Continuous partitioning of neuronal variability

Anuththara Rupasinghe[*1], Adam S. Charles[2], and Jonathan W. Pillow[*1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, 08544
[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

Abstract

Neural variability poses a major challenge to understanding the information content of neural codes. Recent work by Goris et al. [1] has provided new insights into variability in the visual pathway by partitioning it into two components: a stimulus component reflecting sensory tuning, and a modulatory component arising from stimulus-independent fluctuations in excitability. However, a key limitation of this framework is that it lacks a continuous-time interpretation; it applies only to spike counts measured in time bins of a given size, and does not generalize to different bin sizes. Here we overcome this limitation using a new model for continuous-time partitioning of neural variability, the "continuous modulated Poisson" (CMP). This model extends the partitioning model of Goris et al. [1] by modeling a neuron's instantaneous firing rate as the product of a time-varying stimulus drive and a continuous-time stochastic gain process. We apply this model to spike responses from four different visual areas (LGN, V1, V2, and MT) and show that it accurately captures spike train variability across timescales and throughout the visual hierarchy. Moreover, we demonstrate that the modulatory gain process decays according to an exponentiated power law, with higher variance and slower decay at later stages of the visual hierarchy. This model provides new insights into the organization of both stimulus-driven and stimulus-independent modulatory processes and provides a powerful framework for characterizing neural variability across brain regions.

Keywords: neural coding; variability; spike train statistics; Gaussian processes; visual cortex

Introduction

Neural responses exhibit trial-to-trial variability, even in response to repeated presentations of the same stimulus [1–5]. Several hypotheses have been proposed for the source of this variability, including noise in synaptic transmission [6], fluctuations in balanced excitation and inhibition [7], noisy peripheral

[*]Corresponding authors: A. R. (ar0621@princeton.edu) and J. W. P. (pillow@princeton.edu).

sensors [8], recurrent dynamics [9], neural adaptation [10] and synaptic plasticity [11]. To assess the impacts of variability on neural coding, researchers have developed a variety of statistical models that seek to capture both the stimulus dependence and stochasticity of neural spike responses [12–17]. The simplest such model for spike train data is the Poisson process [18, 19], which assumes that spike count variability in disjoint time intervals is independent. Under this model, the variance of the binned spike counts is equal to the mean for any choice of time bin. However, a large literature has shown that this relationship does not hold precisely in many brain areas. Rather, spike count variance commonly exceeds the mean, a phenomenon known as "overdispersion" [1, 7, 20–30].

To account for super-Poisson response variability in the visual pathway, Goris et al [1] introduced the "modulated Poisson" model, which partitions neural variability into a stimulus-driven component and a modulatory gain component. Specifically, the model assumes that spike counts are Poisson distributed, with a mean determined by the product of two terms: a stimulus-related component that reflects the neuron's stimulus tuning, and a stimulus-independent stochastic gain that varies randomly across trials (Fig. 1A). Goris et al [1] showed that this model provides an accurate account of spike count variability across the early visual pathway, with the degree of overdispersion linked to the strength of the modulatory component.

However, an important drawback of the modulated Poisson model is that it applies to binned spike counts but not to spike trains. It describes the variability of spike counts measured in discrete time bins, and thus cannot characterize variability on shorter timescales. Moreover, it does not account for how overdispersion depends on the size of the time bins used to count spikes.

To address this limitation, we begin by formulating a continuous-time version of the modulated Poisson model introduced by Goris et al. [1]. First, we replace the (scalar) stimulus component with a stimulus-dependent time-varying Poisson firing rate. Then, we specify a sample rate for the stochastic gain variable; we assume that the gain is constant within each discrete sample interval and is drawn independently for each interval. For example, we might assume that the sample interval for the modulatory gain is equal to the stimulus duration (Fig. 1B). In this case, the gain is constant for each trial and independent across trials. This produces a gain "process" with a flat temporal auto-correlation, and a Fano factor (variance-to-mean ratio) that grows linearly with the size of the window used to count spikes (Fig. 1B, right). Alternately, we could assume that the sample interval is much shorter than the length of a single trial (Fig. 1C). In this case, the auto-correlation of the gain process is narrow, and the Fano factor curve remains flat for all bin sizes larger than the sample interval (Fig. 1C, right).

These continuous interpretations of the Goris model rely on the assumption that the modulatory gain is constant during each sample interval and then jumps discontinuously between intervals, which seems unrealistic for biological systems. Moreover, such models have limited ability to capture complex relationships between time bin size and Fano factor, given that they can only produce curves that grow linearly and then saturate (Fig. 1B,C).

To overcome these limitations, we introduce the continuous modulated Poisson (CMP) model, a flexible extension of the Goris model to continuous-time spike responses. The CMP model describes spike trains as arising from a Poisson process with a rate given by the product of two continuous-time processes: (1) a "stimulus-driven" component, which describes the firing rate fluctuations induced by the stimulus; and (2) a modulatory "gain" component, which captures stimulus-independent trial-to-trial variability using a stochastic process (Fig. 1D). Both components are modeled as exponentiated Gaus-

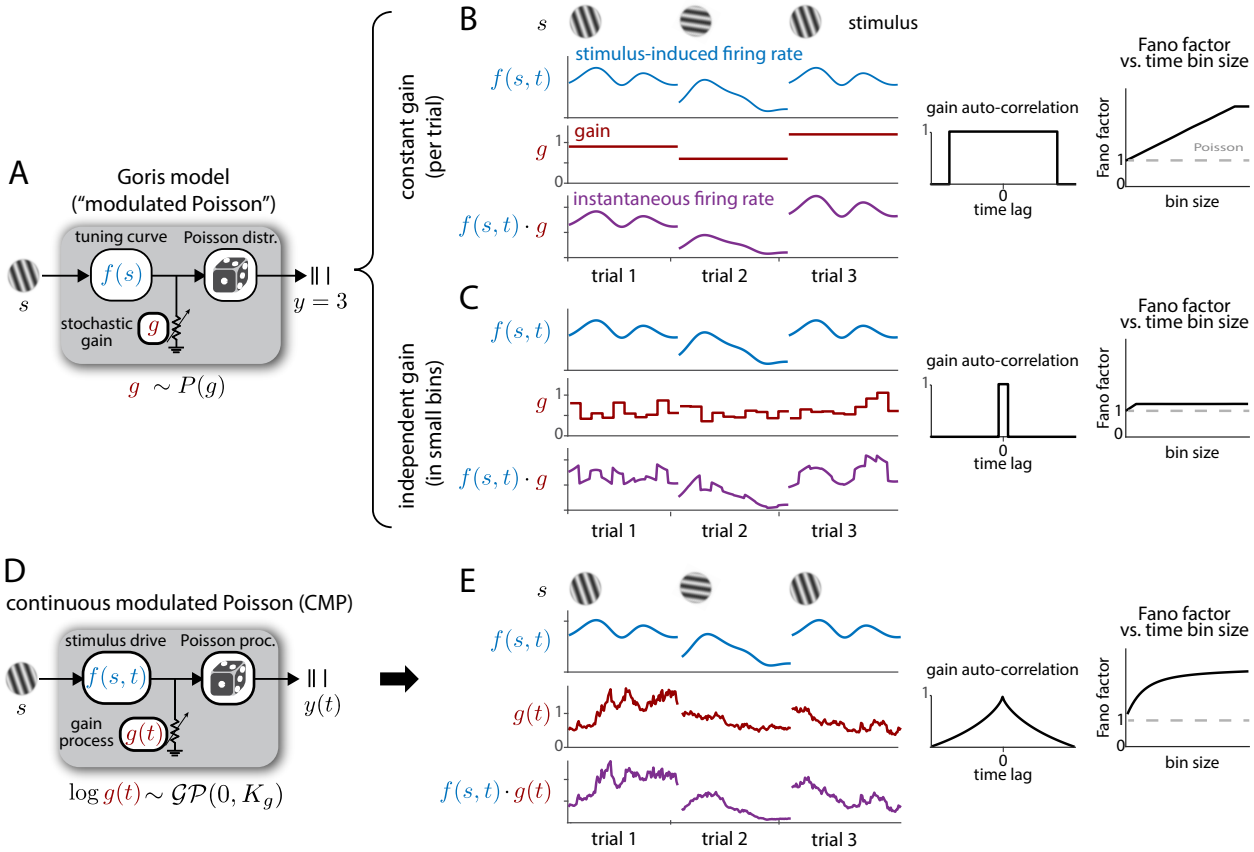

Figure 1: (A) Schematic of the Goris modulated Poisson model [1], which assumes that spike counts (y) are generated by a Poisson process whose rate is the product of two components: a stimulus-dependent drive $f(s)$ and a stimulus-independent gain g. The gain g is a scalar and is independently drawn across trials. (B, C) Two continuous-time extensions of the Goris model. (B) Constant gain model: assumes that the sampling interval of the gain equals the trial duration. Left: stimulus-induced rate, gain process, and instantaneous firing rate for three example trials. Middle: gain autocorrelation function. Right: Fano factor as a function of bin size (dashed line indicates the baseline Poisson model). (C) Independent gain model: assumes that the sampling interval of the gain is much shorter than the length of a single trial. Panels are organized as in (B). (D) Schematic of the proposed Continuous Modulated Poisson (CMP) model. Here, spike trains $y(t)$ are generated by a Poisson process whose rate is the product of a time-varying stimulus drive $f(s,t)$ and a continuous-time stochastic gain process $g(t)$. The gain process $g(t)$ follows a Gaussian process (GP) with temporal correlations governed by an exponentiated power law kernel, and is independently drawn across trials. (E) Panels are organized as in (B), for the CMP model.

sian processes (GPs), providing a flexible framework for capturing smooth, continuous dynamics with tunable temporal structure [31–33].

To capture the temporal correlations in the modulatory gain process, we introduce a novel covariance function that allows these correlations to decay more slowly than those governed by the standard radial

basis (or "Gaussian") kernel. We refer to this as the Exponentiated Power Law (EPL) kernel, as it models autocorrelation decay according to an exponentiated power law with exponent $q \leq 2$ (Fig. 1E, middle). This added flexibility enables the model to more accurately capture nonlinear relationships between bin size and Fano factor (Fig. 1E, right) compared to previous approaches.

We applied the CMP model to spike train recordings from four visual areas (LGN, V1, V2, and MT) [1, 34] and uncovered novel insights into variability across the visual hierarchy. In particular, we observed a systematic increase in the magnitude of gain fluctuations with progression along the visual hierarchy, consistent with prior observations [1]. Furthermore, comparisons of the inferred EPL kernels across visual areas revealed that gain autocorrelation exhibits slower temporal decay at higher levels of the visual hierarchy. A comparison of the variance of the stimulus drive with the variance of the modulatory gain process across neurons further revealed that these two components tend to be negatively correlated in all visual areas. These findings support existing hypotheses while providing new insight into variability structure and demonstrate that the CMP model outperforms existing methods in capturing time-resolved overdispersion. Altogether, CMP offers a principled and flexible framework for partitioning neural variability in continuous time across stimulus conditions and brain areas.

Results

Continuous modulated Poisson (CMP) model

Our proposed model describes spike trains using a doubly stochastic point process model, in which the firing rate of a Poisson process is itself a stochastic process. It describes spike train $y(t)$ elicited in response to a stimulus s using a Poisson process:

$$y(t) \mid \lambda(t) \sim \mathrm{PoissProc}\big(\lambda(t)\big), \tag{1}$$

where the firing rate or "conditional intensity" $\lambda(t)$ is a random variable given by the product of two terms:

$$\lambda(t) = f(s,t) \cdot g(t), \tag{2}$$

where $f(s,t)$ is a deterministic stimulus-dependent term that captures stimulus-locked fluctuations in firing rate, and $g(t)$ is a stochastic gain process that captures stimulus-independent fluctuations in a neuron's excitability (Fig. 1D-E).

