The Annals of Applied Statistics
2021, Vol. 15, No. 1, 41–63
https://doi.org/10.1214/20-AOAS1383

ANALYZING SECOND ORDER STOCHASTICITY OF NEURAL SPIKING UNDER STIMULI-BUNDLE EXPOSURE

BY CHRIS GLYNN[1], SURYA T. TOKDAR[2], AZEEM ZAMAN[3], VALERIA C. CARUSO[4], JEFF T. MOHL[5,*], SHAWN M. WILLETT[6] AND JENNIFER M. GROH[5,†]

[1]*Zillow Research, christophergl@zillowgroup.com*
[2]*Department of Statistical Science, Duke University, surya.tokdar@duke.edu*
[3]*Department of Statistics, Harvard University, azaman@g.harvard.edu*
[4]*Center for Human Growth and Development, University of Michigan, vcaruso@umich.edu*
[5]*Department of Psychology and Neuroscience, Duke University, *jeffrey.mohl@duke.edu; †jmgroh@duke.edu*
[6]*Department of Neurobiology, Duke University, shawn.willett@duke.edu*

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 multiitem 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 point 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 illustrate the utility and accuracy of the DAPP analysis with synthetic data simulation studies. We present real-world evidence of second order stochastic variation with an analysis of monkey inferior colliculus recordings under auditory stimuli.

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)).

Received March 2019; revised June 2020.
Key words and phrases. Spike train, multiple stimuli, dynamic admixture of Poisson processes, Gaussian process, Dirichlet process, Bayesian inference.

For simplicity, and also limited by available experimental data, we restrict this discussion to stimuli bundles consisting of two stimuli which evoke measurably different 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, Ventura and Brown (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 $\lambda_A(t)$ and $\lambda_B(t)$ 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, Ventura and Brown (2005), Ventura, Cai and Kass (2005)). Our focus is on quantifying variability that is specifically activated by the presence of two stimuli at the same time. Disambiguation of second order stochasticity from trial-to-trial variations that are not specific to the AB condition is an additional statistical challenge which is addressed in Section 6.3.

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 $\alpha(t)\lambda_A(t) + (1 - \alpha(t))\lambda_B(t)$, 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 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 $\lambda_A(t)$ and $\lambda_B(t)$ 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 subtrial durations.

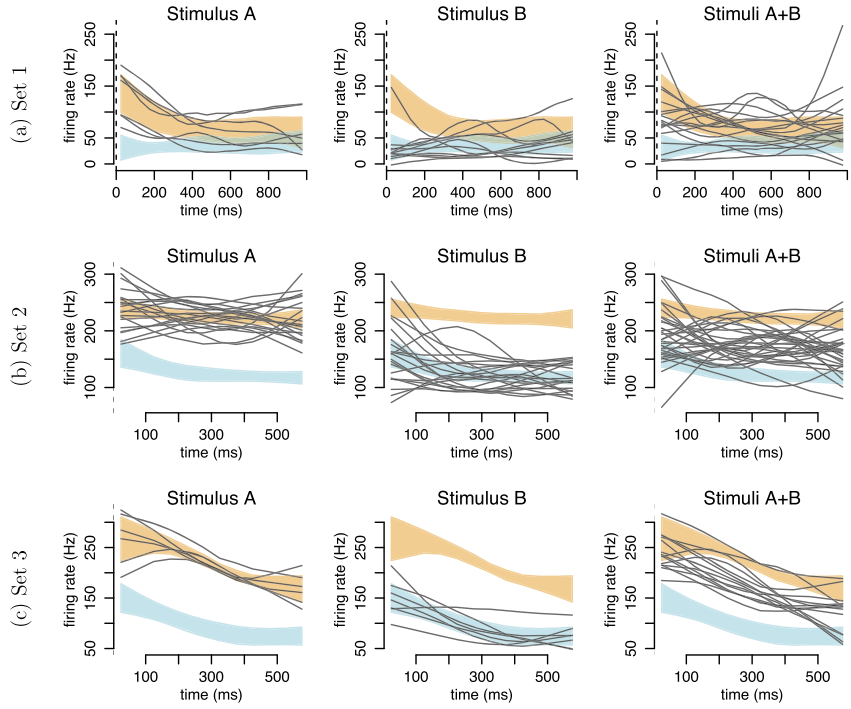

FIG. 1. *Multiple forms of stochasticity in inferior colliculus (IC). Each row corresponds to a distinct experiment set with recording from monkey IC and shows how a cell responds to a triplet of experimental conditions A, B and AB, where A and B each corresponds to an auditory stimulus in the form of 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) Set* 1*: AB responses appear to be a superimposition of A and B responses. (b) Set* 2*: AB responses appear to fluctuate more widely within each trial than A or B responses. (c) Set* 3*: AB responses appear nonwavering and A-like within any trial but partially shifted toward a middling firing rate.*

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 an 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 belonging to three different sets of our experiments. For the cell in Set 1, AB responses appear to be a superimposition of A and B responses (Figure 1), conforming to random selection of signals at the whole trial level. Correspondingly, the AB spike count distribution appears as a mixture of A and B spike count distributions (Figure 2). In contrast, in both Sets 2 and 3 the cells have their AB spike count distribution sit in between their A and B spike count distributions. However, in Set 2, AB responses appear to fluctuate more widely within each trial than A or B responses, whereas, in Set 3, AB responses appear nonwavering and A-like within any trial but partially shifted toward a middling firing rate.

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

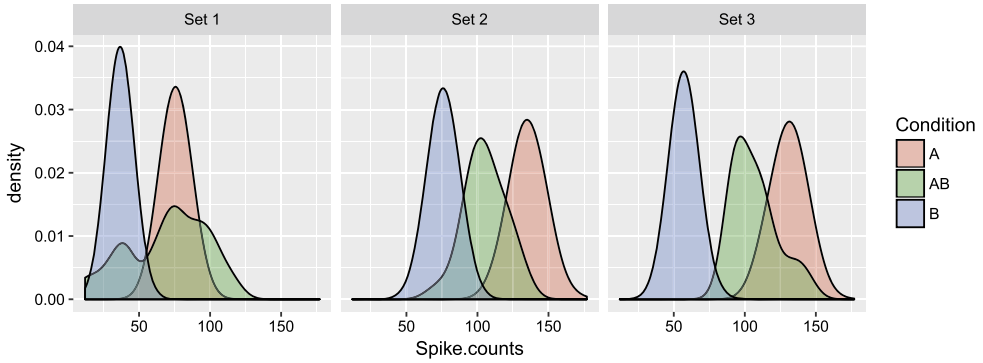

FIG. 2. *Smoothed histograms of whole-trial spike counts of the three IC experiment sets grouped by experimental condition A, B and AB. For each set the AB total spike count distribution sits between the distributions under conditions A and B. But the shape of the AB distribution varies across sets.*

