Reviewed Preprint
v1 • March 16, 2026
Not revised

Reviewed Preprint
v2 • August 4, 2026
Revised by authors

For correspondence:
ar0621@princeton.edu
pillow@princeton.edu

Competing interests: No competing interests declared

Funding: See page 35

Reviewing editor: Timothy D Hanks, University of California, Davis, United States

Continuous partitioning of neuronal variability

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

[1]Princeton Neuroscience Institute, Princeton University, Princeton, United States • [2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, United States

eLife Assessment

This work of **fundamental** significance introduces a novel statistical model of spiking activity that incorporates continuous-time gain modulation. The authors provide **exceptional** evidence that the model outperforms earlier approaches and alternative candidates in capturing spiking responses across multiple visual areas in the macaque. Beyond its methodological contribution, the study offers new insights into how stimulus-driven variability and internally generated gain fluctuations evolve over time and between brain areas. The framework is likely to find broad application beyond the datasets examined here.

https://doi.org/10.7554/eLife.109719.2.sa3

Abstract

Neurons exhibit substantial trial-to-trial variability in response to repeated stimuli, posing a major challenge for understanding the information content of neural spike trains. In the visual cortex, responses show greater-than-Poisson variability, whose origins and structure remain unclear. To address this puzzle, we introduce a continuous, doubly stochastic model of spike train variability that partitions neural responses into a smooth stimulus-driven component and a time-varying stochastic gain process. We applied this model to spike trains from four visual areas (LGN, V1, V2, and MT) and found that the gain process is well described by an exponentiated power law, with increasing amplitude and slower decay at higher levels of the visual hierarchy. The model also provides analytical expressions for the Fano factor of binned spike counts as a function of timescale, linking observed variability to underlying modulatory dynamics. Together, these results establish a principled framework for characterizing neural variability across cortical processing stages.

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 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 (Figure 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, in the original modulated Poisson formulation [1], the modulatory gain is represented as a scalar random variable associated with a counting window or trial. Thus, while the model provides an effective account of spike count overdispersion, it does not explicitly model gain as a continuously timevarying process within a trial. This limits its ability to characterize how gain fluctuations evolve across subtrial timescales. Related work has further examined the dependence of overdispersion on counting-window duration and related time-dependent extensions of the modulated Poisson framework [31, 32]. Here, we build on this line of work by introducing an explicit continuous-time model of within-trial gain fluctuations.

To clarify this distinction, we start by considering two possible continuous-time interpretations of the modulated Poisson framework of Goris et al. [1]. In both cases, we allow the stimulus component to be a stimulus-dependent, time-varying Poisson firing rate and impose a smoothness prior on this stimulus drive, so that the comparison focuses on different assumptions about the temporal structure of the gain. This smoothness prior distinguishes our implementation from related time-dependent variants of the modulated Poisson framework [31, 32], and provides a regularized estimate of the time-varying stimulus component. We then consider two different assumptions about the temporal sampling of the stochastic gain variable. First, the gain may be constant across the full stimulus presentation or trial (Figure 1B). In this case, the gain is independently drawn for each trial, producing a gain process with broad temporal correlations and a Fano factor (the spike-count variance divided by the mean spike count) that grows approximately linearly with the size of the counting window (Figure 1B, right). As a second alternative, the gain may be resampled independently across time bins on a timescale much shorter than the length of a single trial (Figure 1C). In this case, the gain autocorrelation is narrow, and the Fano factor curve saturates for bin sizes larger than the gain sampling interval (Figure 1C, right).

These continuous-time 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. Such piecewise-constant gain dynamics are unlikely to capture the smooth temporal structure of modulatory fluctuations in biological systems. Moreover, these models have limited ability to capture complex relationships between time bin size and Fano factor, given that they primarily produce curves that grow linearly and then saturate (Figure 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 (Figure 1D). Both components are modeled as exponentiated Gaussian processes (GPs), providing a flexible framework for capturing smooth, continuous dynamics with tunable temporal structure [33–35].