We assume that stimulus component $f(s,t)$ varies independently across stimuli and is smooth in time, which we express using a Gaussian process prior:

$$\log f(s,t) \sim \mathcal{GP}(0, K_f), \tag{3}$$

where K_f is the standard radial basis function (RBF) covariance function, which describes the covariance of fluctuations in log firing rate at arbitrary points in time [31]. Specifically, the covariance between the log firing rate at any two time points t_1 and t_2 during stimulus presentation interval $[0, T]$ is given by:

$$K_f(t_1, t_2) = \rho_f \exp\left(-\frac{(t_1 - t_2)^2}{2\ell_f^2}\right), \tag{4}$$

where ρ_f and ℓ_f are hyperparameters controlling variance and length-scale (or smoothness) of the fluctuations in log firing rate, respectively.

Similarly, we model the gain process $g(t)$ as arising from an exponentiated Gaussian process:

$$\log g(t) \sim \mathcal{GP}(0, K_g), \tag{5}$$

where K_g is the covariance function that characterizes temporal correlations in the gain process. Unlike the stimulus component, the gain process is a latent variable that varies independently across stimulus presentations, contributing to greater-than-Poisson variability of the spike response. Thus, if we have N repeated presentations of a fixed stimulus $s = s^*$, we will have a common stimulus component $f(s^*, t)$ and an independent gain component $g_i(t)$ for each repeat $i \in \{1, \ldots, N\}$.

To accurately capture the stochasticity of real neural spike trains, we introduce a novel covariance function for the gain process, which we refer to as the exponentiated power law (EPL) kernel:

$$K_g(t_1, t_2) = \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_1 - t_2}{\ell_g}\right|^q\right), \tag{6}$$

where ρ_g denotes the marginal variance, ℓ_g is the length scale, and q is the power law exponent, which determines the decay of correlations at long timescales. We prove in the SI Appendix that the proposed EPL kernel is a valid covariance function for $0 < q \leq 2$.

Fig. 2 shows example gain covariance functions given by different settings of the hyperparameters ℓ_g (lengthscale) and q (power-law exponent). Note that when $q = 2$ (right column), we recover the RBF covariance function given above (eq. (4)). For $q < 2$, however, the gain process correlations decay more slowly. Note additionally that in the limit of $\ell_g \to \infty$ or $q \to 0$, we obtain the continuous-time Goris model shown in Fig. 1B, where the gain is constant for the entirety of each stimulus presentation interval $[0, T]$.

Partitioning variability with CMP

Under the CMP model, spike train variability can be partitioned into a Poisson component, which is equal to the mean, and a modulatory component, which controls the degree of overdispersion. The continuous-time formulation of the CMP model allows us to compute this partitioning for arbitrary time intervals, and thus to analytically characterize how overdispersion grows with time bin size. In the case of a non-fluctuating stimulus component $f(s, t) = \bar{f}$, the mean spike count in a time bin of size Δ is given by:

$$\text{mean:} \qquad \mathbb{E}[y_\Delta \mid s] = \Delta\left(\bar{f} e^{\frac{\rho_g}{2}}\right) := \Delta\bar{\lambda}, \tag{7}$$

where $\bar{\lambda}$ denotes the mean firing rate under the model. The spike count variance is then given by:

$$\text{variance:} \qquad \mathrm{var}[y_\Delta \mid s] = \Delta\bar{\lambda} + (\Delta\bar{\lambda})^2(\kappa(\Delta) - 1), \tag{8}$$

where

$$\kappa(\Delta) = \frac{1}{\Delta^2}\int_0^\Delta \int_0^\Delta e^{K_g(t_1,t_2)}\, dt_1\, dt_2, \tag{9}$$

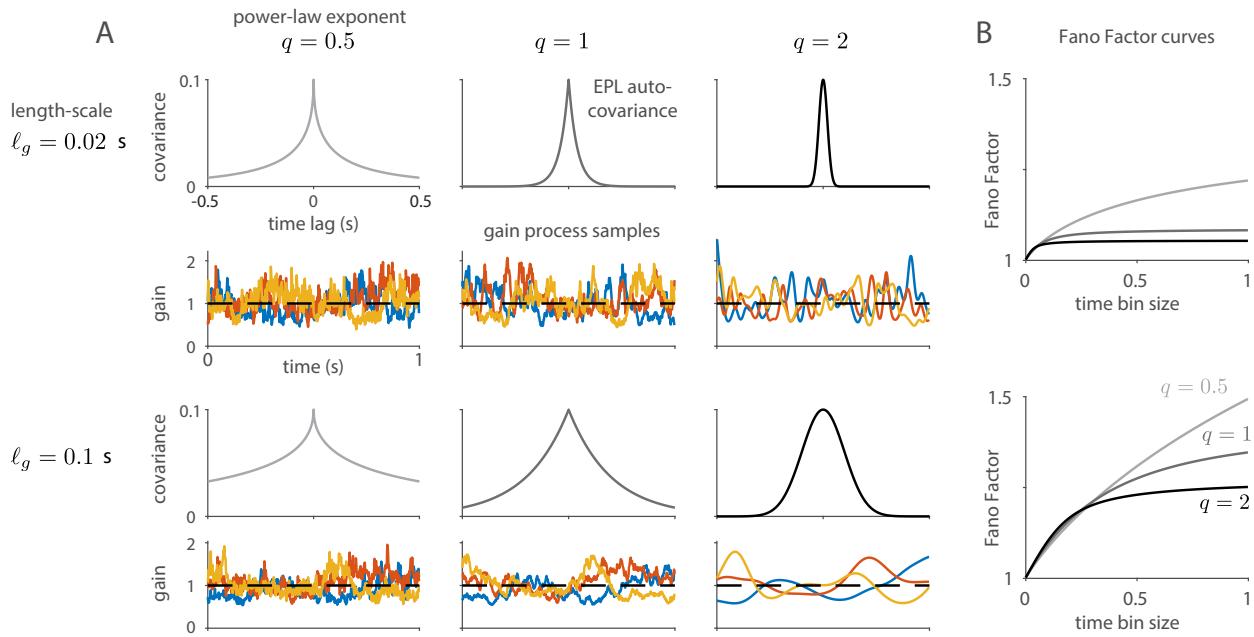

Figure 2: Illustration of different settings of the proposed Exponentiated Power Law (EPL) covariance function with marginal variance $\rho_g = 0.1$. **(A)** EPL auto-covariance functions with different settings of length-scale (top row: $\ell_g = 0.02$, bottom row: $\ell_g = 0.1$) and power-law exponent (columns left to right: $q = 0.5$, $q = 1$, $q = 2$). Colored traces below each auto-covariance show three sample gain processes $g(t)$, obtained by sampling from the corresponding GP and exponentiating. **(B)** Fano Factor curves showing how Fano Factor changes as a function of the bin size used to count spikes, for EPL covariances in the top and bottom rows in (A), respectively.

is the average of the exponentiated covariance function $\exp(K_g)$ over all pairs of times in the interval $[0, \Delta]$. (Note: this formula holds for any covariance function, not just the EPL covariance function we use for K_g). From these two terms, the Fano factor as a function of bin size can be computed as the ratio of variance to mean:

$$\text{Fano factor:} \qquad FF(\Delta) = 1 + \Delta\bar{\lambda}(\kappa(\Delta) - 1). \tag{10}$$

This shows that the Fano Factor approaches 1 in the limit of small bin size Δ, low firing rate $\bar{\lambda}$, or when $\kappa(\Delta)$ approaches 1, which occurs when the variance of the gain process ρ_g goes to zero. Note that $\kappa(\Delta)$ is a decreasing function of Δ, reflecting the decay of correlations in the gain process—it is constant only for a gain process with infinite lengthscale (e.g., Fig. 1B). Thus, the Fano factor of the CMP model grows at most linearly with time bin size. Corresponding expressions for the general case, where $f(s, t)$ is a fluctuating stimulus component, are provided in the Methods (see SI Appendix for derivations).

Model fitting

To fit the CMP model to data, we need to compute the probability over spike counts in finite-sized time bins, given that spike trains are generally measured at a finite sampling rate. Computing this probability involves integrating a conditionally Poisson spike count probability over all realizations of the gain process $g(t)$:

$$P(y_t \mid s) = \int \mathrm{Poiss}\big(y_t \mid \Lambda_t\big) d\mu\big(g(t)\big)\,, \tag{11}$$

where $\Lambda_t = \int_t^{t+\Delta} f(s,\tau) \cdot g(\tau)\, d\tau$ is the expected spike count in the bin $[t, t+\Delta]$ given the stimulus s and gain process $g(t)$, and $\mu\big(g(t)\big)$ is the probabilistic measure of the gain process under the Gaussian process prior (eq. (5)). However, this (infinite-dimensional) integral cannot be computed in closed form, since it involves the product of Poisson and Gaussian distributions. Thus, we introduce an efficient variational inference procedure [35], which allows us to optimize the hyperparameters governing the GP covariance functions and infer the stimulus components $f(s,t)$ for each stimulus (see Methods). The fitting method also provides a log-Gaussian approximation to the posterior over the gain process $g(t)$ for each trial, allowing us to visualize the most likely modulatory stochastic gain trajectory for each repeat of a given stimulus.

To validate our inference procedure, we applied it to a synthetic dataset generated from the CMP model. Fig. S1A shows true and inferred stimulus-drive and gain processes for two repeats of each of two different stimuli, while Fig. S1B compares the true and inferred GP hyperparameters. These results demonstrate that our proposed inference procedure can accurately recover ground-truth hyperparameters, tuning functions, and gain processes on a grid of time bins. Note that the model itself is defined in continuous time and can therefore be consistently applied to data measured at arbitrary temporal resolutions.

Continuous partitioning of variability in the macaque V1

We applied our method to partition response variability in five populations of neurons recorded from the superficial layers of the macaque primary visual cortex (data from [34]). These datasets were collected from three monkeys using drifting grating stimuli presented at $J = 72$ equally spaced orientations. Each grating was displayed for 1280 ms, followed by a 1280 ms blank period, resulting in a total trial duration of 2.56 s. Each grating orientation was repeated $N = 50$ times. In total, we analyzed spiking activity from 654 isolated units (147, 113, 113, 133, and 148 neurons across the five datasets).

To flexibly capture different tuning profiles, we assumed that each stimulus condition has its own set of hyperparameters for the stimulus-driven component (i.e., for $j = 1, \ldots, J$ stimuli, $\rho_f = \rho_f^{(j)}$ and $\ell_f = \ell_f^{(j)}$). In contrast, because the gain process is stimulus-independent, we assumed that the gain hyperparameters are shared across all stimulus conditions. This formulation provides a unified and parsimonious framework for modeling trial-to-trial variability across different stimuli.

Fig. 3A shows the inferred time-domain activity of a representative neuron from dataset 1, using a temporal grid resolution of $dt = 0.02$ s, for three different drift orientations (0^0, 105^0, and 280^0). The rows (top to bottom) show the PSTH derived from the data, the estimated stimulus drive, and–for two different trials per orientation–the inferred gain processes, firing rates, and observed spike rasters.

