\lambda_{\mathrm{A}}(t_m^*) + (1-\alpha_{jm})\lambda_{\mathrm{B}}(t_m^*)})$ and set $Y_{jm}^{\mathrm{B}} = X_{jm}^{\mathrm{AB}} - Y_{jm}^{\mathrm{A}}$, where $\alpha_{jm} = 1/(1 + e^{-\eta_{jm}})$.
2. Draw $\bar{Z}_{jm}^{\mathrm{A}} \sim Poi(w(1 - \alpha_j(t_m^*))\lambda_{\mathrm{A}}(t_m^*))$ and set $Z_{jm}^{\mathrm{A}} = Y_{jm}^{\mathrm{A}} + \bar{Z}_{jm}^{\mathrm{A}}$.
3. Draw $\bar{Z}_{jm}^{\mathrm{B}} \sim Poi(w\alpha_j(t_m^*)\lambda_{\mathrm{B}}(t_m^*))$ and set $Z_{jm}^{\mathrm{B}} = Y_{jm}^{\mathrm{B}} + \bar{Z}_{jm}^{\mathrm{B}}$

Step 2. Parameter-expanded Gibbs update of $\{\boldsymbol{\eta}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}$. We can write

$$\begin{aligned} p(\{\boldsymbol{\eta}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} &\mid \mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}}, \kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}, \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) \\ &= p(\{\boldsymbol{\eta}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}} \mid \mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}}, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}) \\ &\propto \prod_{j=1}^{n_{\mathrm{AB}}} \left\{ \prod_{m=1}^{M} Bin\left(y_{jm}^* \middle| N_{jm}, \frac{1}{1 + e^{-\eta_{jm}}} \right) \right\} N(\boldsymbol{\eta}_j | \phi_j \mathbf{1}, \psi_j \mathbf{C}_{\ell_j}) P_{\mathcal{L}, \boldsymbol{\pi}_j}(\ell_j), \end{aligned}$$

where $y_{jm}^* = Y_{jm}^{\mathrm{A}} + Z_{jm}^{\mathrm{B}} - Y_{jm}^{\mathrm{B}}$, $N_{jm} = Z_{jm}^{\mathrm{A}} + Z_{jm}^{\mathrm{B}}$, $\mathbf{1}$ is the M-dimensional vector of ones, and, $\mathbf{C}_{\ell_j}$ is an $M \times M$ covariance matrix with (m, m')-th element $C_{\ell_j}^{\mathrm{SE}}(t_m^*, t_{m'}^*)$. The factorization over j allows us to make parallel Gibbs updates of $(\boldsymbol{\eta}_j, \ell_j)$ across $j = 1, \dots, n_{\mathrm{AB}}$, which can be accomplished by the Pólya-gamma parameter augmentation trick of Polson et al. (2013). This proceeds in two steps:

1. Draw $\omega_{jm} \sim PG(N_{jm}, \eta_{jm})$, independently across $m = 1, \ldots, M$. Here $PG(b, c)$ denote the Pólya-gamma distribution with shape parameter b and tilting parameter c. Set $\mathbf{\Omega}_j = \mathrm{diag}(\omega_{j1}, \ldots, \omega_{jM})$.

2. Update $(\boldsymbol{\eta}_j, \ell_j)$ based on the local model:

$$\bar{\mathbf{y}}_j|(\boldsymbol{\eta}_j, \ell_j) \sim N(\boldsymbol{\eta}_j, \mathbf{\Omega}_j^{-1}), \quad \boldsymbol{\eta}_j|\ell_j \sim N(\phi_j \mathbf{1}, \psi_j \mathbf{C}_{\ell_j}), \quad \ell_j \sim P_{\mathcal{L}, \boldsymbol{\pi}_j}$$