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 process (Ferguson (1973)) 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, $0 \le t \le T$. 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 jth such process is given by

$$\lambda_j(t) = \alpha_j(t)\lambda_{\mathrm{A}}(t) + \{1 - \alpha_j(t)\}\lambda_{\mathrm{B}}(t), \quad t \in [0, T],$$

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

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

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

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

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

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

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 the across-trial variance of the curve at any single time point. On the other hand, for $\ell = 160\%T$, the within-trial deviation equals $0.03 \cdot \psi\sigma_0^2$, that is, 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:

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

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

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

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), for example, 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) offer 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 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

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

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 distribution 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 $\mathcal{L} = \{0.16T/N : N \in \{0.01, 0.5, 1, 2, 3, 4\}\}$. 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 $\mathcal{L} = \{40, 53.3, 80, 160, 320, 16{,}000\}$.

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

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

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 $\mathrm{DP}(\kappa G)$ with precision $\kappa > 0$ and base probability measure G 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 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, that is, $\pi_i \approx 1$ 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 of the Supplementary Material (Glynn et al. (2021)) for more details.

4.3. *An unconventional choice of the base distribution.* We deviate from common practice in choosing the base measure G equaling the law of $(\phi^*, \psi^*, \boldsymbol{\pi}^*)$ where

$$\boldsymbol{\pi}^* \sim Dir(a_1, \ldots, a_L), \qquad \psi^* \sim Be(b_1, b_2), \qquad \phi^* | \psi^* \sim N\big(0, \sigma_0^2(1 - \psi^*)\big). \tag{8}$$

This choice of G 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; see Griffin (2010), Tokdar and Martin (2019) for similar constructions. 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 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 $P_{\mathcal{L},\boldsymbol{\pi}}$. We choose $b_1 = b_2 = 1$, opting for a noninformative uniform distribution for the new draws ψ^* of the variance component of each cluster.

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, for example, 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 mth 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 mth 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 mth subinterval for the jth trial under experimental condition e, where, $m = 1, \ldots, M$, $j = 1, \ldots, n_e$, $e \in \{A, B, AB\}$. Under the approximation given by (10), our data model now looks as follows:

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

and all these random variables are independent of each other given λ_A, λ_B and α_j, $j = 1, \ldots, n_{AB}$. 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 $e \in \{A, B, AB\}$.

Notice that only the AB trial data $\mathbf{X}^{AB}$ 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 $\mathbf{X}^{AB}$ 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 $\mathbf{X}^A$ and $\mathbf{X}^B$ in a preprocessing step. We also discuss how the uncertainty in these estimates may be incorporated in the the second stage analysis of $\mathbf{X}^{AB}$.

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 $\eta_j \sim \mathrm{GP}(\phi_j, \psi_j C_{\ell_j}^{SE})$ such that $\alpha_j(t) = [1 + \exp\{-\eta_j(t)\}]^{-1}$, $t \in [0, T]$, $j = 1, \ldots, n_{AB}$. 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,

$$p\big(\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}, \kappa, \{\eta_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{AB}} | \mathbf{X}^{AB}, \lambda_A, \lambda_B\big)$$

$$\propto p\big(\kappa, \{\boldsymbol{\eta}_j, \phi_j, \psi_j, \boldsymbol{\pi}_j, \ell_j\}_{j=1}^{n_{AB}} | \mathbf{X}^{AB}, \lambda_A, \lambda_B\big)$$

$$\times \prod_{j=1}^{n_{\mathrm{AB}}} p(\eta_j|\boldsymbol{\eta}_j, \phi_j, \psi_j, \ell_j)$$

$$\times p\big(\mathbb{Q}_{\phi,\psi,\boldsymbol{\pi}}|\kappa, \{\phi_j, \psi_j, \boldsymbol{\pi}_j\}_{j=1}^{n_{\mathrm{AB}}}\big),$$

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 $X_{jm}^{\mathrm{AB}} \sim Poi(w\{\alpha_j(t_m^*) \times \lambda_{\mathrm{A}}(t_m^*) + (1 - \alpha_j(t_m^*))\lambda_{\mathrm{B}}(t_m^*)\})$ as

$$\begin{aligned}
\big(Z_{jm}^{\mathrm{A}}, Z_{jm}^{\mathrm{B}}\big) &\sim Poi\big(w\lambda_{\mathrm{A}}(t_m^*)\big) \times Poi\big(w\lambda_{\mathrm{B}}(t_m^*)\big),\\
\big(Y_{jm}^{\mathrm{A}}, Y_{jm}^{\mathrm{B}}\big) &\sim Bin\big(Z_{jm}^{\mathrm{A}}, \alpha_j(t_m^*)\big) \times Bin\big(Z_{jm}^{\mathrm{B}}, 1 - \alpha_j(t_m^*)\big),\\
X_{jm}^{\mathrm{AB}} &= Y_{jm}^{\mathrm{A}} + Y_{jm}^{\mathrm{B}},
\end{aligned}$$

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

$$(11)\quad Z \sim Poi(\mu), \qquad Y|Z \sim Bin(Z, p) \quad \Longrightarrow \quad (Y, Z - Y) \sim Poi(p\mu) \times Poi\big((1-p)\mu\big).$$

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

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

where $\mathbf{Z}^e = (Z_{jm}^e : 1 \le j \le n_{\mathrm{AB}}, 1 \le m \le M)$, $\mathbf{Y}^e = (Y_{jm}^e : 1 \le j \le n_{\mathrm{AB}}, 1 \le m \le M)$, $e \in \{\mathrm{A}, \mathrm{B}\}$. Algorithm 1 gives a schematic representation of our Markov chain sampler. All technical details are provided in Appendix A of the Supplementary Material (Glynn et al. (2021)).