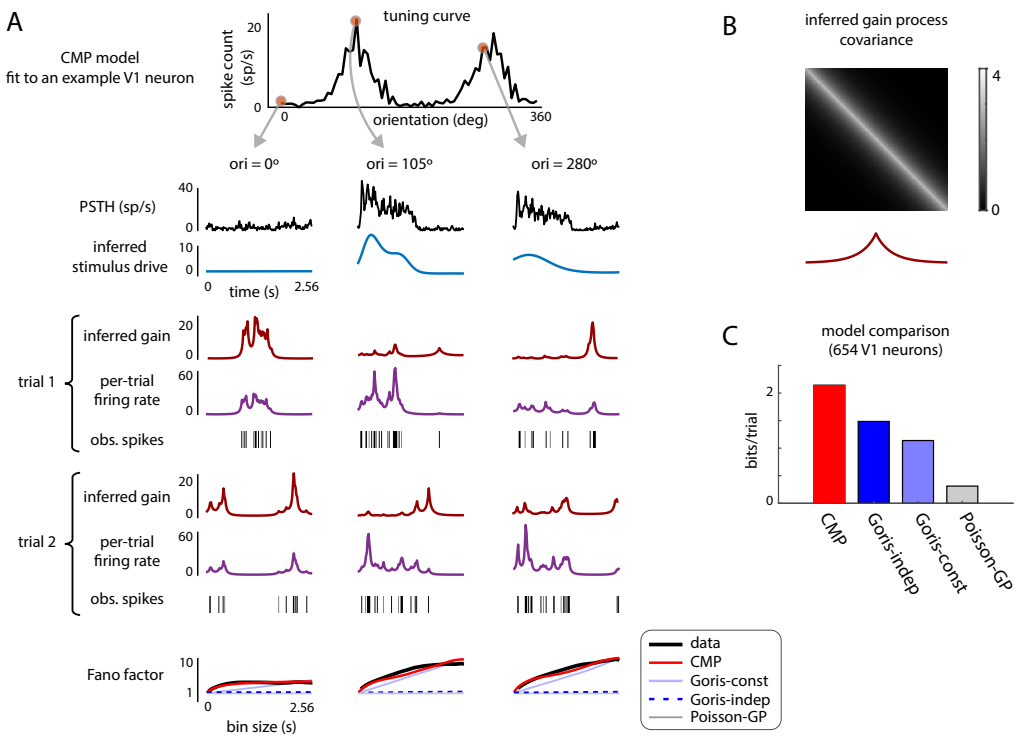

Figure 3: Performance comparisons of different models on neural population data recorded from macaque primary visual cortex under 72 drifting sinusoidal gratings (data from [34]). **(A)** Top to bottom: Tuning curve of a sample neuron; for three selected orientations (0^0,105^0,280^0), the PSTH, inferred stimulus drive, inferred gain (trial 1), inferred firing rate (trial 1), observed spike raster (trial 1), inferred gain (trial 2), inferred firing rate (trial 2), observed spike raster (trial 2), and Fano factor vs. bin size curves for real data (black), CMP model (red), Goris-independent gain model (dark blue dashed), Goris-constant gain model (light blue), and Poisson-GP model (gray). **(B)** EPL gain covariance (top) and autocorrelation (bottom) under the CMP model for the neuron in panel A. **(C)** Log-likelihood improvement of each model relative to the baseline Poisson model, computed on held-out data across all 654 V1 neurons and 72 stimulus orientations.

While the estimated stimulus drive closely matches the PSTH, the inferred gain processes capture trial-specific temporal fluctuations that are clearly visible in the spike rasters. Notably, when the neuron is strongly driven by the stimulus (middle column), the instantaneous per-trial firing rate is dominated by the stimulus drive. In contrast, when the neuron is weakly driven (left column), the stochastic gain process plays a larger role in shaping the per-trial firing rate. These observations indicate that the CMP model meaningfully partitions response variability across stimulus conditions. Fig. 3B shows the estimated EPL gain covariance, corresponding to optimal hyperparameters $\rho_g = 2.9$, $\ell_g = 0.93$, and $q = 0.95$.

The continuous-time formulation of the CMP model allows variability to be partitioned at arbitrary temporal resolutions, enabling analytical characterization of the relationship between overdispersion and bin size (see Materials and Methods for details). The bottom row of Fig. 3A shows Fano factor curves

derived from the CMP model (red), compared with curves computed from the data (black), the Goris model with independent gain (dark blue dashed), the Goris model with constant gain (light blue), and the Poisson-GP model (gray). The two variants of the Goris model are continuous-time extensions of the original modulated Poisson model[1]: the independent gain model assumes that the sampling interval of the gain process is much shorter than the trial duration, whereas the constant gain model assumes that the gain remains fixed throughout the entire trial. The Poisson-GP model corresponds to the CMP model without a gain component—that is, it includes only a stimulus drive (see Materials and Methods for details of all model variants). These results show that the CMP model closely matches the overdispersion observed in real data across orientations and bin sizes, outperforming all other methods. While both Goris model variants produce overdispersion [1], the independent gain model yields Fano factor curves that are largely flat, and the constant gain model produces curves that increase linearly with bin size—neither of which captures the true empirical relationship.

We next compared model performance using the average log-likelihood improvement relative to the baseline Poisson model on held-out data from all 654 V1 neurons across all 72 orientations (Fig. 3C). The CMP model achieved the highest predictive performance across all neurons. In Fig. S2, we further compared the CMP model to three variants with different gain kernels: CMP-RBF (where the EPL kernel in eq. (6) is replaced by the standard RBF kernel [31]), CMP-Matérn 0, and CMP-Matérn 1 (with Matérn kernels of order 0 and 1, respectively). These results demonstrate that the EPL kernel provides the best fit to the gain structure in the data.

To evaluate the utility of sharing gain hyperparameters across stimuli, we compared the CMP model with a variant in which the gain hyperparameters $\{\rho_g, \ell_g, q\}$ are fit independently for each of the $j = 1, \ldots, J$ stimulus conditions, denoted by $\{\rho_g^{(j)}, \ell_g^{(j)}, q^{(j)}\}$. Fig. S3A shows the inferred activity under both models, and Fig. S3B presents the corresponding log-likelihoods. The performance of the two models was nearly identical, suggesting that the CMP model can robustly capture trial-to-trial variability across stimuli using a single shared set of gain hyperparameters.

Finally, we tested the consistency of CMP inference under different temporal resolutions by comparing results across three grid resolutions: $dt = 0.02$ s, $dt = 0.05$ s, and $dt = 0.1$ s. Fig. S4A shows that the estimated gain hyperparameters are largely invariant to grid resolution. Fig. S4B presents log-likelihood comparisons across models and resolutions, confirming that CMP consistently outperforms other models regardless of the grid resolution used for inference.

Continuous partitioning of variability along the visual hierarchy

Next, we evaluated model performance across different areas of the visual hierarchy by applying each method to neural responses recorded in the lateral geniculate nucleus (LGN), V1, V2, and MT (data from [34]). The recordings were obtained using drifting sinusoidal gratings of the preferred size and speed, varying either in $J = 12$ spatial frequency (ranging from 0 to 10 cycles/deg) or in $J = 16$ equally spaced drift direction. Each grating was presented for one second and repeated at least $N = 5$ times. We analyzed spiking activity from 48, 376, 167, and 127 neurons from LGN, V1, V2, and MT, respectively, using a grid resolution of $dt = 0.02$ s.

Fig. 4A shows the average log-likelihood improvement on held-out data, relative to the baseline Poisson model, for each method across all four areas. The CMP model consistently outperformed all other

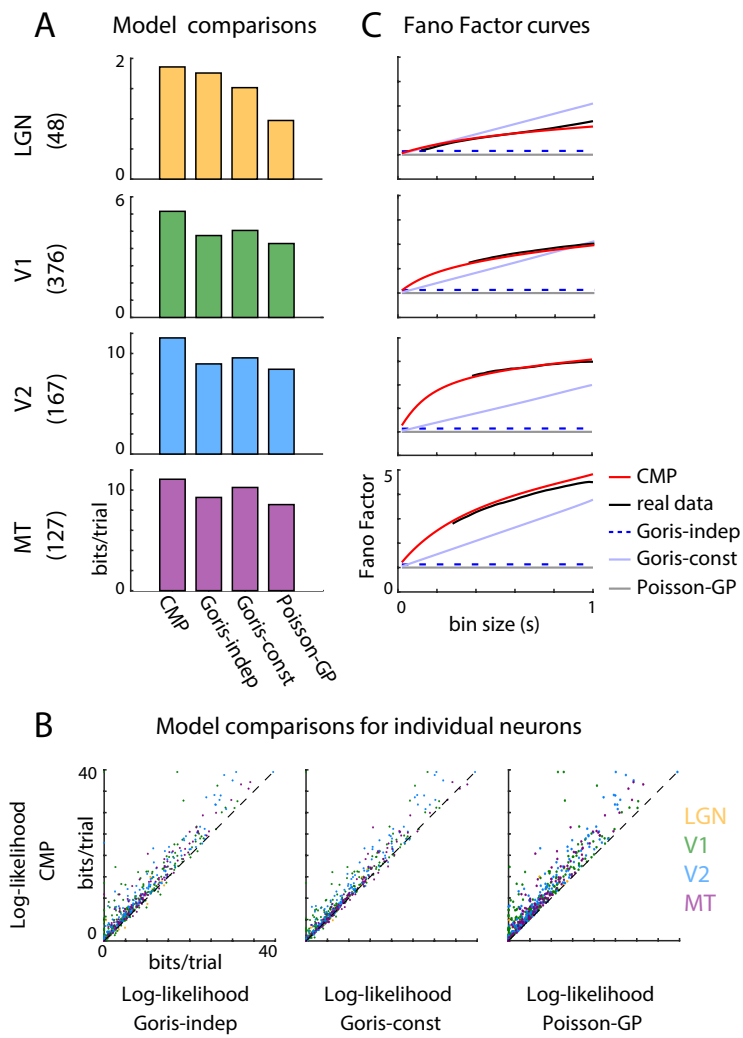

Figure 4: Performance comparisons of different models on neuronal population data recorded from different areas along the visual hierarchy—the lateral geniculate nucleus (LGN), V1, V2 and MT—under drifting sinusoidal gratings of preferred size and speed, varying in spatial frequency (12 frequencies from 0 to 10 cycles/deg) or in drift direction (16 directions) (data from [1]). **(A)** Log-likelihood improvement of each model relative to the baseline Poisson model, averaged across neurons in each population. Rows from top to bottom: LGN, V1, V2, MT. The number of neurons per area is indicated in parentheses. **(B)** Log-likelihood comparison of the CMP model versus the Goris independent gain model (left), Goris constant gain model (middle), and Poisson-GP model (right) for individual neurons. Colors indicate cortical areas: LGN (yellow), V1 (green), V2 (blue), MT (purple). **(C)** Fano factor vs. bin size curves averaged across each population. Black: real data; red: CMP; dark blue dashed: Goris-independent gain; light blue: Goris-constant gain; gray: Poisson-GP.

models across the visual hierarchy. To examine performance at the single-neuron level, Fig. 4B presents scatter plots comparing the log-likelihoods of the CMP model with those of the Poisson-GP model, the Goris independent gain model, and the Goris constant gain model. These results show that the CMP model systematically achieves better fits across individual neurons in all regions.

Next, we characterized trial-to-trial variability by analyzing the mean–variance relationship as a function of bin size. Fig. 4C shows Fano factor curves, averaged over neurons in each population, for real data and the various models. Notably, the CMP model closely follows the overdispersion patterns observed in real data across visual areas and timescales, again outperforming alternative methods. In Fig. S5, we extend this analysis by including three additional CMP variants: CMP-RBF, CMP-Matérn 0, and CMP-Matérn 1. These comparisons further confirm that the standard CMP model provides superior performance across all visual areas.

Finally, to examine the importance of including a GP prior on the stimulus drive, we compared the performance of the Goris model with and without this component. (Note that the original modulated Poisson model in [1] does not include a prior on the stimulus drive; thus, the version without the prior is more directly comparable to their formulation.) Fig. S6 presents the log-likelihood comparisons from these analyses, which clearly show that incorporating a smoothness prior on the stimulus drive generally improves the model's ability to fit held-out data. It is also noteworthy that even the Poisson-GP model—which lacks a stochastic gain component—outperforms both variants of the Goris model when the GP prior on the stimulus drive is removed (Fig. S6A–C). These results highlight that the smoothness prior we place on the stimulus drive is crucial for accurately capturing neural variability (Fig. S6D).

Variations in gain and stimulus-drive along the visual hierarchy

Finally, we compared the distributions of the inferred CMP hyperparameters across the visual areas LGN, V1, V2, and MT, as shown in Fig. 5. Panels A, B, and C show the distributions of the inferred gain power-law exponent q, gain length scale ℓ_g, and gain variance ρ_g, respectively, for each neuronal population. Fig. 5D presents the median gain covariance and its corresponding autocorrelation function for each area.