   where $\bar{\mathbf{y}}_j = (\bar{y}_{j1}, \ldots, \bar{y}_{jM})$ with $\bar{y}_{jm} = y^*_{jm} - N_{jm}/2$. That is, with $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jL})$, one first draws ℓ_j from $\{\ell^*_1, \ldots, \ell^*_L\}$ according to probabilities $(q^j_1, \ldots, q^j_L)$ where $q^j_i \propto \pi_{ji} N(\bar{y}_j|\phi_j \mathbf{1}, \psi_j \mathbf{C}_{\ell^*_i} + \mathbf{\Omega}_j^{-1})$. Then one draws $\boldsymbol{\eta}_j$ from $N(\mathbf{m}, \mathbf{S})$ where $\mathbf{S} = (\mathbf{\Omega}_j + \psi_j^{-1} \mathbf{C}_{\ell_j}^{-1})^{-1}$ and $\mathbf{m} = \mathbf{S}(\mathbf{\Omega}_j \bar{\mathbf{y}}_j + \phi_j \psi_j^{-1} \mathbf{C}_{\ell_j}^{-1} \mathbf{1})$

Step 3. Parameter-expanded Gibbs update of $\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}$. We sequence through $j = 1, \ldots, n_{\mathrm{AB}}$ and update (ψ_j, ψ_j) given all other parameter values. At any such instance $j = i \in \{1, \ldots, n_{\mathrm{AB}}\}$,

$$\begin{aligned} p(\phi_i, &\psi_i, \boldsymbol{\pi}_i | \mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}}, \kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j \neq i}, \boldsymbol{\eta}_i, \ell_i, \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) \\ &= p(\phi_i, \psi_i, \boldsymbol{\pi}_i | \kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j \neq i}, \boldsymbol{\eta}_i, \ell_i) \\ &\propto p(\boldsymbol{\eta}_i | \phi_i, \psi_i, \ell_i) p(\ell_i | \boldsymbol{\pi}_i) p(\phi_i, \psi_i, \boldsymbol{\pi}_i | \kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j \neq i}) \end{aligned}$$

with $p(\boldsymbol{\eta}_i|\phi_i, \psi_i, \ell_i) = N(\boldsymbol{\eta}_i|\phi_i \mathbf{1}, \psi_i \mathbf{C}_{\ell_i})$, and, by the Pólya urn scheme representation of the Dirichlet process,

$$p(\phi_i, \psi_i, \boldsymbol{\pi}_i | \kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j \neq i}) = \frac{1}{\kappa + n_{\mathrm{AB}} - 1} \left\{ \kappa G_\kappa + \sum_{c=1}^{K^-} n_{-i,c} \delta_{(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c)} \right\}$$

where $(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c)$, $c = 1, \ldots, K^-$, are the distinct elements in the collection $\{(\psi_j, \psi_j, \boldsymbol{\pi}_j) : j \neq i\}$ with $n_{-i,c}$ giving the number of times the c-th distinct element appears in the collection. Hence, Algorithm 8 of Neal (2000), with a given auxiliary sample size r, can be used to update $(\phi_i, \psi_i, \boldsymbol{\pi}_i)$ as follows:

1. Draw r additional pairs $(\phi^*_{K^-+h}, \psi^*_{K^-+h}, \boldsymbol{\pi}^*_{K^-+h}) \sim G_\kappa$, $h = 1, \ldots, r$. If $(\phi_i, \psi_i, \boldsymbol{\pi}_i) \notin \{(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c) : 1 \leq c \leq K^-\}$, then replace the first additional draw $(\phi^*_{K^-+1}, \psi^*_{K^-+1}, \boldsymbol{\pi}^*_{K^-+1})$ with $(\phi_i, \psi_i, \boldsymbol{\pi}_i)$.

2. Draw an index c_i from $\{1, \ldots, K^- + r\}$ according to probabilities $p_c \propto s_c \cdot P_{\mathcal{L}, \boldsymbol{\pi}^*_c}(\ell_i) \cdot N(\boldsymbol{\eta}_i|\phi^*_c \mathbf{1}, \psi^*_c \mathbf{C}_{\ell_i})$, $1 \leq c \leq K^- + r$, where $s_c = n_{-i,c}$ for $c = 1, \ldots, K^-$, and, $s_c = \kappa/r$ for $c = K^- + 1, \ldots, K^- + r$.