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, Ventura and Cai (2003), Rao and Teh (2011), Truccolo et al. (2005)). 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 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:

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

for some a_m^e, b_m^e, $m = 1, \ldots, M$, which depend only on $\mathbf{X}^e$, that is, 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 $\lambda_e(t_m^*) \sim Ga(a_m^e + \sum_j Z_{jm}^e, b_m^e + n_{\mathrm{AB}})$

Input: Binned spike counts $\mathbf{X}^{\mathrm{AB}}$ from AB trials, and, λ_{A} and λ_{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 may be 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, Scott and Windle (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 κ 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, (ϕ_c^*, ψ_c^*) is updated by a combination of an independent proposal Metropolis–Hastings update for ψ_c^*, followed by a draw of ϕ_c^* from a normal distribution.
6. Save current parameter values as the sth sample draw.

end

Algorithm 1: Schematic description of the Markov chain sampler

independently across $e \in \{\mathrm{A}, \mathrm{B}\}$ and $m = 1, \ldots, M$. These draws could be made right after Step 1 of Section 5.2.1.

We fix the parameters a_m^e, b_m^e 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_m^e/b_m^e) and variance $(a_m^e/(b_m^e)^2)$ 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 $\Pi_e^*(\lambda_e(t_1^*), \ldots, \lambda_e(t_M^*)|\mathbf{X}^e)$, $e \in \{\mathrm{A}, \mathrm{B}\}$ 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_m^e, b_m^e 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 $(\lambda_e(t_1^*), \ldots, \lambda_e(t_M^*))$.

REMARK 6 (Computation time). We have currently implemented DAPP with the R programming language. An R package neuromplex has been published on The Comprehensive R Archive Network.[3] We ran DAPP analyses on an Apple MacBook Pro with 2.5 GHz Intel

[3] Available at https://CRAN.R-project.org/package=neuromplex.

Core i7 processor and 16 GB 1600 MHz memory. It took about 105 (230) seconds to accumulate 5000 iterations of the Markov chain sampler with $n_{\mathrm{AB}} = 20$ (50) AB trials and $M = 20$ bins. With $M = 40$ bins, the run time increased to 130 (260) seconds.

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 $(\phi_c^*, \psi_c^*, \boldsymbol{\pi}_c^*)$, $c = 1, \ldots, K$, among the collection $\{(\phi_j, \psi_j, \boldsymbol{\pi}_j) : j = 1, \ldots, n_{\mathrm{AB}}\}$.

6. Second order stochasticity in inferior colliculus.

6.1. *Data*. The neural data reported here comes from electrophysiological recordings described and analyzed in Caruso et al. (2018). Briefly, the activity of individual neurons in the inferior colliculus (IC) was recorded while two monkeys listened for sounds and made eye movements to their locations. 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, at which point 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, one at 742 Hz and another at a frequency that differed by a ratio of 1.22 or an integer power of that ratio. Each distinct double-sound experiment set was uniquely identified by the cell under recording (determined by monkey and the day of the experiment) and the frequency and the location of the second sound. Associated single sounds were derived 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 or 1000 ms since sound onset in which the sounds were on, but the monkey was maintaining fixation. All conditions were randomly interleaved.

A total of 1484 distinct double-sound sets were retained after dropping those where the cell was nonresponsive to either of the two sounds (see Caruso et al. (2018), for more details). From these, 698 sets were included for DAPP analysis after ascertaining that they had at least five trials in each of the three experimental conditions and that the distribution of the total spike counts under condition A could be statistically distinguished from that under B (log intrinsic Bayes factor of at least 3 under a Poisson distributional assumption and Jeffreys' prior on the Poisson rates). This is a larger group than the 362 sets reported in Caruso et al. (2018) who further screened out sets where either the A or the B total spike count distribution failed a Poisson goodness of fit test. We exclude this step in favor of a novel assessment of the suitability of a DAPP analysis of the AB trials, reported in Section 6.3.

6.2. *Case study with three example sets.* For illustrative purposes we begin with DAPP analysis results for the three example experiment sets referred to in Figures 1 and 2. Our primary goal is to demonstrate that DAPP is able to tease apart different modes of second order stochasticity from real data. A secondary goal is to illustrate what kind of precise quantitative inference one may draw from such analyses. The latter is explored further with a full analysis of the 698 sets in the next subsection and a comprehensive simulation study presented in the next section.

All three example sets came from monkey Y but involved three different cells as recordings were done on different days and involved different electrode insertions. Set 1 had a 500 Hz sound at 24 degrees to the right (A) and a 742 Hz sound at six degrees to the left (B). Set 2 also had 500 Hz (A) and 742 Hz (B) sounds, but their locations were at six degrees to the left and 24 degrees to the right, respectively. Signal A for Set 3 was a 903 Hz sound at 24 degrees to the left, with a 742 Hz sound at six degrees to the right being signal B. Set 1 was recorded over a 1000 ms response period, while the other two sets each had a 600 ms response horizon. Set 1 had seven A trials, 11 B trials and 18 AB trials. The corresponding counts were 21, 21 and 33 for Set 2, and, 6, 7, and 14 for Set 3. Recall that Figure 2 shows smoothed histograms of whole trial spike counts grouped by conditions A, B and AB. For each set the AB distribution appears to sit between the distributions under conditions A and B, conforming with the DAPP assumption.

Figure 3 offers a visualization of DAPP analyses of the three sets. Along with estimates of the α curves underlying the recorded AB trials, we also show 20 sample draws of the weight curve from the posterior predictive distribution. Additionally, we use 1000 such draws to compute posterior predictive distributions of the three broad features of the weight curves: (1) the range of α defined as $\text{range}(\alpha) = \max_{t\in[0,T]}\alpha(t) - \min_{t\in[0,T]}\alpha(t)$, (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.

It is apparent that the three sets exhibit different patterns of second order stochasticity. The plots of the distribution of $\text{range}(\alpha)$ indicate that, while both Sets 1 and 3 produce mostly flat α curves, Set 2 includes a good mix of wavy curves. Also, Set 1 and Set 3 are distinct from one another in the distribution of $\bar{\alpha}$. For Set 1 the weight curves have a noticeably higher concentration near the extremes of 0 and 1, while for Set 3 they concentrate in the middle with a slight tilt toward 1.

Such marginal distributions, however, fail to capture the whole picture. For example, it is unclear whether the bulge in the middle of the $\bar{\alpha}$ distribution of Set 3 is due to an increased concentration of flat weight curves around the midpoint or is driven by the wavy weight curves which, by design, have to have an $\bar{\alpha}$ near the center. To resolve such entanglements and to extract more precise quantitative summaries out of the DAPP analysis, we adopt a *labeling* convention for the weight curves and study posterior predictive distributions of the resulting labels.