The gain power-law exponent (Fig. 5A) increases along the visual hierarchy: from LGN to V1 ($p < 0.001$, sign test) and from V1 to V2 ($p < 0.001$), although there is a slight decrease from V2 to MT ($p < 0.05$). The gain length scale (Fig. 5B) also shows a progressive increase: from LGN to V1 ($p < 0.05$), from V1 to V2 ($p < 0.05$), and from V2 to MT ($p < 0.001$). Additionally, the median gain autocorrelation (Fig. 5D) exhibits slower temporal decay and a broader central lobe in higher visual areas. Together, these results suggest that rapid temporal fluctuations in trial-to-trial variability diminish as information propagates along the visual hierarchy.

In contrast, the gain variance ρ_g (Fig. 5C) increases systematically along the hierarchy: from LGN to V1 ($p < 0.001$), and from V1 to V2 ($p < 0.001$). This trend indicates that the magnitude of gain fluctuations becomes larger at later stages of the visual pathway, consistent with prior observations[1].

Next, we examined the distribution of stimulus-drive variance, averaged across stimuli, defined as: $\bar{\rho}_f = \frac{1}{J}\sum_{j=1}^{J}\rho_f^{(j)}$, as shown in Fig. 5E. A larger value of $\bar{\rho}_f$ generally indicates stronger stimulus encoding by the neuron. In contrast to the gain variance trend, the stimulus-drive variance tends to decrease along the visual pathway: from LGN to V1 ($p < 0.001$), and from V1 to V2 ($p < 0.05$). This suggests a reduction in the strength of stimulus-driven encoding as information moves through the visual hierarchy.

To further investigate the relationship between stimulus-drive and gain variability, Fig. 5F shows scatter

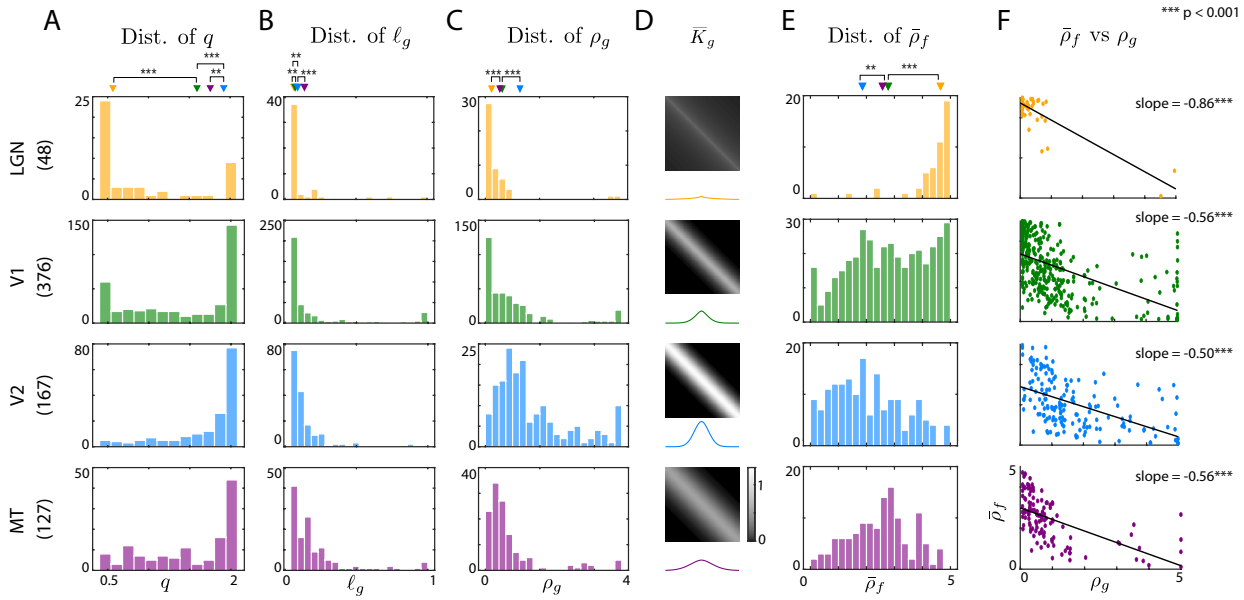

Figure 5: Comparison of inferred CMP hyperparameters across visual areas: the lateral geniculate nucleus (LGN), V1, V2, and MT (data from [1]). Rows top to bottom: LGN (yellow), V1 (green), V2 (blue), and MT (purple). The number of neurons per area is indicated in parentheses. **(A)** Distribution of gain power-law exponent q across neurons. Triangles indicate the median. **(B)** Distribution of gain length scale ℓ_g across neurons. **(C)** Distribution of gain variance ρ_g across neurons. **(D)** Median gain covariance $\overline{K}_g$ across the neuronal population (top) and gain auto-correlation (bottom). **(E)** Distribution of the average stimulus-drive variance ($\bar{\rho}_f = \frac{1}{J}\sum_{j=1}^{J}\rho_f^{(j)}$) across neurons. **(F)** Scatter plots of $\bar{\rho}_f$ versus ρ_g across neurons, with linear regression fits (black) and slopes annotated. $^{***}p < 0.001$.

plots of $\bar{\rho}_f$ versus ρ_g for individual neurons in each area. Each panel includes a linear regression line with the slope indicated. Across all visual areas, there is a statistically significant negative correlation between the two quantities ($p < 0.001$, t-test), indicating that neurons with weaker stimulus-driven components tend to have stronger gain variability, and vice versa. This finding aligns with recent reports of an anti-correlation between stimulus encoding strength and trial-to-trial, stimulus-independent variability in sensory neurons [33, 36–38].

Discussion

Advances in high-density electrophysiological recording techniques [39, 40] have enabled large-scale recordings of neuronal activity across multiple brain areas under diverse experimental conditions. Analyses of sensory responses to repeated stimuli consistently reveal trial-to-trial variability in neural spiking activity. This variability may arise from several sources, including synaptic transmission noise, gain modulation along sensory pathways, and dynamic effects such as adaptation or synaptic plasticity [6–11]. Capturing such variability is central to understanding neural coding [12–14].

While Poisson spiking models have provided a foundational framework [18, 19], numerous studies

have shown that cortical neurons often exhibit overdispersion–greater variability than expected under Poisson statistics [1]. Although several recent models have been developed to account for this greater-than-Poisson variability [23–29, 32, 33, 41, 42], few are designed for detailed, time-resolved analysis of trial-to-trial variability.

Here, we introduced the Continuous Modulated Poisson (CMP) model, a principled probabilistic framework for modeling spiking variability in continuous time. The CMP model builds upon the modulated Poisson model [1] by describing the spike rate at each time point as the product of two independent processes: a stimulus-driven component and a stochastic gain process. Both components are modeled as exponentiated Gaussian processes (GPs), enabling flexible characterization of their smooth temporal dynamics. This model belongs to the class of log-Gaussian Cox processes [43], as it constitutes a doubly stochastic Poisson process. We allow the stimulus-driven components to have condition-specific hyperparameters, while the gain hyperparameters are shared across conditions, supporting efficient and interpretable joint inference.

To accurately capture the long-range correlations observed in gain fluctuations, we introduced the Exponentiated Power Law (EPL) kernel–a new covariance function that generalizes the commonly used radial basis function (RBF) kernel–and demonstrated that it consistently outperforms standard alternatives such as the RBF and Matérn kernels in our application. We also developed a scalable variational inference method to jointly estimate the latent stimulus-driven and stochastic gain components, along with their hyperparameters, from spiking data. This approach was validated using synthetic experiments.

When applied to recordings from macaque V1 [34], the CMP model accurately partitioned response variability across time, trials, and stimulus conditions. It outperformed alternative approaches, including continuous-time extensions of the modulated Poisson model, in predicting held-out spiking activity and capturing the empirical Fano factor curves over time. Applying the CMP model to data from multiple stages of the visual hierarchy (LGN, V1, V2, and MT) [1], we again observed superior performance across areas and consistent predictions of overdispersion.

Analysis of the inferred CMP hyperparameters across visual areas revealed systematic trends. The temporal smoothness of gain fluctuations increased along the hierarchy, suggesting that trial-to-trial variability becomes more slowly varying in higher-order areas. In contrast, gain variance increased, consistent with an accumulation of shared noise or contextual modulation. Meanwhile, the variance of the stimulus-driven component decreased, consistent with increased selectivity and stabilization of tuning in downstream regions. A consistent negative correlation between the magnitudes of gain variance and stimulus-drive variance was observed across all areas, indicating that neurons with strong stimulus encoding tend to exhibit lower trial-to-trial variability, and vice versa. This observation aligns with previous findings of an anti-correlation between signal strength and variability in sensory neurons [33, 36–38]. These results have broader implications for understanding sensory coding, revealing how the brain balances fidelity and flexibility by modulating trial-to-trial variability across space and time [44]. They also align with broader theories in systems neuroscience, such as efficient coding [45] and probabilistic inference [46], where internal variability may be structured rather than random.

Our method complements black-box models by providing interpretable estimates of both stimulus-driven and modulatory influences on spiking variability. It is particularly well-suited for experimental settings with repeated stimuli and temporally structured variability, such as studies of sensory tuning or motor

control [1, 34, 47]. The CMP model extends prior work by enabling continuous-time analysis of spike trains, jointly modeling stimulus-driven and stimulus-independent components across conditions, and allowing full inference over both latent processes and their hyperparameters. While we found that the exponentiated power-law (EPL) kernel performed best in our experiments, kernel selection may depend on the task-specific temporal structure of neural variability. More broadly, the EPL kernel introduced here is generally applicable and may support future Gaussian process–based modeling efforts beyond neuroscience.

This work also opens several directions for future research. First, extensions to high-dimensional population activity [33] may allow joint modeling of shared variability across neurons. Second, adapting the model for alternative link functions [26] could allow it to generalize to different statistical structures. Third, developing inference algorithms compatible with calcium imaging data [38] would expand its utility to widely used optical recording techniques. Fourth, integration with behavioral covariates [48] or state variables such as attention [49] or arousal [50] could improve inference of nonstationary trial structure. Finally, while our implementation focused on neural spike trains, the general formulation of the CMP model could be extended to other domains involving overdispersed count data [51–53].

To facilitate reproducibility and dissemination, we have released our implementation of the CMP model on GitHub [54]. In summary, our proposed CMP framework provides a scalable, flexible, and interpretable approach for modeling neuronal spike variability, revealing new insights into the structure of trial-to-trial fluctuations and their organization across the visual hierarchy.

Materials and Methods

CMP model inference

We start by recalling the general CMP forward model:

$$
\begin{aligned}
y(t) \mid s &\sim \mathrm{PoissProc}\big(f(s,t)\cdot g(t)\big),\\
\log f(s,t) &\sim \mathcal{GP}(0,K_f), K_f(t_1,t_2) = \rho_f \exp\left(-\frac{(t_1-t_2)^2}{2\ell_f^2}\right),\\
\log g(t) &\sim \mathcal{GP}(0,K_g), K_g(t_1,t_2) = \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_1-t_2}{\ell_g}\right|^q\right).
\end{aligned}
$$

In our inference framework, we consider a setting with multiple repeats ($n = 1,\dots,N$) for a given stimulus condition, with data potentially collected across different stimulus conditions ($j = 1,\dots,J$). We perform inference on a uniform grid of time points ($t = 1,\dots,T$), with grid size set to dt seconds. Note that the model itself is defined in continuous time and can therefore be consistently applied to data measured at arbitrary temporal grid resolutions.