Step 4. Parameter-expanded Gibbs update of κ. Now, let $(\phi^*_c, \psi^*_c, \boldsymbol{\pi}^*_c)$, $c = 1, \ldots, K$, denote the distinct elements in the full collection $\{(\phi_j, \psi_j, \boldsymbol{\pi}_j) : j = 1, \ldots, n_{\mathrm{AB}}\}$. Following

Escobar and West (1995), we write

$$\begin{aligned}
&p(\kappa|\mathbf{Z}^{\mathrm{A}}, \mathbf{Z}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{A}}, \mathbf{Y}^{\mathrm{B}}, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{\mathrm{AB}}}, \mathbf{X}^{\mathrm{AB}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) \\
&\quad \propto p(\kappa) p(\{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}|\kappa) \\
&\quad \propto e^{-\kappa} \frac{\kappa^K}{\prod_{j=1}^{n_{\mathrm{AB}}}(\kappa + j - 1)} \prod_{c=1}^{K} \{\kappa(1-\psi_c^*)^{\kappa-1}\} \\
&\quad \propto B(\kappa, n_{\mathrm{AB}}) \kappa^{2K} e^{-b\kappa} \\
&\quad = \int_0^1 \omega^{\kappa-1}(1-\omega)^{n_{\mathrm{AB}}-1} \kappa^{2K} e^{-b\kappa} d\omega
\end{aligned}$$

where $b = 1 - \sum_{c=1}^{K} \log(1-\psi_c^*)$ and $B(a,b) = \int_0^1 \omega^{a-1}(1-\omega)^{b-1} d\omega = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ denotes the beta function. Therefore the conditional posterior density function of κ is identical to its marginal under the joint density function $p(\omega, \kappa) \propto \omega^{\kappa-1}(1-\omega)^{n_{\mathrm{AB}}-1}\kappa^{2K}e^{-b\kappa}$, $\omega \in (0,1)$, $\kappa > 0$. Consequentially, a valid Gibbs update of κ is obtained by the following two steps:

1. Draw $\omega \sim p(\omega|\kappa) = Be(\omega|\kappa, n_{\mathrm{AB}})$, and, then
2. Draw $\kappa \sim p(\kappa|\omega) \propto \omega^{\kappa}\kappa^{2K}e^{-b\kappa} = Ga(\kappa|2K+1, b - \log(\omega))$

Step 5. Other updates. Following the recommendation of Neal (2000), we carry out additional Gibbs like updates for the cluster specific parameters $(\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*)$, $c = 1, \ldots, K$, to improve mixing of the Markov chain sampler. For any cluster c, let $S_c = \{1 \le j \le n_{\mathrm{AB}} : (\phi_j, \psi_j, \boldsymbol{\pi}_j) = (\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*)\}$ denote the collection of AB trials belonging to that cluster. The conditional posterior distribution of $(\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*)$, given all other model parameters and latent variables, then depend only on the reduced model:

$$\boldsymbol{\eta}_j \sim N(\phi_c^* \mathbf{1}, \psi_c^* \mathbf{C}_{\ell_j}), \quad \ell_j \sim P_{\mathcal{L}, \boldsymbol{\pi}_c^*}, \; j \in S_c, \quad (\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*) \sim G_\kappa.$$

First, by utilizing the multinomial-Dirichlet conjugacy, we update $\boldsymbol{\pi}_c^*$ by making a draw from $Dir(a_1 + n_1^c, \ldots, a_L + n_L^c)$ where $n_i^c = \#\{j \in S_c : \ell_j = \ell_i^*\}$, $i = 1, \ldots, L$. Next, calculate four *summary statistics* for the cluster:

$$u_c = \sum_{j \in S_c} \mathbf{1}^T \mathbf{C}_{\ell_j}^{-1} \mathbf{1}, \quad v_c = \sum_{j \in S_c} \mathbf{1}^T \mathbf{C}_{\ell_j}^{-1} \boldsymbol{\eta}_j, \quad w_c = \sum_{j \in S_c} \boldsymbol{\eta}_j^T \mathbf{C}_{\ell_j}^{-1} \boldsymbol{\eta}_j, \text{ and, } z_c = \frac{v_c}{u_c}.$$