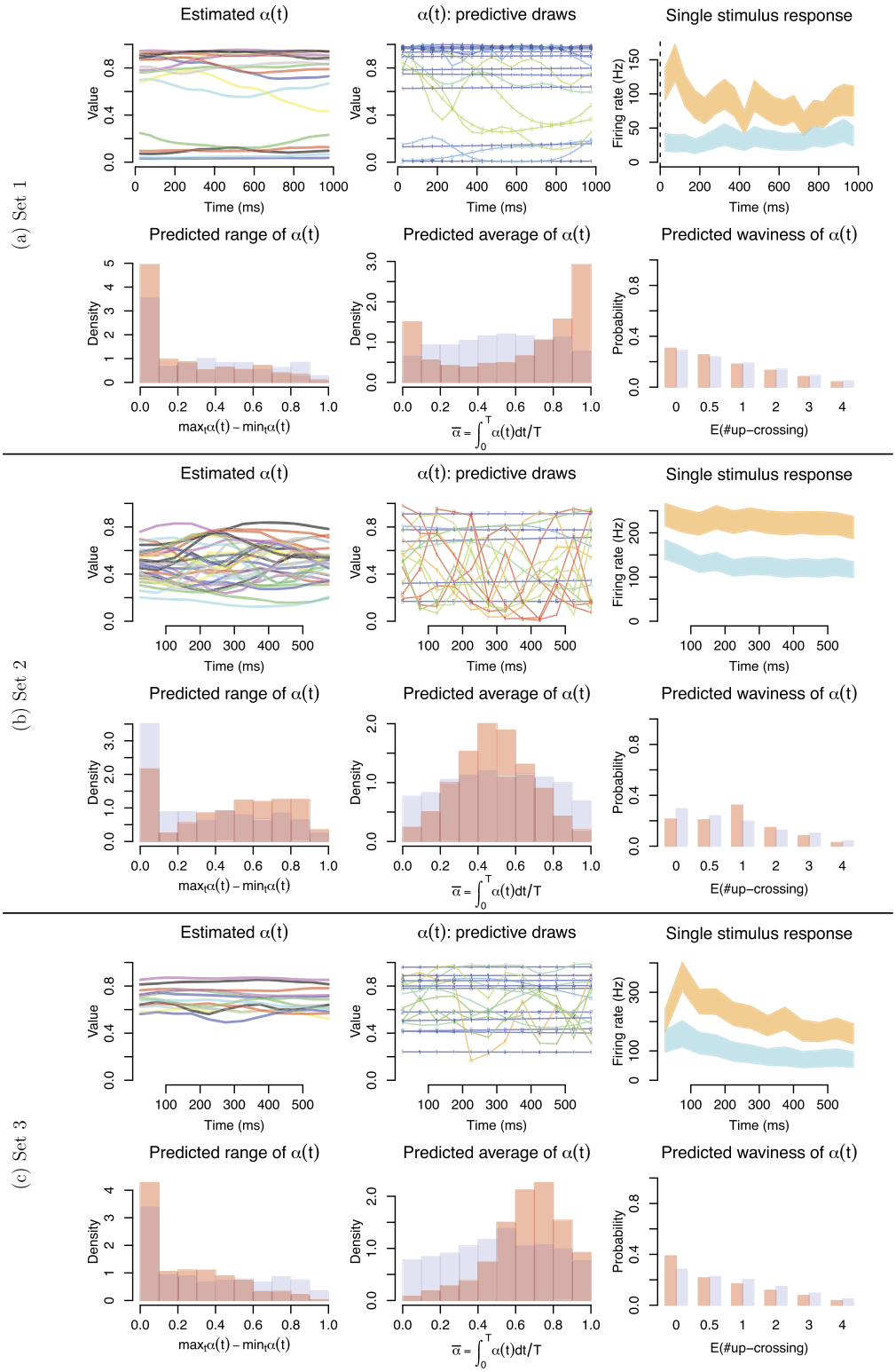

FIG. 3. *Visual summary of inference for three example IC sets. For each set, top panels show (left to right) estimated* $\alpha(t)$ *for recorded AB trials, 20 posterior predictive draws of* $\alpha(t)$ *in future hypothetical trials (red/yellow for wavy/mostly wavy; blue/green for flat/nearly flat), inference on* λ_A *(orange) and* λ_B *(cyan) depicted by 95% credible bands. Bottom panels show (left to right) marginal posterior predictive distributions of three features of the weight curves: range, long-term average and waviness (blue for prior, red for posterior).*

TABLE 1
Label distributions for three example IC sets, along with inferred second order stochastic variability types. Prior figures are shown in parentheses

Cell	flat-A	flat-B	flat-Mid	Wavy	Unlabeled	Type
1	52%(24%)	29%(24%)	12%(31%)	7%(21%)	40%(46%)	flat-A + flat-B
2	13%(27%)	13%(27%)	32%(29%)	42%(18%)	62%(51%)	flat-Mid + Wavy
3	41%(26%)	4%(26%)	51%(29%)	4%(19%)	51%(52%)	flat-A + flat-Mid

We label an α as *wavy* or *flat* based on whether range(α) is larger than 0.8 or smaller than 0.15, leaving it unlabeled otherwise. Each flat α curve is further sublabeled as *flat-A* or *flat-B* or *flat-Mid* based on whether $\bar{\alpha}$ is between $(0.75, 1)$ or $(0, 0.25)$ or $(0.25, 0.75)$. Table 1 shows posterior predictive distributions of these labels computed from 1000 draws of α. DAPP is most decisive for Set 1, being able to label 60% of the posterior predictive draws, whereas offering a more modest labeling success for the other two sets.

From Table 1 it appears that Set 1 exhibits a second order stochasticity pattern somewhat similar to random selection. Of the labeled draws, a total of 93% are flat, with 52% flat-A and 29% flat-B. That is, of the future AB trials that could be clearly labeled, more than half the time the cell would respond like it is responding only to sound A. But, in about every third of these trials, its response will be more resembling of its sound B spiking activity.

In contrast, Set 3, which also has a high propensity of producing flat α curves (96% of the labeled draws), appears to have a different pattern than random selection. With flat-Mid (51%) and flat-A (41%) being the two dominant labels, the underlying cell appears to exhibit a spiking activity with a firing rate that is either similar to λ_A or is a nondynamic weighted average of λ_A and λ_B where the weights could be evenly balanced between the two signals.