Let $y_{t,j,n}$ denote the spike count for the tth time bin, j^{th} stimulus condition, and n^{th} repeat. Further, let $\mathbf{f}_j = [f_{1,j}\dots f_{T,j}]^\top$ and $\mathbf{g}_{j,n} = [g_{1,j,n}\dots g_{T,j,n}]^\top$ denote the corresponding stimulus-driven and stochastic gain processes. We also assume that each stimulus condition has its own set of hyperparameters for the stimulus-driven component (i.e., for $j = 1,\dots,J$ stimuli, $\rho_f = \rho_f^{(j)}$ and $\ell_f = \ell_f^{(j)}$). In

contrast, because the gain process is stimulus-independent, we assume that the gain hyperparameters are shared across all stimulus conditions.

Let $\tilde{\mathbf{f}}_j = \log \mathbf{f}_j$ and $\tilde{\mathbf{g}}_{j,n} = \log \mathbf{g}_{j,n}$ represent the log-transformed stimulus-driven and gain components, for notational brevity. Note that, following the CMP generative model, the latent variables $\mathbf{f} = \tilde{\mathbf{f}}_{1:J}$ and $\mathbf{g} = \tilde{\mathbf{g}}_{1:J,1:N}$, along with the hyperparameters $\theta = \{\rho_g, \ell_g, q, \rho_f^{(j)}, \ell_f^{(j)}, j = 1, \ldots, J\}$, are all unknown quantities that need to be estimated given the spiking observations $\mathbf{y} = y_{1:T,1:J,1:N}$.

Frequency domain transformation of stimulus-driven components

The radial basis function (RBF) kernel diagonalizes in the Fourier domain, a property previously exploited to improve computational efficiency [33, 55]. Motivated by this, we adopt a frequency-domain representation of the stimulus-driven components ($\mathbf{f}$) in our inference procedure. Accordingly, we relate the time-domain components to their corresponding Fourier-domain components:

$$\tilde{\mathbf{f}}_{\mathbf{j}} = \mathbf{B}^{(\mathrm{j})} \mathbf{w}_{\mathbf{j}},$$

where $\mathbf{B}^{(\mathrm{j})} \in \mathbb{R}^{T \times M_j}$ (noting that typically $M_j \ll T$ due to pruning of unnecessary Fourier coefficients that do not substantially contribute to explaining variability in $\tilde{\mathbf{f}}_j$ [33, 55]) represents the standard real orthonormal inverse discrete Fourier transform matrix. Then, with the diagonalization of the RBF kernel [33, 55], the Fourier-transformed stimulus-driven components $\mathbf{w}_j = \left[w_{1,j}, w_{2,j}, \ldots, w_{M_j,j}\right]^\top$ are distributed as independent Gaussian random variables with the following parameters:

$$w_{m,j} \sim \mathcal{N}\left(0, \tilde{k}_m^{(j)}\right),$$

where $\tilde{k}_m^{(j)} = \sqrt{2\pi} \rho_f^{(j)} \ell_f^{(j)} \exp\left(-\frac{1}{2} {f_m^{(j)}}^2 {\ell_f^{(j)}}^2\right)$, with $f_m^{(j)}$ denoting the mth Fourier frequency [33, 55]. Thus, we use the frequency-domain stimulus-driven components $\mathbf{w} = \mathbf{w_{1:J}}$ in place of their time-domain counterparts $\mathbf{f}$, reducing computational complexity by avoiding matrix inversions and enabling pruning ($M_j \ll T$). We do not apply frequency-domain inference to the gain components $\mathbf{g}$, as the heavy tails of the EPL kernel do not support effective pruning.

Variational Inference

To jointly infer the latent variables $\mathbf{w}$ and $\mathbf{g}$ and hyperparameters θ, we introduce an efficient inference procedure based on Variational Inference (VI) [35], which aims to find an approximate distribution $q(\mathbf{w}, \mathbf{g})$ for the intractable true posterior $p(\mathbf{w}, \mathbf{g}|\mathbf{y}, \theta)$. In VI, the approximate posterior is first introduced into the data log-likelihood by rewriting it as [32]:

$$\begin{aligned}
\log p(\mathbf{y}|\theta) &= \mathbb{E}_q\left[\log\left(\frac{p(\mathbf{y}, \mathbf{w}, \mathbf{g}|\theta)}{q(\mathbf{w}, \mathbf{g})} \cdot \frac{q(\mathbf{w}, \mathbf{g})}{p(\mathbf{w}, \mathbf{g}|\mathbf{y}, \theta)}\right)\right] \\
&= \underbrace{\mathbb{E}_q\left[\log \frac{p(\mathbf{y}, \mathbf{w}, \mathbf{g}|\theta)}{q(\mathbf{w}, \mathbf{g})}\right]}_{\mathcal{L}(q)} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\mathbf{w}, \mathbf{g})}{p(\mathbf{w}, \mathbf{g}|\mathbf{y}, \theta)}\right]}_{D_{KL}(q||p)},
\end{aligned}$$

where $D_{KL}(q||p)$ is the Kullback-Leibler divergence between the variational approximation and the true posterior. Since $D_{KL}(q||p)$ is non-negative, $\mathcal{L}(q)$ provides a lower bound on the model evidence, and is therefore called the Evidence Lower Bound (ELBO). Maximizing the ELBO is equivalent to minimizing $D_{KL}(q||p)$, thus finding the variational approximation closest to the true posterior.

Variational distributions and the ELBO

We assume that the variational distributions $q(\mathbf{w}, \mathbf{g})$ factorizes into independent Gaussian distributions over stimuli $j = 1, \ldots, J$ and trials $n = 1, \ldots, N$:

$$q(\mathbf{w}, \mathbf{g}) = \prod_{j=1}^{J} \left(q(\mathbf{w_j}) \prod_{n=1}^{N} q(\mathbf{g_{j,n}}) \right),$$
$$q(\tilde{\mathbf{g}}_{\mathbf{j,n}}) \sim \mathcal{N}(\boldsymbol{\mu}_{j,n}, \boldsymbol{\Sigma}_{j,n}) \text{ and } q(\mathbf{w_j}) \sim \mathcal{N}(\boldsymbol{\nu}_j, \boldsymbol{\Omega}_j),$$

where $\boldsymbol{\Sigma}_{j,n} \in \mathbb{R}^{T \times T}$ and $\boldsymbol{\Omega}_j \in \mathbb{R}^{M_j \times M_j}$ are full covariance matrices. Note that, by the property of linear transformations of Gaussian random variables [56], this in turn implies that the variational distributions over the time-domain stimulus-driven components are also Gaussian:

$$q(\tilde{\mathbf{f}}_{\mathbf{j}}) = q(\mathbf{B}^{(\mathrm{j})}\mathbf{w_j}) \sim \mathcal{N}\left(\mathbf{B}^{(\mathrm{j})}\boldsymbol{\nu}_j, \mathbf{B}^{(\mathrm{j})}\boldsymbol{\Omega}_j\mathbf{B}^{(\mathrm{j})\top}\right).$$

Accordingly, the ELBO can be derived as:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{y}, \mathbf{w}, \mathbf{g}|\theta)\right] - \mathbb{E}_q\left[\log q(\mathbf{w}, \mathbf{g})\right] \\
&= -\sum_{t,j,n} dt \exp\Big((\mathbf{B}^{(\mathrm{j})}\boldsymbol{\nu}_j)_t + (\mathbf{B}^{(\mathrm{j})}\boldsymbol{\Omega}_j\mathbf{B}^{(\mathrm{j})\top})_{t,t}/2 + (\boldsymbol{\mu}_{j,n})_t \\
&\quad + (\boldsymbol{\Sigma}_{\mathbf{j,n}})_{t,t}/2\Big) + \sum_{t,j,n} y_{t,j,n}\left((\mathbf{B}^{(\mathrm{j})}\boldsymbol{\nu}_j)_t + (\boldsymbol{\mu}_{j,n})_t\right) \\
&\quad - \frac{JN}{2}\log|K_g| - \frac{1}{2}\sum_{j,n} \mathrm{trace}\left(K_g^{-1}\left(\boldsymbol{\mu}_{j,n}\boldsymbol{\mu}_{j,n}^\top + \boldsymbol{\Sigma}_{j,n}\right)\right) \\
&\quad - \frac{1}{2}\sum_{j,m}\log\tilde{k}_m^{(j)} - \frac{1}{2}\sum_{j,m}\left(\tilde{k}_m^{(j)}\right)^{-1}\left(\nu_{m,j}^2 + (\boldsymbol{\Omega}_j)_{m,m}\right) \\
&\quad + \sum_{j,n}\frac{1}{2}\log|\boldsymbol{\Sigma}_{j,n}| + \sum_j \frac{1}{2}\log|\boldsymbol{\Omega}_j| + \frac{1}{2}\sum_j M_j + \text{const.}
\end{aligned} \tag{12}$$

Joint inference of variational parameters and hyperparameters

In this section, we outline our proposed iterative scheme for the joint inference of variational parameters and hyperparameters, as illustrated in Algorithm 1. We alternate between optimizing the stimulus-drive hyperparameters $\theta_f^{(j)} = \{\rho_f^{(j)}, \ell_f^{(j)}\}$ for $j = 1, \ldots, J$, and the gain hyperparameters $\theta_g = \{\rho_g, \ell_g, q\}$, until convergence. Each optimization maximizes the ELBO in Eq. (12), using Algorithm 2 (described in the next section) to infer the conditional variational parameters. We employed the built-in MATLAB function *fmincon*, a gradient-based solver for constrained nonlinear multivariable problems, to perform each maximization. To ensure stability, convergence, and valid covariance matrices, we constrained the hyperparameters to the following ranges: $\rho_g, \rho_f^{(1:J)} \in [0.05, 6]$, $\ell_g, \ell_f^{(1:J)} \in [0.05, 10]$, and $q \in [0.4, 2]$.

Algorithm 1 Joint inference of variational parameters and hyperparameters

Inputs: Spiking observations $y_{1:T,1:J,1:N}$

Outputs: Estimated hyperparameters $\theta = \{\rho_g, \ell_g, q, \rho_f^{(1:J)}, \ell_f^{(1:J)}\}$, and variational parameters $\boldsymbol{\nu}_{1:J}$, $\boldsymbol{\Omega}_{1:J}$, $\boldsymbol{\mu}_{1:J,1:N}$, and $\boldsymbol{\Sigma}_{1:J,1:N}$

Initialization: $\delta = 1$, $\theta_g = \{\rho_g = 0.1, \ell_g = 0.1, q = 1\}$ $\theta_f^{(1:J)} = \{\rho_f^{(1:J)} = 0.1, \ell_f^{(1:J)} = 1\}$

while $\delta > 10^{-5}$ **do**
  $\theta^{\text{prev}} = \theta$
  Fix $\theta_g = \{\rho_g, \ell_g, q\}$
  for $j = 1, \ldots, J$ **do**
    Optimize $\theta_f^{(j)} = \{\rho_f^{(j)}, \ell_f^{(j)}\}$ by maximizing the ELBO in Eq. (12), using Algorithm 2 to derive the conditional variational parameters
  Fix $\theta_f^{(1:J)} = \{\rho_f^{(1:J)}, \ell_f^{(1:J)}\}$
  Optimize $\theta_g = \{\rho_g, \ell_g, q\}$ by maximizing the ELBO in Eq. (12), using Algorithm 2 to derive the conditional variational parameters
  $\delta = ||\theta - \theta^{\text{prev}}||/||\theta||$
for $j = 1, \ldots, J$ **do**
  Estimate the final settings of $\boldsymbol{\nu}_j$, $\boldsymbol{\Omega}_j$, $\boldsymbol{\mu}_{j,1:N}$, and $\boldsymbol{\Sigma}_{j,1:N}$ based on the final estimates of the hyperparameters $\theta^{(j)} = \{\theta_g, \theta_f^{(j)}\}$ using Algorithm 2

Inferring variational parameters given hyperparameters

Here, we outline the procedure for inferring the variational parameters by optimizing the ELBO for a given set of hyperparameters θ (i.e., assuming K_g and $\tilde{k}_{1:M_j}^{(j)}$ are known), as summarized in Algorithm 2. Note that the ELBO in Eq. (12) is independent across stimulus conditions $j = 1, \ldots, J$, when conditioned on θ. Therefore, we focus on estimating the variational parameters corresponding to the jth stimulus: $\boldsymbol{\nu}_j$, $\boldsymbol{\Omega}_j$, $\boldsymbol{\mu}_{j,1:N}$, and $\boldsymbol{\Sigma}_{j,1:N}$. We maximize the ELBO to find the optimal settings of these variational parameters, alternating between updating the stimulus-drive variational parameters ($\boldsymbol{\nu}_j$, $\boldsymbol{\Omega}_j$) and the gain variational parameters ($\boldsymbol{\mu}_{j,1:N}$, $\boldsymbol{\Sigma}_{j,1:N}$) until convergence. Each maximization step uses a Newton–Raphson procedure.