From the reduced model and the choice of G_κ, we write $p(\phi_c^*, \psi_c^*|-) = p(\psi_c^*|-)p(\phi_c^*|\psi_c^*, -)$ where

$$p(\phi_c^*|\psi_c^*, -) \propto N(\phi_c^*|0, 1-\psi_c^*) \prod_{j \in S_c} N(\boldsymbol{\eta}_j|\phi_c^* \mathbf{1}, \psi_c^* \mathbf{C}_{\ell_j}) \propto N(\phi_c^*|m_c, s_c^2) \tag{16}$$

with $s_c^2 = \psi_c^*(1-\psi_c^*)/\{\psi_c^* + (1-\psi_c^*)u_c\}$, $m_c = (1-\psi_c^*)v_c/\{\psi_c^* + (1-\psi_c^*)u_c\}$, and,

$$\begin{aligned}
p(\psi_c^*|-) &\propto Be(\psi_c^*|1, \kappa) \int N(\phi_c^*|0, 1-\psi_c^*) \prod_{j \in S_c} N(\boldsymbol{\eta}_j|\phi_c^* \mathbf{1}, \psi_c^* \mathbf{C}_{\ell_j}) d\phi_c^* \\
&\propto Be(\psi_c^*|2, \kappa) N\left(z_c|0, \psi_c^* u_c^{-1} + 1 - \psi_c^*\right) IG\left(\psi_c^* \,\middle|\, \tfrac{M|S_c|-1}{2}, \tfrac{w_c - z_c^2 u_c}{2}\right).
\end{aligned} \tag{17}$$

The shape parameter of the inverse-gamma distribution on the last expression is well defined whenever the cluster size $|S_c|$ is at least two. The rate parameter is always well defined by the Cauchy-Schwartz theorem since $\langle(\mathbf{a}_j : j \in S_c), (\mathbf{b}_j : j \in S_c)\rangle = \sum_{j \in S_c} \mathbf{a}_j^T \mathbf{C}_{\ell_j}^{-1} \mathbf{b}_j$ defines an inner product on $(\mathbb{R}^M)^{S_c}$.

Consequently, we update ψ_c^* by a Metropolis-Hastings step where we propose ψ_c' from the inverse-gamma distribution in (17), and, accept the proposal with probability given by the Hastings ratio $\min\{1, f(\psi_c')/f(\psi_c^*)\}$ where $f(\psi) = Be(\psi|2,\kappa)N(z_c|0, \psi u_c^{-1} + 1 - \psi)$. Next ϕ_c^* is updated by making a draw from the normal distribution on the extreme right of (16).

Appendix B Dirichlet process prior on $\mathbb{Q}$

An alternative prior specification on $\mathbb{Q}$ is to directly assign it a Dirichlet process prior distribution. This would amount to modeling $\mathbb{Q}$ as

$$\mathbb{Q} = \sum_{h=1}^{\infty} \omega_h \delta_{(\phi_h^*, \psi_h^*, \ell_h^*)}$$

where the atoms $(\phi_h^*, \psi_h^*, \ell_h^*)$, $h = 1, 2, \ldots$, are drawn independently from a base measure G_κ' on $(-\infty, \infty) \times (0, \infty) \times \mathcal{L}$, and, the weights ω_h, $h = 1, 2, \ldots$, are given by the stick-breaking construction as detailed in Section 4.2. To mimic the specification of Section 4.3, the base measure G_κ' could be taken as in (8), but with the part $\boldsymbol{\pi}^* \sim Dir(a_1, \ldots, a_K)$ now replaced with a matching counterpart $\ell^* \sim \mathcal{P}_{\mathcal{L}^*, \bar{a}}$ where $\bar{a}$ gives the probability vector obtained by normalizing $a = (a_1, \ldots, a_K)$.

In contrast to the DAPP specification, this alternative formulation enforces a hard coupling between ℓ and (ϕ, ψ) by removing the intermediary $\boldsymbol{\pi}$ in (6). Here, all AB trials in a cluster must share a common value for each of these three parameters. While this specification appears simpler than ours, and, a hard coupling may be scientifically meaningful, it leads to more challenging posterior computation. Specifically, in our experience, an adaptation of Algorithm 1 to the alternative formulation, leads to considerably poorer mixing of the resulting Markov chain sampling.