In comparison to Sets 1 and 3, Set 2 has a much higher likelihood of producing a wavy α curve. Among the labeled draws, the two dominating groups are wavy (42%) and flat-Mid (32%). The cell in Set 2 appears to have a different bimodal response pattern under the associated AB exposure: it either dynamically swings between its A and B firing patterns or holds steady at a balanced average of the two single sounds firing rates. Note that a wavy α with range(α) > 0.8 must make at least one switch between the ranges 0–10% and 90–100%, a behavior consistent with within trial random interleaving. Also, from Figure 3 the posterior distribution of the expected up-crossing count has a modest concentration around 1, indicating that about one complete switch (back and forth) is likely to occur within a single wavy AB trial.

The last column (Type) of Table 1 offers a concise summary of the label distribution by assigning each set with a tag consisting of the labels with at least 20% posterior predictive probability. We refer to this tag as the inferred *type* of second order stochasticity of each set. Set 1 is inferred to be a "flat-A + flat-B" type, underlining its random selection like second order stochastic variation. Set 2 is a "flat-Mid + wavy" type, and Set 3 is a "flat-A + flat-Mid" type. These type tags emphasize the main modes of second order stochasticity for each set, offering a simple yet meaningful summary of the DAPP inference.

6.3. *A comprehensive analysis of IC neurons.* Some caution is warranted before a full scale analysis of the IC data could be carried out with DAPP. The fundamental premise of a DAPP analysis is to quantify any second order stochastic variation in AB trials against the benchmarks set by A and B trials. But these benchmarks are set only under a Poisson model, suppressing any additional trial-to-trial variability that may have been present in the single sound data. If such variability were present but suppressed in estimating λ_A and λ_B, it is

possible the same source of variation could be mistakenly identified as causing information-encoding second order stochastic variation in the double-sound activity! Note that such trial-to-trial variation need not arise only from variable, unaccounted for stimuli present in the experimental environment. It could arise purely internally as manifest by the prevalence of burstiness in which neural spikes tend to spontaneously bunch up and space out in alternating short time intervals (Tokdar et al. (2010)). While one hopes that the DAPP analysis is performed with a bin width large enough to average out such additional local variability, it is important to safeguard the analysis against the possibility of suppressing substantial trial-to-trial variation in the single sound data.

To address this, we preprocessed each of the 698 IC sets by employing DAPP on two fake copies of each set. In one fake copy, data from AB trials was replaced with data from A trials; the other copy had B trials' data in place of the AB trials. We ran DAPP with 50 ms bin width on each copy and retrieved the type tag as outlined in Section 6.2. Under the hypothesis that single sound behavior was only first-order stochastic, at least at the scale of the chosen bin width, one would expect to retrieve a "flat-A" type for the first fake copy and "flat-B" for the second, and this should happen with high precision, that is, with only a small percentage of unlabeled draws. With a 40% cut-off on the unlabeled percentage, a total of 159 sets returned the expected type tag for each fake copy. Recall that the type codification uses only a 20% threshold on each label, making this screening process extremely sensitive. Even though many sets were screened out by this preprocessing step, we considered the remaining 159 sets to offer considerable amount of data to address the central questions: does second order stochastic variation exist in IC? If yes, then are there multiple modes through which such variation could manifest?

We then analyzed the actual data from the AB trials of these remaining 159 sets with DAPP. Figure 4 shows their inferred types. The most common inferred type (57 sets) was "flat-X + flat-Mid" where X was either A (37) or B (20). We consolidate these two types into one because the A vs. B labelling of the signals was essentially arbitrary in our experiments. The important detail is that this type is made up of flat weight curves which were either in the middle or to one extreme. The next dominating type was "flat-X" (47 sets), where X was either A (32) or B (15). Notice that this type essentially encodes a lack of second order stochasticity in AB trials. Here, the double-sound spiking activity is nearly identical to one of the two single-sound activities, in a winner-take-all manner. A third well-represented type was "flat-X + flat-Mid + flat-wavy" (25 sets) where, again, X could be either A (13) or B (12). Each of these sets included a fair amount of wavy weight curves along with flat curves that were either in the middle or toward one extreme. Ten other types were also inferred, but each was taken up by a small subgroup of the remaining 30 sets. Random selection, which is encoded by the type tag "flat-A + flat-B," was assigned to seven sets (including Set 1 presented in Section 6.2), whereas random interleaving ("wavy") was not assigned to any.

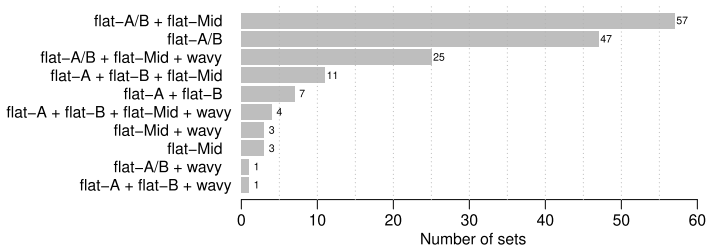

FIG. 4. *Inferred second order stochastic variation types for* 159 *IC sets. Types are consolidated to suppress the arbitrariness of the A and B labeling of the single sounds. For example, the top group consolidates "flat-A + flat-Mid" and "flat-B + flat-Mid" into a single group. Notice that all groups, except for "flat-A/B," correspond to some mode and degree of second order variability.*

From our analysis we conclude that the IC neurons exhibit a fair amount of second order stochastic variation (all types other than "flat-A" or "flat-B" and, potentially, "flat-Mid"). Additionally, the modes of this variation can be remarkably different from pure random selection or interleaving, as highlighted by the prevalence of tags that include "flat-Mid." Also, AB spiking activity seems to be biased toward one of the two signals even when it is not entirely winner-take-all. This is underlined by the presence of either "flat-A" or "flat-B," but not both, in the three dominating types.

7. Simulation study with synthetic data. Here, we report results from simulation experiments in which we assessed the accuracy of the labeling and tagging methods introduced above. Five "true" dynamic admixture types were considered:

Type 1 (60% flat-A + 40% flat-B). The cell always produces flat weight curves $\alpha(t) \equiv \alpha$, with the magnitude α drawn either uniformly from $(0.85, 0.95)$ with probability 60% or uniformly from $(0.05, 0.15)$ with probability 40%.

Type 2 (100% wavy). 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 a (also in ms) is drawn uniformly from $(0, b)$.

Type 3 (50% flat-Mid + 50% wavy). 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 $(0.45, 0.55)$. The sinusoidal curves are drawn exactly as in Type 2.