First, consider the estimation of the stimulus-drive variational parameters $\boldsymbol{\nu}_j$ and $\boldsymbol{\Omega}_j$, given the estimated gain variational parameters from the previous iteration. The gradient and Hessian of Eq. (12) with respect to $\boldsymbol{\nu}_j$ can be derived as:

$$\begin{aligned}
\nabla_{\boldsymbol{\nu}_j}\mathcal{L} &= \sum_n \mathbf{B}^{(\text{j})\top}\left(\mathbf{y}_{j,n} - \widehat{\boldsymbol{\lambda}}_{j,n}\right) - \left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1}\boldsymbol{\nu}_j, \\
\nabla^2_{\boldsymbol{\nu}_j}\mathcal{L} &= -\sum_n \mathbf{B}^{(\text{j})\top}\widehat{\boldsymbol{\Lambda}}_{j,n}\mathbf{B}^{(\text{j})} - \left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1},
\end{aligned} \tag{13}$$

where $\mathbf{y_{j,n}} = [y_{1,j,n}, y_{2,j,n}, \ldots, y_{t,j,n}]^\top$, $\widehat{\boldsymbol{\lambda}}_{j,n} = \left[\widehat{\lambda}_{1,j,n}, \widehat{\lambda}_{2,j,n}, \ldots, \widehat{\lambda}_{t,j,n}\right]^\top$, $\widehat{\lambda}_{t,j,n} = dt\exp\Big((\mathbf{B}^{(\text{j})}\boldsymbol{\nu}_j)_t + (\mathbf{B}^{(\text{j})}\boldsymbol{\Omega}_j\mathbf{B}^{(\text{j})\top})_{t,t}/2 + (\boldsymbol{\mu}_{j,n})_t + (\boldsymbol{\Sigma}_{j,n})_{t,t}/2\Big)$, $\widehat{\boldsymbol{\Lambda}}_{j,n}$ is a diagonal matrix with the tth diagonal entry given by $\widehat{\lambda}_{t,j,n}$, and $\widetilde{\mathbf{K}}^{(j)}$ is a diagonal matrix with the mth diagonal entry equal to $\tilde{k}_m^{(j)}$. Further, taking derivatives in Eq. (12) with respect to $\boldsymbol{\Omega}_j$ yields:

$$\nabla_{\boldsymbol{\Omega}_j}\mathcal{L} = -\frac{1}{2}\sum_n \mathbf{B}^{(\text{j})\top}\widehat{\boldsymbol{\Lambda}}_{j,n}\mathbf{B}^{(\text{j})} + \frac{1}{2}\left(\boldsymbol{\Omega}_j\right)^{-1} - \frac{1}{2}\left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1},$$

Algorithm 2 Estimation of variational parameters given hyperparameters

Inputs: Spiking observations $\mathbf{y}_{j,1:N} = y_{1:T,j,1:N}$ and hyperparameters $\theta^{(j)} = \{\rho_g, \ell_g, q, \rho_f^{(j)}, \ell_f^{(j)}\}$
Outputs: Estimated $\boldsymbol{\nu}_j$, $\boldsymbol{\Omega}_j$, $\boldsymbol{\mu}_{j,1:N}$, and $\boldsymbol{\Sigma}_{j,1:N}$
Initialization: $\epsilon = 1$, $\widehat{\boldsymbol{\lambda}}_{j,1:N} = \mathbf{y}_{j,1:N}$, $\boldsymbol{\nu}_j = \mathbf{0}$, $\boldsymbol{\Omega}_j = \mathbf{0}$, $\boldsymbol{\mu}_{j,1:N} = \mathbf{0}$, and $\boldsymbol{\Sigma}_{j,1:N} = \mathbf{0}$

Get K_g, $\widetilde{\mathbf{K}}^{(j)}$ and $\mathbf{B}^{(\mathrm{j})}$ from $\theta^{(j)}$
while $\epsilon > 10^{-5}$ **do**
  $\widehat{\boldsymbol{\lambda}}^{\text{prev}}_{j,1:N} = \widehat{\boldsymbol{\lambda}}_{j,1:N}$
  Estimate $\boldsymbol{\nu}_j$ using the Newton–Rhapson Method with:
  $\nabla_{\boldsymbol{\nu}_j}\mathcal{L} = \sum_n \mathbf{B}^{(\mathrm{j})\top}\left(\mathbf{y}_{j,n} - \widehat{\boldsymbol{\lambda}}_{j,n}\right) - \left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1}\boldsymbol{\nu}_j$
  $\nabla^2_{\boldsymbol{\nu}_j}\mathcal{L} = -\sum_n \mathbf{B}^{(\mathrm{j})\top}\widehat{\boldsymbol{\Lambda}}_{j,n}\mathbf{B}^{(\mathrm{j})} - \left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1}$
  Update $\boldsymbol{\Omega}_j = -\left(\nabla^2_{\boldsymbol{\nu}_j}\mathcal{L}\right)^{-1}$
  for $n = 1, \ldots, N$ **do**
    Update $\widehat{\boldsymbol{\lambda}}_{j,n} = \left[\widehat{\lambda}_{1,j,n}, \widehat{\lambda}_{2,j,n}, \ldots, \widehat{\lambda}_{t,j,n}\right]^\top$, where $\widehat{\lambda}_{t,j,n} = dt \exp\left((\mathbf{B}^{(\mathrm{j})}\boldsymbol{\nu}_j)_t + (\mathbf{B}^{(\mathrm{j})}\boldsymbol{\Omega}_j\mathbf{B}^{(\mathrm{j})\top})_{t,t}/2\right) \times \exp\left((\boldsymbol{\mu}_{j,n})_t + (\boldsymbol{\Sigma}_{j,n})_{t,t}/2\right)$
  for $n = 1, \ldots, N$ **do**
    Estimate $\boldsymbol{\mu}_{j,n}$ using the Newton–Rhapson Method with:
    $\nabla_{\boldsymbol{\mu}_{j,n}}\mathcal{L} = \left(\mathbf{y}_{j,n} - \widehat{\boldsymbol{\lambda}}_{j,n}\right) - K_g^{-1}\boldsymbol{\mu}_{j,n}$
    $\nabla^2_{\boldsymbol{\mu}_{j,n}}\mathcal{L} = -\widehat{\boldsymbol{\Lambda}}_{j,n} - K_g^{-1}$
    Update $\boldsymbol{\Sigma}_{j,n} = -\left(\nabla^2_{\boldsymbol{\mu}_{j,n}}\mathcal{L}\right)^{-1}$
    Update $\widehat{\boldsymbol{\lambda}}_{j,n}$ as defined above
  $\epsilon = ||\widehat{\boldsymbol{\lambda}}_{j,1:N} - \widehat{\boldsymbol{\lambda}}^{\text{prev}}_{j,1:N}||/||\widehat{\boldsymbol{\lambda}}_{j,1:N}||$

and setting this derivative to zero results in:

$$\boldsymbol{\Omega}_j = \left(\sum_n \mathbf{B}^{(\mathrm{j})\top}\widehat{\boldsymbol{\Lambda}}_{j,n}\mathbf{B}^{(\mathrm{j})} + \left(\widetilde{\mathbf{K}}^{(j)}\right)^{-1}\right)^{-1} = -\left(\nabla^2_{\boldsymbol{\nu}_j}\mathcal{L}\right)^{-1}. \tag{14}$$

Accordingly, we update the stimulus-drive variational mean $\boldsymbol{\nu}_j$ using Newton–Raphson optimization with the gradient and Hessian in Eq. (13), and subsequently update the stimulus-drive variational covariance $\boldsymbol{\Omega}_j$ using the Hessian as in Eq. (14).

Next, we update the gain variational parameters independently for $n = 1, \ldots, N$ using the updated stimulus-drive variational parameters, following an equivalent Newton–Raphson procedure, with the gradient and Hessian relative to $\boldsymbol{\mu}_{j,n}$ given by:

$$\nabla_{\boldsymbol{\mu}_{j,n}}\mathcal{L} = \left(\mathbf{y}_{j,n} - \widehat{\boldsymbol{\lambda}}_{j,n}\right) - K_g^{-1}\boldsymbol{\mu}_{j,n},$$
$$\nabla^2_{\boldsymbol{\mu}_{j,n}}\mathcal{L} = -\widehat{\boldsymbol{\Lambda}}_{j,n} - K_g^{-1},$$

and the update rule of $\boldsymbol{\Sigma}_{j,n}$ given by:

$$\boldsymbol{\Sigma}_{j,n} = \left(\widehat{\boldsymbol{\Lambda}}_{j,n} + K_g^{-1}\right)^{-1} = -\left(\nabla^2_{\boldsymbol{\mu}_{j,n}}\mathcal{L}\right)^{-1}.$$

Following the above derivations, we iterate between updating the stimulus-drive variational parameters and gain variational parameters until convergence. The combined iterative procedure for variational parameter inference is summarized in Algorithm 2.

Log-likelihood derivation of the CMP model

While we optimize the ELBO for parameter inference, accurately and fairly comparing models requires computing the log-likelihood on held-out data. To estimate this for the CMP model, we used the importance sampling procedure described below. Note that,

$$\log\left(p\left(\mathbf{y}|s\right)\right) = \sum_{j,n} \log\left(p\left(\mathbf{y}_{j,n}|s\right)\right), \text{ where,} \tag{15}$$

$$p\left(\mathbf{y}_{j,n}|s\right) = \int p\left(\mathbf{y}_{j,n}|\mathbf{g}_{j,n}, s\right) p\left(\mathbf{g}_{j,n}\right) d\mathbf{g}_{j,n}. \tag{16}$$

Observing that the integral in Eq.(16) is intractable, we used importance sampling [57] to approximate it. We employed the variational distribution $q\left(\tilde{\mathbf{g}}_{\mathbf{j},\mathbf{n}}\right)$ as the proposal distribution,

$$\begin{aligned} p\left(\mathbf{y}_{j,n}|\mathbf{B}_j\right) &= \mathbb{E}_{q(\mathbf{g}_{j,n})}\left[p\left(\mathbf{y}_{j,n}|\mathbf{g}_{j,n}, s\right)\frac{p\left(\mathbf{g}_{j,n}\right)}{q\left(\mathbf{g}_{j,n}\right)}\right] \\ &\approx \frac{1}{I}\sum_{i=1}^{I} p\left(\mathbf{y}_{j,n}|\mathbf{g}_{j,n}^{(i)}, s\right)\frac{p\left(\mathbf{g}_{j,n}^{(i)}\right)}{q\left(\mathbf{g}_{j,n}^{(i)}\right)}, \end{aligned}$$

where $\log \mathbf{g}_{j,n}^{(i)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{j,n}, \boldsymbol{\Sigma}_{j,n}\right)$. Defining:

$$L_{j,n}^{(i)} := \log\left(p\left(\mathbf{y}_{j,n}|\mathbf{g}_{j,n}^{(i)}, s\right)\right) + \log\left(p\left(\mathbf{g}_{j,n}^{(i)}\right)\right) - \log\left(q\left(\mathbf{g}_{j,n}^{(i)}\right)\right)$$

and applying the log-sum-exp trick for numerical stability, we obtained:

$$\log\left(p\left(\mathbf{y}_{j,n}|s\right)\right) \approx \log\left(\sum_{i=1}^{I}\exp\left(L_{j,n}^{(i)} - L_{j,n}^{\mathsf{max}}\right)\right) + L_{j,n}^{\mathsf{max}} - \log\left(I\right),$$

where $L_{j,n}^{\mathsf{max}} = \max_i L_{j,n}^{(i)}$. We set $I = 1000$ and used the above formulation to estimate the log-likelihood of the spiking observations in the nth test trial of the jth stimulus, and finally derived the joint log-likelihood of all test data according to Eq. (15).