To more precisely 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 described by the standard radial basis (or "Gaussian") kernel. We refer to this as the Exponentiated Power Law (EPL) kernel, which models autocorrelation decay according to an exponentiated power law with exponent $q \leq 2$ (Figure 1E, middle). This added flexibility enables the model to more accurately capture nonlinear relationships between bin size and Fano factor

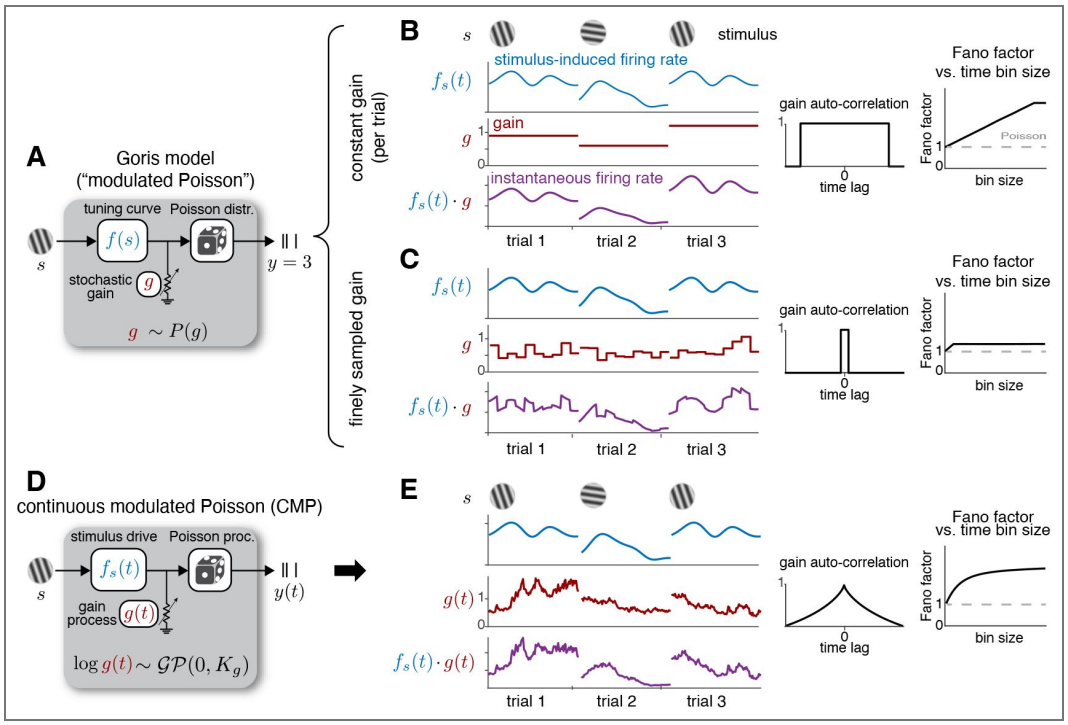


Figure 1.

(A) Schematic of the Goris modulated Poisson model [1], in which spike counts (y) arise from a Poisson process with rate given by 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.

(Figure 1E, right) compared to previous approaches. Moreover, the CMP framework provides closed-form expressions for how the Fano factor scales with time bin size, offering a direct way to probe how variability accumulates and transforms across different temporal resolutions.

We applied the CMP model to spike train recordings from four visual areas (LGN, V1, V2, and MT) [1, 36] 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.

In summary, this study makes five main contributions. (1) We extend classical variability partitioning approaches from spike-count models to a continuous-time spike-train framework, allowing neural variability to be characterized as a function of temporal scale rather than fixed counting windows. (2) We introduce the Exponentiated Power Law (EPL) kernel, which models slow, heavy-tailed correlations in modulatory gain and links them to distinct timescales of variability. (3) We derive analytic expressions for the Fano factor as a function of time bin size, enabling variability to be partitioned across temporal scales within a single unified model. (4) We develop a scalable inference framework that jointly estimates stimulus-driven and modulatory components directly from spike train data. (5) By applying CMP to large-scale recordings across multiple stages of the visual hierarchy, we reveal systematic trends in gain dynamics and stimulus modulation, providing new insights into how variability is structured and propagated across cortical processing stages. Altogether, CMP preserves the interpretability of classical variability-partitioning models while extending them to continuous-time spike-train data, enabling time-resolved analysis of gain dynamics 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 [37], in which the firing rate of a Poisson process is itself a stochastic process. Specifically, we model a spike train $y(t)$ elicited in response to a stimulus s using a Poisson process:

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

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

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

with $f_s(t)$ representing the stimulus-dependent component that captures stimulus-locked fluctuations in firing rate, and $g(t)$ denoting a stochastic gain process that accounts for stimulus independent fluctuations in neuronal excitability (Figure 1D-E).

We assume that the stimulus component $f_s(t)$ varies independently across stimuli and evolves smoothly in time. To express this smoothness, we place a Gaussian process (GP) prior over the log firing rate:

$$\log f_s(t) \sim \mathcal{GP}\left(0, K_f^{(s)}\right) \tag{3}$$

where $K_f^{(s)}$ is the standard radial basis function (RBF) covariance function, which describes the covariance of fluctuations in log firing rate at arbitrary points in time [33]. 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:

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

where $\rho_f^{(s)}$ and $\ell_f^{(s)}$ are hyperparameters controlling, respectively, the variance and temporal smoothness (length scale) of fluctuations in the log firing rate. To flexibly capture diverse tuning profiles, each stimulus condition is assigned its own set of stimulus-related hyperparameters.

The gain process $g(t)$ is modeled analogously as 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 is a latent variable that varies independently across repeated presentations of the same stimulus, thereby inducing greater-than-Poisson variability in the spike counts. For a fixed stimulus $s = s^*$ presented N times, the model includes a shared stimulus component $f_{s^*}(t)$ and an independent gain process $g_i(t)$ for each repeat $i \in \{1, \ldots, N\}$. Also, in contrast to the stimulus drive, because the gain process reflects stimulus-independent variability, its hyperparameters are shared across all stimulus conditions. This formulation provides a unified and parsimonious framework for modeling trial-to-trial variability across stimuli within a single generative model.

To more accurately model the stochastic structure of real neural spike trains, we introduce a novel covariance function for the gain process, referred 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 temporal length scale, and q is the power-law exponent controlling the rate of correlation decay over long timescales. We show in the Appendix 1 that this kernel defines a valid covariance function for $0 < q \leq 2$. Because the gain reflects stimulusindependent variability, its hyperparameters are shared across all stimulus conditions. This formulation provides a unified and parsimonious generative model that captures both smooth, stimulus-locked activity and structured trial-to-trial variability across stimuli.

Figure 2 illustrates example gain covariance functions under different settings of the hyperparameters ℓ_g (length scale) and q (power-law exponent). When $q = 2$ (right column), the EPL kernel reduces to the standard RBF covariance function defined in Equation 4. For $q < 2$, the correlations exhibit heavier tails than in the RBF case, decaying more slowly at long time lags. In the limiting case of $\ell_g \to \infty$ or $q \to 0$, the gain becomes effectively constant over each stimulus presentation interval [0, T], corresponding to the continuous-time Goris model shown in Figure 1B.