Type 4 (50% flat-B + 50% wavy). This is identical to Type 3 except that, for the flat curves, the time invariant magnitude is drawn uniformly from $(0.05, 0.15)$.

Type 5 (60% flat-A + 40% flat-Mid). This is identical to Type 1 except that the magnitudes of the flat curves are drawn, with a 60–40 split, either from $(0.85, 0.95)$ or from $(0.45, 0.55)$.

For each true type, 100 data sets were generated each with $n_A = n_B = 20$ trials for each single stimulus condition and $n_{AB} \in \{20, 50\}$ trials for the double-stimuli condition. For each trial the response horizon was taken to be $T = 1000$ ms, and the instantaneous single stimulus firing rates (in Hz) were taken to be

$$\lambda_B(t) = S \cdot 40e^{-2t/T}, \qquad \lambda_A(t) = 4\lambda_B(t) + S \cdot 40e^{-0.2t/T}, \quad t \in (0, T),$$

where $S \in \{1, 1.5\}$ was used to manipulate the overall signal strength. These choices produced average firing rates of $105 \cdot S$ Hz for the A trials and $17 \cdot S$ Hz for the B trials. We also ran experiments with $n_A = n_B = 50$, but the results were not substantially different from the results presented here with smaller single stimulus trial size and hence are omitted from further discussion.

Each synthetic dataset was analyzed by the DAPP method, with spike counts aggregated over 50 ms bins. The posterior predictive draws were then processed to generate a distribution of the four labels: flat-A, flat-B, flat-Mid and wavy, amongst those that could be labeled. We also tallied the proportion of unlabeled draws. The distribution of the four labels was then compared to the true type, and the total variation distance between the two probability vectors was calculated as a measure of estimation error. For example, if, for a data set in Type 1, the distribution of labels were 52% flat-A, 29% flat-B, 12% flat-Mid and 7% wavy, then the predicted distribution vector would be $\hat{p} = (0.52, 0.29, 0.12, 0.07)$, whereas the true distribution was $p = (0.6, 0.4, 0, 0)$, with a total-variation distance $= 0.5 \cdot \sum_{j=1}^{4} |p_j - \hat{p}_j| = 0.19$. We also tabulated whether or not the inferred second order stochasticity type matched the true type; see Appendix C in the Supplementary Material (Glynn et al. (2021)) for additional visualization.

TABLE 2
Statistical performance of DAPP measured by total variation error in estimating the label distribution (Error), rate of accurate recovery of the second order stochastic variability type (Recovery) and, lack of precision in labeling (Unlabeled)

lSet	True type	Signal	n_{AB}	Error	Recovery	Unlabeled
l1	60% flat-A + 40% flat-B	Base	20	14%	92%	34%
l			50	9	100	28
l		1.5x	20	12	97	33
l			50	9	99	26
l2	100% wavy	Base	20	9	100	44
l			50	3	100	38
l		1.5x	20	7	100	45
l			50	2	100	42
l3	50% flat-Mid + 50% wavy	Base	20	33	39	57
l			50	25	53	52
l		1.5x	20	24	65	52
l			50	13	95	46
l4	50% flat-B + 50% wavy	Base	20	19	86	43
l			50	10	100	32
l		1.5x	20	16	96	39
l			50	11	100	32
l5	60% flat-A + 40% flat-Mid	Base	20	15	83	39
l			50	10	93	36
l		1.5x	20	14	82	35
l			50	10	92	32

Table 2 gives a summary of the results. DAPP appears to have offered a fairly decent performance across the board, except for set 3 with weak signal strength. Overall, performance improved when either signal strength improved or one had more double trials. Set 3, which included a mix of flat and wavy weight curves all with similar long term averages, presented the hardest challenge to DAPP which failed to resolve the two clusters when either signal strength was weak or only a limited number of double trials were available. For this set, DAPP mostly overestimated the percentage of "wavy" curves (bias $= 19\%$ for $S = 1$ and $n_{AB} = 20$) and underestimated the same for "flat-mid" curves (bias $= -28\%$). In 57% of the replicates, the recovered type was just "wavy". DAPP's inability to disambiguate the two types was reflected in a fairly high percentage of unlabelled curves. In contrast, DAPP delivered a much more confident and accurate inference for Set 4 which has a similar mix of flat and wavy curves but with distinct long term averages.

The results for Set 5 are intriguing. While the overall estimation errors were comparable to Set 1, the recovery rates were substantially smaller. These rates failed to improve with higher signal strength, however the unlabeled percentage did go down. Interestingly, for the entirety of Set 5, every replicate that was misclassified got assigned to "flat-A." Putting together the results from Sets 3 and 5, we conclude that DAPP may suppress "flat-Mid" in favor of "flat-X" or "wavy" when signals are weak or sample sizes are small.

8. Discussion. We have introduced here 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-process 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 and codify how a cell is likely to respond in future hypothetical trials under a stimuli bundle exposure. Our analysis of monkey inferior colliculus recording with auditory signals presents substantial evidence of the existence of second order stochastic variation in the natural world and the utility of the DAPP analysis in identifying different modes of such variations. However, several comments are in order to further appreciate the development presented here.

REMARK 7 (Subjective choices). In its current form the DAPP analysis requires the user 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, an increased number of bins adds to computing cost, though the computing complexity increases only linearly in the number of bins. Under moderate signal strength, DAPP analyses results are fairly robust to the choice of the bin width. This was verified by repeating the simulation study reported in Section 7 with 25 ms bin width. However, when single stimulus firing rates are weak and/or sample sizes are small, DAPP results could be sensitive to the choice of bin width. For example, a DAPP analysis of the IC sets with a 25 ms bin width mostly resulted in similar label distributions but exhibited a mild suppression of the "flat-Mid" percentage which was redistributed to either "wavy" or "flat-X."

Recall that a similar suppression of "flat-Mid" was reported in Section 7. The two suppression phenomena are related because both stem from a common source, weak signal strength. For the IC analysis with 25 ms bins, the weakening of the signal strength is an artifact of our two stage estimation of the λ_A and λ_B curves. For a smaller bin size with only few spikes in each bin, the second stage prior in (12), which assumes conditional independence of the curve values across bins, does not offer adequate smoothing or anchoring at the single-sound estimates. Final estimates of λ_A and λ_B are noisy with large credible bands. This results in diminished separation between the two single sound benchmarks, effectively reducing the overall signal strength.