Figure 7 compares Monte Carlo estimates of the posterior predictive distribution of ℓ from three independent runs of the MCMC in fitting the DAPP model or its alternative to the synthetic data set from Experiment 3 (Section 6.1). Each Markov chain was run for 10,000 iterations of which first 1,000 draws were discarded, and, 1,000 samples were saved from the remainder of the chain by thinning it uniformly. We measured Monte Carlo error as $\max_c \|p^c - \bar{p}\|_1$, where p^c denotes the posterior predictive probability vector from chain c, $c \in \{1, 2, 3\}$ and $\bar{p}$ is the average across the three chains. For the DAPP model, the Monte Carlo error is 0.07 whereas for the alternative model it equals 0.37.

References

Adler, R. J. and J. E. Taylor (2009). Random fields and geometry. Springer Science & Business Media.

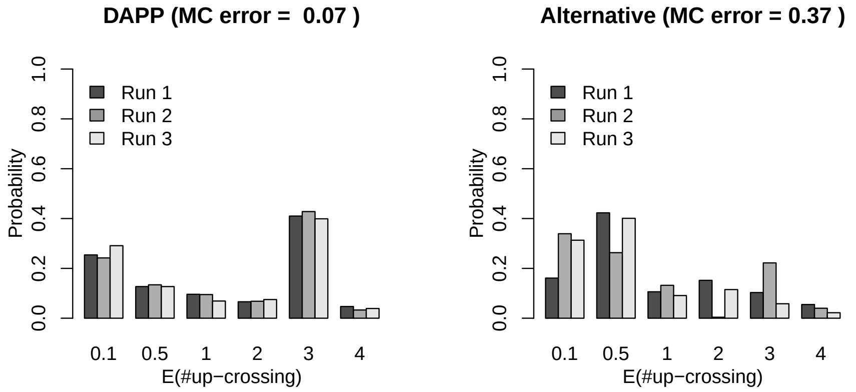


Figure 7: Evidence of poor MCMC mixing when a Dirichlet process prior is assigned directly to $\mathbb{Q}$, as opposed to the hierarchical formulation adopted in DAPP. Left panel shows estimates of the posterior predictive distribution of ℓ from three independent MCMC runs of DAPP. Right panel shows the same for the alternative specification. The data set from Experiment 3 is used for model fitting.

Caruso, V. C., J. T. Mohl, C. Glynn, J. Lee, S. M. Willett, A. Zaman, A. F. Ebihara, R. Estrada, W. A. Freiwald, S. T. Tokdar, et al. (2018). Single neurons may encode simultaneous stimuli by switching between activity patterns. Nature communications 9(1), 2715.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. Journal of the american statistical association 90(430), 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1(2), 209–230.

Friedman, J. H. (1984). A variable span smoother. Technical report, Stanford Univ CA lab for computational statistics.

Gerstein, G. L. and N. Y.-S. Kiang (1960). An approach to the quantitative analysis of electrophysiological data from single neurons. Biophysical Journal 1(1), 15.

Kass, R. E., V. Ventura, and E. N. Brown (2005). Statistical issues in the analysis of neuronal data. Journal of neurophysiology 94(1), 8–25.

Kass, R. E., V. Ventura, and C. Cai (2003). Statistical smoothing of neuronal data. Network-Computation in Neural Systems 14(1), 5–16.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics 9(2), 249–265.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. Journal of the American Statistical Association 108(504), 1339–1349.

Rao, V. and Y. W. Teh (2011). Gaussian process modulated renewal processes. In Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 2474–2482. Curran Associates Inc.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Ventura, V., C. Cai, and R. E. Kass (2005). Trial-to-trial variability and its effect on time-varying dependency between two neurons. Journal of neurophysiology 94(4), 2928–2939.

Wolpert, R. L. and K. Ickstadt (1998). Poisson/gamma random field models for spatial statistics. Biometrika 85(2), 251–267.