Partitioning variability with CMP

Here we provide the general formula for partitioning variability into a Poisson component, equal to the mean, and a modulatory component, based on the CMP model in the setting where the stimulus drive $f(s,t)$ is a time-varying signal. The continuous-time formulation of the CMP model allows this partitioning to be computed for arbitrary time windows, enabling an analytic characterization of how overdispersion grows with the bin size used to count spikes. For a time bin of size Δ, the mean is given by:

$$\text{mean}[y_\Delta \mid s] = e^{\rho_g/2}\int_0^\Delta f(s,t)dt,$$

and the variance is:

$$\text{var}[y_\Delta \mid s] = e^{\rho_g/2}\int_0^\Delta f(s,t)dt$$

$$+ e^{\rho_g} \int_0^\Delta \int_0^\Delta f(s,t_1) f(s,t_2) \left(e^{K_g(t_1,t_2)} - 1 \right) dt_1 dt_2.$$

From these expressions, we can compute the Fano factor as a function of bin size, which is simply the ratio of variance to mean:

$$FF(\Delta) = 1 + \frac{e^{\rho_g/2} \int_0^\Delta \int_0^\Delta f(s,t_1) f(s,t_2) \left(e^{K_g(t_1,t_2)} - 1 \right) dt_1 dt_2}{\int_0^\Delta f(s,t) dt}.$$

The details of the derivation are provided in the SI Appendix. To summarize statistics at a given bin size Δ over the entire observation window T, we use the following approach: we apply a sliding window of R consecutive samples (duration Δ), shift it systematically through all K time samples (duration T), and average the computed statistics across all such windows.

Alternative models considered for performance comparison

We compared the performance of the CMP model with the alternative models described below.

Baseline Poisson

All log-likelihood figures reported in this paper are expressed as log-likelihood gains of each method relative to the baseline Poisson model. The baseline Poisson model assumed spike counts across trials are generated by a Poisson process:

$$y_{t,j,n} \sim \text{Poisson}\left(\lambda_{t,j}\right),$$

and, to compensate for sparsity, we further assumed a Gamma prior on the firing rate $\lambda_{t,j} \sim \text{Gamma}\left(\alpha, \beta\right)$, where α and β are hyperparameters. Since the Gamma distribution is conjugate to the Poisson distribution [58], the posterior distribution is also Gamma-distributed:

$$p\left(\lambda_{t,j} | y_{t,j,1:N}\right) \sim \text{Gamma}\left(\alpha + \sum_{n=1}^{N} y_{t,j,n}, \beta + N\right).$$

Thus, we found the MAP estimate of the firing rate by the mode of the posterior Gamma distribution, given by:

$$\widehat{\lambda}_{t,j} = \frac{\alpha + \sum_{n=1}^{N} y_{t,j,n} - 1}{\beta + N}.$$

To fix the hyperparameters, we performed moment matching with the mean $= \frac{\alpha}{\beta}$ and variance $= \frac{\alpha}{\beta^2}$. Note further that the mean and variance conditioned on $\lambda_{t,j}$ are identical for this model, resulting in a Fano factor of 1.

Poisson-GP

This model includes only the stimulus-driven components, assuming that trial-to-trial variability is fully described by the stochastic nature of the Poisson process:

$$y(t) \sim \text{Poisson}\left(\lambda_t = f(s,t)\right).$$

The stimulus drive itself is modeled by a Gaussian process, similar to the CMP model:

$$\log f(s,t) \sim \mathcal{GP}(0, K_f), K_f(t_1, t_2) = \rho_f \exp\left(-\frac{(t_1 - t_2)^2}{2\ell_f^2}\right).$$

We inferred the hyperparameters and latent stimulus-driven components of this model using a simplified version of our proposed variational inference procedure. As in the baseline Poisson model, the mean and variance conditioned on the stimulus are identical, yielding a Fano factor of 1.

Goris-indep

This model extends the modulated Poisson framework of Goris et al.[1] to continuous time. The Goris model in[1] considered spike counts within a single time bin. For spiking observations from the jth stimulus presentation with $t = 1, \ldots, T$ time bins, we assumed that the stimulus-independent gain is independent across time bins with shared variance $\sigma_{g_j}^2$, i.e.,

$$\begin{aligned} y_{t,j,n} &\sim \text{Poisson}(\lambda_{t,j,n}), \\ \lambda_{t,j,n} &= g_{t,j,n} f_{t,j} dt, \end{aligned}$$

where $g_{t,j,n} \sim \text{Gamma}(r_j, s_j)$, with shape parameter $r_j = \frac{1}{\sigma_{g_j}^2}$ and scale parameter $s_j = \sigma_{g_j}^2$. Marginalizing over the gain variables yields a product of independent Negative Binomial distributions:

$$\begin{aligned} p(y_{1:T,j,1:N} | r_j, s_j) &= \int_0^\infty p(y_{1:T,j,1:N}, \lambda_{1:T,j,1:N} | r_j, s_j)) d\lambda_{1:T,j,1:N} \\ &= \prod_{t,n=1}^{T,N} \text{NegativeBinomial}\left(y_{t,j,n} \middle| r_j, \frac{1}{1 + s_{t,j}}\right), \end{aligned} \tag{17}$$

where $s_{t,j} = \sigma_{g_j}^2 f_{t,j} dt$. To ensure smoothness of the stimulus drive, we set $s_{t,j} = \sigma_{g_j}^2 f_{t,j} dt = f_{t,j} dt / r_j = \hat{\lambda}_{t,j} / r_j$, where $\hat{\lambda}_{t,j}$ denotes the firing rates estimated by the Poisson-GP model. For $j = 1, \ldots, J$, we searched for the values of r_j that minimize the negative log-likelihood of the spiking observations $y_{1:T,j,1:N}$. Details of Eq. (17) and the Fano factor for this model are provided in the SI Appendix.

Goris-const

This is another extension of the Goris model [1] to continuous time. Here, we assumed the gain is constant across time bins within a trial:

$$y_{t,j,n} \sim \text{Poisson}(\lambda_{t,j,n}),$$

$$\lambda_{t,j,n} = g_{j,n} f_{t,j} dt,$$

where $g_{j,n} \sim \mathrm{Gamma}(r_j, s_j)$ with shape parameter $r_j = \frac{1}{\sigma^2_{g_j}}$ and scale parameter $s_j = \sigma^2_{g_j}$. Further, we assumed that the total spike count across all time bins in a trial, $Y_{j,n} = \sum_t y_{t,j,n}$, also shares the same gain variance $\sigma^2_{g_j}$, i.e, $Y_{j,n} \sim \mathrm{Poisson}(\widetilde{g}_{j,n} f_j dtT)$ with $\widetilde{g}_{j,n} \sim \mathrm{Gamma}(r_j, s_j)$. We found $\sigma^2_{g_j}$ by fitting for $Y_{j,n}$ across training data $n = 1, \ldots, N$ as in [1], and set the stimulus encoding component $f_{t,j} dt = \hat{\lambda}_{t,j}$, where $\hat{\lambda}_{t,j}$ represents the firing rates estimated by the Poisson-GP model. Details of the mean-variance relationship and log-likelihood derivations for this model are included in the SI Appendix.

CMP-RBF

This model follows the same generative structure as the CMP model, except that the gain covariance is parameterized by the standard RBF kernel:

$$K_g(t_1, t_2) = \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_m - t_n}{\ell_g}\right|^2\right).$$

CMP-Matérn

This variant also follows the CMP framework, with the gain covariance parameterized by a Matérn kernel of order 0 in the CMP-Matérn-0 model:

$$K_g(t_1, t_2) = \rho_g \exp\left(-\left|\frac{t_1 - t_2}{\ell_g}\right|\right),$$

and a Matérn kernel of order 1 in the CMP-Matérn-1 model:

$$K_g(t_1, t_2) = \rho_g \left(1 + \sqrt{3}\left|\frac{t_1 - t_2}{\ell_g}\right|\right) \exp\left(-\sqrt{3}\left|\frac{t_1 - t_2}{\ell_g}\right|\right).$$

We used the same variational inference procedure as the CMP model to infer the hyperparameters and latent variables of the CMP-RBF and CMP-Matérn models. The log-likelihood and mean-variance derivations for these models are analogous to those of the CMP model.

Datasets studied

In this paper, we applied the CMP model to two datasets recorded from anesthetized, paralyzed, adult macaque monkeys.

Multi-electrode array recordings from the primary visual cortex

Complete details about this dataset are available in [34]. In brief, extracellular recordings were obtained using a 10-by-10 fixed silicon microelectrode array (length 1mm, spaced 400 μm) implanted in the superficial layers of the primary visual cortex (area V1) of three adult male macaque monkeys. The animals were stimulated with drifting sinusoidal gratings at full contrast, with the orientation of the

grating varied around the clock in steps of 5 degrees, yielding a total of 72 orientations. Each stimulus was presented for 1,280ms and followed by a blank screen of mean luminance for another 1,280ms, repeated for 50 trials per stimulus condition. We analyzed 654 isolated units from five datasets in this paper (147, 113, 113, 133, and 148 units in each dataset). For each unit and each stimulus, we randomly selected $N = 45$ trials for training (i.e., for hyperparameter and stimulus-drive inference) and used the remaining $N = 5$ trials to derive the test log-likelihood.

Single-electrode recordings along the visual hierarchy

Full details regarding these datasets are available in [1]. Briefly, extracellular recordings were made from 18 adult macaque monkeys of either sex, in different areas along the visual hierarchy: the lateral geniculate nucleus (LGN), V1, V2, and MT. The animals were presented with drifting sinusoidal gratings of the preferred size and speed, varying either in spatial frequency (12 spatial frequencies, ranging from 0 to 10 cycles/deg) or in drift direction (16 equally spaced directions). Each grating was presented for 1,000 ms and repeated at least five times. We analyzed the activity of 48, 376, 167, and 127 cells from each population, respectively. For each unit and each stimulus, we randomly selected $N = 1$ trial to derive the test log-likelihood and used the remaining trials for training the model.

Acknowledgments

We are grateful to Arnulf Graf, Robbe Goris, Eero Simoncelli, and J. Anthony Movshon for providing the datasets used in this study. AR was supported by NIH T32 institutional training grant (T32MH065214). ASC was supported by NSF CAREER award 2340338. JWP was supported by grants from the Simons Collaboration on the Global Brain (SCGB AWD543027) the NIH BRAIN initiative (R01DA056404, and R01EB02694), the National Eye Institute (NEI) of the National Institutes of Health (NIH) (R01EY033064), and a U19 NIH-NINDS BRAIN Initiative Award (5U19NS123716).

References

[1] R L T Goris, J A Movshon, and E P Simoncelli. Partitioning neuronal variability. *Nature Neuroscience*, 17(6):858–865, 2014.

[2] George J. Tomko and Donald R. Crapper. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Research*, 79(3):405–418, 1974. ISSN 0006-8993. doi: https://doi.org/10.1016/0006-8993(74)90438-7. URL https://www.sciencedirect.com/science/article/pii/0006899374904387.

[3] G.H. Henry, P.O. Bishop, R.M. Tupper, and B. Dreher. Orientation specificity and response variability of cells in the striate cortex. *Vision Research*, 13(9):1771–1779, 1973. ISSN 0042-6989. doi: https://doi.org/10.1016/0042-6989(73)90094-1. URL https://www.sciencedirect.com/science/article/pii/0042698973900941.