Based on our numerical experiments, we recommend a conservative choice of the bin width relative to the single stimulus firing rates and sample sizes; see also relevant discussion in Kass and Ventura (2006), Shimazaki and Shinomoto (2007). The latter paper offers a data driven choice of an optimal bin width. For this to be suitable for DAPP, one does need to reconcile possibly different choices made for the A and the B spiking activities, and additional reconciliations will be needed to fix a common choice of the bin width under which multiple sets could be analyzed. An alternative and practical rule of thumb is to choose the bin size so that, for either single stimulus, the trial-aggregated spike count in each bin exceeds a target count N, corresponding to a $100/\sqrt{N}\%$ coefficient of variation under the Poisson assumption. A small coefficient of variation ensures that the single stimulus data provides a stronger anchoring of the benchmarks λ_A and λ_B, making them less vulnerable to the weak second-stage prior. For IC data, a 50 ms bin width gave a 15% median coefficient of variation, whereas the same figure for 25 ms bins was 21%.

A more technically satisfying remedy to this problem could be obtained by replacing the product gamma prior in (12) with a smoothness inducing prior for the single stimulus firing rates. Such a choice, in general, could increase the computing cost manifold. A potential trade-off may be obtained with the conditionally autoregressive gamma priors discussed in Wolpert and Ickstadt (1998). This will be addressed in future research.

A second set of subjective choices is needed in setting the thresholds for the codification scheme introduced in Section 6.2. We intend these thresholds to be chosen by the neuroscientist, not the statistician. For example, we have used a high bar for the "wavy" label to codify the phenomenon of random, within-trial interleaving. A lower threshold could be used if the goal were to simply detect any level of within-trial, dynamic variations that may not necessarily match the notion of the cell encoding only one signal at a time.

REMARK 8 (Alternative methods). Note that one could employ much simpler tests to assess whether a given AB trial had an underlying weight curve that was flat or time-varying. For example, once the data has been binned, and, λ_A and λ_B have been estimated, one could calculate a Bayes factor for trial j between a "flat" model: $\alpha_j(t_1^*) = \cdots = \alpha_j(t_M^*) = \alpha_j \sim Unif(0,1)$ and a "wavy" model: $\alpha_j(t_m^*) \sim Unif(0,1)$, independently across $m = 1, \ldots, M$. Such an analysis suffers from three major drawbacks: (1) repeating this for all AB trials would invariably require some reconciliation that is not exactly aligned with multiplicity adjustment, unless one chooses either "flat" or "wavy" as a *null* hypothesis; (2) the "wavy" model is a weak competitor without a smoothness assumption which would be much more demanding to compute with and will require additional reconciliation across trials; (3) most importantly, such flat-vs.-wavy tests do not offer any meaningful codification of the modes of second order stochastic variation.

Toward a more structured version of such an approach, we considered a hierarchical state space model as follows. We assumed that for each AB trial with probability p_{flat} one had a flat weight curve: $\alpha_j(t_1^*) = \cdots = \alpha_j(t_M^*)$ with the common value drawn from a distribution π_{flat} on $[0,1]$. Otherwise, $(\alpha_j(t_m^*) : m = 1, \ldots, M)$ was a Markov chain on the binary state space $\{0,1\}$, governed by an initial distribution and a transition probability matrix. The latter two quantities could be parametrized by a $p_0 \in (0,1)$ giving the probability of starting at 0 and $q_{01}, q_{10} \in (0,1)$ giving the probabilities of the two possible state changes. The parameters p_{flat}, π_{flat}, p_0, q_{01} and q_{10} were assumed unknown but shared across trials.

In analyzing the IC data with this hierarchical hidden Markov model, we estimated p_{flat} to be larger than 80% for 146 of the 159 IC sets considered in Section 6.3. Estimated p_{flat} was smaller than 40% for only two sets, all of which were of the type "flat-X + flat-Mid + wavy" under DAPP. These estimates were obtained under a Bayesian analysis of the state space model, with π_{flat} modeled as $\pi_{\text{flat}} = \sum_{k=1}^{10} \pi_k Unif((k-1)/10, k/10)$, where $(\pi_1, \ldots, \pi_{10})$ was assigned a Dirichlet prior. Additionally, to avoid ambiguity between a flat α and a Markov switching α that happened to not make any switch at all, we used informative priors on q_{01}, q_{10} ensuring that the induced prior on the run length of either state, measured in milliseconds, concentrated on $(T/4, 3T/4)$ with a median at $T/2$.

These results are not surprising because the Markov model gives a precise encoding of random interleaving which has been shown by our DAPP analysis to be not well supported by the IC data. The important point to note here is that the above state space model does not allow any other form of waviness, and, hence, for each trial the evidence toward "flat" is stronger than that toward "wavy." When these evidences across trials are aggregated together through a shared probability p_{flat}, the overall evidence toward flatness becomes overwhelming. For example, if there were 2:1 evidence toward "flat" in each of 20 AB trials then, under a uniform prior, p_{flat} would be estimated to be as high as 93%!

The only way to prevent such overwhelming aggregation of evidence toward flatness is to allow for more flexible waviness structures. At the same time, it is crucial to share such structures across trials to allow borrowing of vital information that enhances the shared evidence toward waviness. DAPP achieves these twin goals by entertaining an unknown discrete probability distribution $\mathbb{P}$ that is designed to concentrate on a small number of waviness patterns selected out of a large and varied collection. In our understanding, any comparable statistical

analysis approach will have to embrace a similar level of modeling complexity. To the best of our knowledge, no such reasonable alternative currently exists in the literature.

REMARK 9 (Model deficiencies). The overarching assumption of dynamic averaging can explain only 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, that is, producing rates outside the range of single stimulus response rates, cannot be analyzed with the current version of the DAPP model. A possible generalization could be to model $\lambda_j(t) = \alpha(t)\lambda_A(t) + \beta(t)\lambda_B(t)$, where $\alpha(t)$, $\beta(t)$ are two weight curves bounded between 0 and 1 but are not restricted to satisfy $\alpha(t) + \beta(t) = 1$. While such a model could be easily computed with using our current machinery, it is not clear how to resolve the issue of nonidentifiability that arises without the restriction on the sum. A second possibility is to retain our affine combination approach but allow $\alpha(t)$ to take values outside of $(0, 1)$. This relaxation will require substantial work on the computational side as the Poisson thinning results needed for our current computation scheme no longer apply.