[4] M. Carandini. Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS biology*, 2(9):e264, 2004.

[5] Ilan Dinstein, David J. Heeger, and Marlene Behrmann. Neural variability: friend or foe? *Trends in Cognitive Sciences*, 19(6):322–328, 2015. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2015.04.005. URL https://www.sciencedirect.com/science/article/pii/S1364661315000911.

[6] C Allen and C F Stevens. An evaluation of causes for unreliability of synaptic transmission. *Proceedings of the National Academy of Sciences*, 91(22):10380–10383, 1994. doi: 10.1073/pnas.91.22.10380. URL https://www.pnas.org/doi/abs/10.1073/pnas.91.22.10380.

[7] Michael N. Shadlen and William T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience*, 18 (10):3870–3896, 1998. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.18-10-03870.1998. URL https://www.jneurosci.org/content/18/10/3870.

[8] David M. Schneeweis and Julie L. Schnapf. The photovoltage of macaque cone photoreceptors: Adaptation, noise, and kinetics. *Journal of Neuroscience*, 19(4):1203–1216, 1999. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.19-04-01203.1999. URL https://www.jneurosci.org/content/19/4/1203.

[9] C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996. doi: 10.1126/science.274.5293.1724. URL https://www.science.org/doi/abs/10.1126/science.274.5293.1724.

[10] Colin W.G. Clifford, Michael A. Webster, Garrett B. Stanley, Alan A. Stocker, Adam Kohn, Tatyana O. Sharpee, and Odelia Schwartz. Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131, 2007. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2007.08.023. URL https://www.sciencedirect.com/science/article/pii/S0042698907003756.

[11] Daniel E. Feldman. Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, 32(1):33–55, 2009. doi: 10.1146/annurev.neuro.051508.135516. URL https://doi.org/10.1146/annurev.neuro.051508.135516. PMID: 19400721.

[12] Donald H. Perkel, George L. Gerstein, and George P. Moore. Neuronal spike trains and stochastic point processes: I. the single spike train. *Biophysical Journal*, 7(4):391–418, 1967. ISSN 0006-3495. doi: https://doi.org/10.1016/S0006-3495(67)86596-2. URL https://www.sciencedirect.com/science/article/pii/S0006349567865962.

[13] Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004. URL https://doi.org/10.1152/jn.00697.2004. PMID: 15356183.

[14] Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. In Paul Cisek, Trevor Drew, and John F. Kalaska, editors, *Computational Neuroscience: Theoretical Insights into Brain Function*, volume 165 of *Progress in Brain Research*, pages 493–507. Elsevier, 2007. doi: https://doi.org/10.1016/S0079-6123(06)65031-0. URL https://www.sciencedirect.com/science/article/pii/S0079612306650310.

[15] J. W. Pillow. Time-rescaling methods for the estimation and assessment of non-poisson neural encoding models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1473–1481. MIT Press, 2009.

[16] L. Paninski, E. N. Brown, S. Iyengar, R. E. Kass, Liam Paninski, Emery N Brown, Satish Iyengar, and Robert E Kass. Statistical models of spike trains. *Stochastic methods in neuroscience*, 24: 278–303, 2009.

[17] Alison I. Weber and Jonathan W. Pillow. Capturing the dynamical repertoire of single neurons with generalized linear models. *Neural Computation*, 29(12):3260–3289, 2017. doi: 10.1162/neco_a_01021. URL https://doi.org/10.1162/neco_a_01021.

[18] Henry C. Tuckwell. *Introduction to Theoretical Neurobiology*. Cambridge Studies in Mathematical Biology. Cambridge University Press, 1988.

[19] Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: exploring the neural code*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262181746.

[20] N. V. Swindale and D. E. Mitchell. Comparison of receptive field properties of neurons in area 17 of normal and bilaterally amblyopic cats. *Experimental Brain Research*, 99(3):399–410, Jan 1994. ISSN 1432-1106. doi: 10.1007/BF00228976. URL https://doi.org/10.1007/BF00228976.

[21] Roland Baddeley, L. F. Abbott, Michael C. A. Booth, Frank Sengpiel, Tobe Freeman, Edward A. Wakeman, and Edmund T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1389):1775–1783, 1997. doi: 10.1098/rspb.1997.0246. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1997.0246.

[22] P. Lánskỳ and J. Vaillant. Stochastic model of the overdispersion in the place cell discharge. *Biosystems*, 58(1):27–32, 2000.

[23] Jonathan Pillow and James Scott. Fully bayesian inference for neural models with negative-binomial spiking. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/b55ec28c52d5f6205684a473a2193564-Paper.pdf.

[24] Yuanjun Gao, Lars Busing, Krishna V Shenoy, and John P Cunningham. High-dimensional neural spike train analysis with generalized count linear dynamical systems. In C. Cortes, N. Lawrence,

D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/9996535e07258a7bbfd8b132435c5962-Paper.pdf.

[25] Ian H. Stevenson. Flexible models for spike count data with both over- and under- dispersion. *Journal of Computational Neuroscience*, 41(1):29–43, Aug 2016. ISSN 1573-6873. doi: 10.1007/s10827-016-0603-y. URL https://doi.org/10.1007/s10827-016-0603-y.

[26] Adam S. Charles, Mijung Park, J. Patrick Weller, Gregory D. Horwitz, and Jonathan W. Pillow. Dethroning the Fano Factor: A Flexible, Model-Based Approach to Partitioning Neural Variability. *Neural Computation*, 30(4):1012–1045, 04 2018. ISSN 0899-7667. doi: 10.1162/neco_a_01062. URL https://doi.org/10.1162/neco_a_01062.

[27] R. Barbieri, M. C. Quirk, L. M. Frank, M. A. Wilson, and E. N. Brown. Construction and analysis of non-poisson stimulus-response models of neural spiking activity. *Journal of Neuroscience Methods*, 105(1):25–37, 2001.

[28] Qi She, Xiaoli Wu, Beth Jelfs, Adam S. Charles, and Rosa H. M. Chan. An efficient and flexible spike train model via empirical bayes. *IEEE Transactions on Signal Processing*, 69:3236–3251, 2021. doi: 10.1109/TSP.2021.3076885.

[29] Cina Aghamohammadi, Chandramouli Chandrasekaran, and Tatiana A. Engel. A doubly stochastic renewal framework for partitioning spiking variability. *bioRxiv*, 2024.

[30] A Charles and J Pillow. Additive continuous-time joint partitioning of neural variability. In *2018 Conference on Cognitive Computational Neuroscience. https://doi. org/10.32470/CCN*, 2018.

[31] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X.

[32] Yuan Zhao and Il Memming Park. Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains. *Neural Computation*, 29(5):1293–1316, 05 2017. ISSN 0899-7667. doi: 10.1162/NECO_a_00953. URL https://doi.org/10.1162/NECO_a_00953.

[33] Stephen Keeley, Mikio Aoi, Yiyi Yu, Spencer Smith, and Jonathan W Pillow. Identifying signal and noise structure in neural population activity with gaussian process factor models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13795–13805. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9eed867b73ab1eab60583c9d4a789b1b-Paper.pdf.

[34] A B A Graf, A Kohn, M Jazayeri, and J A Movshon. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neuroscience*, 14(2):239–245, Feb 2011. doi: 10.1038/nn.2733. URL https://doi.org/10.1038/nn.2733.

[35] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. doi: https://doi.org/10.1023/A:1007665907178.

[36] Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100—105, April 2020. ISSN 0028-0836. URL https://doi.org/10.1038/s41586-020-2130-2.

[37] Ramon Bartolo, Richard C. Saunders, Andrew R. Mitz, and Bruno B. Averbeck. Information-limiting correlations in large neural populations. *Journal of Neuroscience*, 40(8):1668–1678, 2020. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2072-19.2019. URL https://www.jneurosci.org/content/40/8/1668.

[38] Anuththara Rupasinghe, Nikolas Francis, Ji Liu, Zac Bowen, Patrick O Kanold, and Behtash Babadi. Direct extraction of signal and noise correlations from two-photon calcium imaging of ensemble neuronal activity. *eLife*, 10:e68046, jun 2021. ISSN 2050-084X. doi: 10.7554/eLife.68046. URL https://doi.org/10.7554/eLife.68046.

[39] James J. Jun, Nick Steinmetz, Joshua H. Siegle, Daniel J. Denman, Marius Bauza, Brian Barbarits, Albert K. Lee, Costas Anastassiou, Alexandru Andrei, Cagatay Aydin, Mladen Barbic, Tim Blanche, Vincent Bonin, Joao Couto, Barundeb Dutta, Sergey Gratiy, Diego Gutnisky, Michael Häusser, Bill Karsh, and Timothy D. Harris. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551:232–236, 11 2017. doi: 10.1038/nature24636.

[40] Jiangang Du, Timothy J. Blanche, Reid R. Harrison, Henry A. Lester, and Sotiris C. Masmanidis. Multiplexed, high density electrophysiology with nanofabricated neural probes. *PLOS ONE*, 6(10): 1–11, 10 2011. doi: 10.1371/journal.pone.0026204.

[41] David Liu and Máté Lengyel. A universal probabilistic spike count model reveals ongoing modulation of neural variability. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Sacha Sokoloski, Amir Aschner, and Ruben Coen-Cagli. Modelling the neural code in large populations of correlated neurons. *eLife*, 10:e64615, oct 2021. ISSN 2050-084X. doi: 10.7554/eLife.64615. URL https://doi.org/10.7554/eLife.64615.

[43] D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955. doi: https://doi.org/10.1111/j.2517-6161.1955.tb00188.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1955.tb00188.x.

[44] A. Aldo Faisal, Luc P. J. Selen, and Daniel M. Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, Apr 2008. ISSN 1471-0048. doi: 10.1038/nrn2258. URL https://doi.org/10.1038/nrn2258.

[45] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[46] Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2016.09.038. URL https://www.sciencedirect.com/science/article/pii/S0896627316306390.

[47] Mark M Churchland, Byron M Yu, Stephen I Ryu, Gopal Santhanam, and Krishna V Shenoy. Neural variability in premotor cortex provides a signature of motor preparation. *J Neurosci*, 26(14): 3697–3712, April 2006.

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

[49] Marlene R. Cohen and John H. R. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12):1594–1600, Dec 2009. ISSN 1546-1726. doi: 10.1038/nn.2439. URL https://doi.org/10.1038/nn.2439.

[50] Matthew J. McGinley, Martin Vinck, Jacob Reimer, Renata Batista-Brito, Edward Zagha, Cathryn R. Cadwell, Andreas S. Tolias, Jessica A. Cardin, and David A. McCormick. Waking state: Rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2015.09.012. URL https://www.sciencedirect.com/science/article/pii/S0896627315007692.

[51] Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.

[52] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press, 2 edition, 2013.

[53] J k. Lindsey. *Modelling Frequency and Count Data*. Oxford University Press, 04 1995. ISBN 9780198523314. doi: 10.1093/oso/9780198523314.001.0001. URL https://doi.org/10.1093/oso/9780198523314.001.0001.

[54] Anuththara Rupasinghe. Continous partitioning of neuronal variability: Matlab codes. https://github.com/Anuththara-Rupasinghe/CMP, 2024.

[55] Mikio C. Aoi and Jonathan W. Pillow. Scalable bayesian inference for high-dimensional neural receptive fields. *bioRxiv*, 2017. doi: 10.1101/212217.

[56] Yung Liang Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

[57] Surya T. Tokdar and Robert E. Kass. Importance sampling: a review. *WIREs Computational Statistics*, 2(1):54–60, 2010. doi: https://doi.org/10.1002/wics.56. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.56.

[58] Persi Diaconis and Donald Ylvisaker. Conjugate Priors for Exponential Families. *The Annals of Statistics*, 7(2):269 – 281, 1979. doi: 10.1214/aos/1176344611. URL https://doi.org/10.1214/aos/1176344611.