Additionally, 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 interspiking interval distributions observed in reality. For example, a Poisson process model is unable 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 inability 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 two ms.

A second issue with the Poisson assumption is its inability to account for *overdispersion* where the variance of the spiking activity is larger than its mean. As pointed out in Section 6.3, suppressing such overdispersion in the single stimulus data with a Poisson model could lead to unsound conclusions about second order variability under multistimuli exposure. While the fake-copy preprocessing step introduced in the same section offers some counter-measure, it may result in a substantial loss of data. Instead of screening out data with overdispersion, it will be useful to account for such additional trial-to-trial variability by either extending the DAPP model, where the Poisson assumption is replaced with a negative-binomial assumption, or by adopting other types of point process models (Kass and Ventura (2001), Ramezan, Marriott and Chenouri (2016)).

REMARK 10 (Joint analysis). In Section 6.3 we carried out separate analyses of the 159 IC sets. Some of these sets corresponded to recordings from the same neural cell, and it may be useful to consider model-based reconciliation of the corresponding analyses. This could potentially be carried out by extending the DAPP model to a multiset joint analyses framework where the core generating distributions $\tilde{\mathbb{Q}}^{(s)}$, across sets $s = 1, \ldots, S$, are jointly modeled as a hierarchical Dirichlet process (Teh et al. (2006)). A second critical issue to be looked into in the future is whether and how second order stochastic variation correlates across a neural population. While our current results offer substantial evidence toward such variation at the level of a single neuron, a complete picture of how the brain represents multiple stimuli could be obtained only by understanding how such variations are synchronized between multiple neurons recorded simultaneously. However, designing a practicable multicell DAPP model would involve significant effort, as correlation specifications must encompass all types of second order variability, spanning both within-trial and across-trial time scales.

REMARK 11 (More than two stimuli). It is relatively straightforward to generalize our approach to the case where the stimuli bundle consists of more than two stimuli. The jth

multistimuli trial can be modeled as a draw from a Poisson process with instantaneous firing rate curve: $\lambda_j(t) = \sum_{k=1}^{K} \alpha_k(t)\lambda_{A_k}(t)$, where λ_{A_k} is the firing rate curve for the kth stimulus presented alone (Condition A_k) and, $\alpha_k(t)$ are weight curves restricted to satisfy: $\alpha_1(t) + \cdots + \alpha_k(t) = 1$. The current prior specification and posterior computation easily extends to this situation, with binomial models replaced with multinomial models in Step 1 of Algorithm 1. However, what is less obvious is how one would analyze data and interpret results when various partial combinations of such stimuli are considered. Should the weight curves under ABC condition be related to weight curves under AB, BC or AC conditions? Modeling and analyzing the relationship between activities recorded under many different stimuli combinations remains a formidable challenge in both statistics and neuroscience.

These challenges notwithstanding, the DAPP analysis framework presented in this paper offers an important first step toward understanding, modeling and estimating second order stochasticity. Section 6.3 presents strong evidence of the prevalence of second order stochasticity in the primate brain. It also demonstrates 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.

Acknowledgments. We thank the Editor, the Associate Editor and two reviewers for helpful comments. Research reported in this article was supported by the National Institutes of Health under award numbers R01DC013906 and R01DC016363.

SUPPLEMENTARY MATERIAL

Supplementary file (DOI: 10.1214/20-AOAS1383SUPP; .pdf). Appendices A, B and C appear in Glynn et al. (2021).

REFERENCES

ADLER, R. J. and TAYLOR, J. E. (2009). *Random Fields and Geometry*. *Springer Monographs in Mathematics*. Springer, New York. MR2319516

CARUSO, V. C., MOHL, J. T., GLYNN, C., LEE, J., WILLETT, S. M., ZAMAN, A., EBIHARA, A. F., ESTRADA, R., FREIWALD, W. A., TOKDAR, S. T. and GROH, J. M. (2018). Single neurons may encode simultaneous stimuli by switching between activity patterns. *Nat. Commun.* **9** 2715.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FRIEDMAN, J. H. (1984). A variable span smoother. Technical report, Stanford Univ. CA Lab for Computational Statistics.

GERSTEIN, G. L. and KIANG, N. Y.-S. (1960). An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophys. J.* **1** 15.

GLYNN, C., TOKDAR, S., ZAMAN, A., CARUSO, V., MOHL, J., WILLETT, S. and GROH, J. (2021). Supplement to "Analyzing second order stochasticity of neural spiking under stimuli-bundle exposure." https://doi.org/10.1214/20-AOAS1383SUPP

GRIFFIN, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Anal.* **5** 45–64. MR2596435 https://doi.org/10.1214/10-BA502

KASS, R. E. and VENTURA, V. (2001). A spike-train probability model. *Neural Comput.* **13** 1713–1720.

KASS, R. E. and VENTURA, V. (2006). Spike count correlation increases with length of time interval in the presence of trial-to-trial variation. *Neural Comput.* **18** 2583–2591.

KASS, R. E., VENTURA, V. and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J. Neurophysiol.* **94** 8–25.

KASS, R. E., VENTURA, V. and CAI, C. (2003). Statistical smoothing of neuronal data. *Netw. Comput. Neural Syst.* **14** 5–16.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 https://doi.org/10.1080/01621459.2013.829001

RAMEZAN, R., MARRIOTT, P. and CHENOURI, S. (2016). Skellam process with resetting: A neural spike train model. *Stat. Med.* **35** 5717–5729. MR3580935 https://doi.org/10.1002/sim.7127

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

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHIMAZAKI, H. and SHINOMOTO, S. (2007). A method for selecting the bin size of a time histogram. *Neural Comput.* **19** 1503–1527. MR2316960 https://doi.org/10.1162/neco.2007.19.6.1503

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

TOKDAR, S. T. and MARTIN, R. G. (2019). Bayesian test of normality versus a Dirichlet process mixture alternative. *Sankhya, Ser. B*. To appear.

TOKDAR, S., XI, P., KELLY, R. C. and KASS, R. E. (2010). Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *J. Comput. Neurosci.* **29** 203–212.

TRUCCOLO, W., EDEN, U. T., FELLOWS, M. R., DONOGHUE, J. P. and BROWN, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* **93** 1074–1089.

VENTURA, V., CAI, C. and KASS, R. E. (2005). Trial-to-trial variability and its effect on time-varying dependency between two neurons. *J. Neurophysiol.* **94** 2928–2939.

WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267. MR1649114 https://doi.org/10.1093/biomet/85.2.251