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

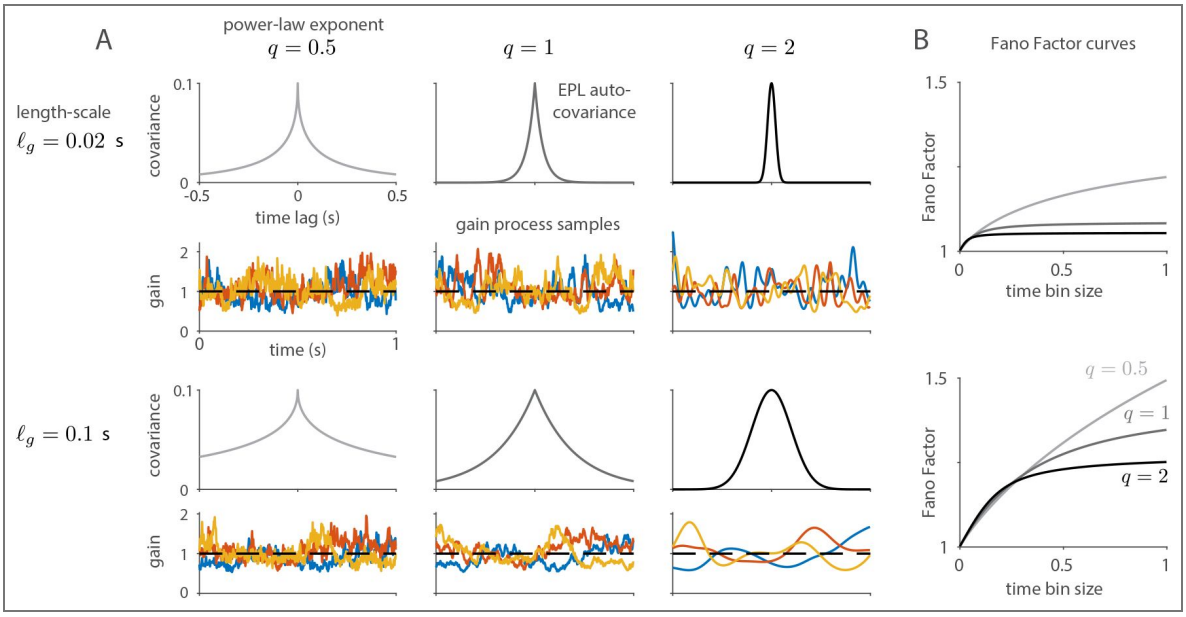


Figure 2. Illustration of different settings of the proposed Exponentiated Power Law (EPL) covariance function with marginal variance ρ_g = 0.1.

(A) EPL auto-covariance functions with different settings of length-scale (top row: ℓ_g = 0.02, bottom row: ℓ_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.

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:

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

where λ denotes the mean firing rate under the model. The spike count variance is then given by:

$$\mathrm{var}\left[y_\Delta \mid s\right] = \Delta\lambda + (\Delta\lambda)^2(\kappa(\Delta) - 1) \tag{8}$$

where

$$\kappa(\Delta) = \frac{1}{\Delta^2}\int_0^\Delta \int_0^\Delta e^{K_g(t_1,t_2)}dt_1dt_2, \tag{9}$$

is the average of the exponentiated covariance function exp(K_g) over all pairs of times in the interval [0, Δ]. (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:

$$FF(\Delta) = 1 + \Delta\lambda(\kappa(\Delta) - 1). \tag{10}$$

This shows that the Fano Factor approaches 1 in the limit of small bin size Δ, low firing rate λ, or when κ(Δ) approaches 1, which occurs when the variance of the gain process ρ_g goes to zero. Note that κ(Δ) is a decreasing function of Δ, reflecting the decay of correlations in the gain process, and it remains constant only for a gain process with infinite lengthscale (for example, Figure 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 and Materials (see Appendix 2 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 $g(t)$:

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

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 + Δ] given the stimulus s and gain process $g(t)$, and $\mu(g(t))$ is the probabilistic measure of the gain process under the Gaussian process prior (Equation 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 [38], 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 and Materials). 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, as shown in Figure 3. We simulated spiking activity over a duration of $T = 5$ s for $S = 3$ stimulus conditions, with $N = 50$ repeated trials per stimulus, following the CMP generative model. We then fit the CMP model to the simulated spike trains using a temporal grid size of $dt = 0.05$ s and estimated both the latent processes and GP hyperparameters (see Methods and Materials for details). Figure 3A illustrates the true and inferred stimulus-drive and gain processes for two representative trials from two different stimulus conditions, along with the corresponding firing rates and observed spike counts. Figure 3B compares the true and inferred Gaussian process hyperparameters. These results show that the proposed inference framework accurately recovers the ground-truth stimulus drives, gain processes, and GP hyperparameters on a discretized time grid. 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 [36]). These datasets were collected from three monkeys using drifting grating stimuli presented at $S = 72$ 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).

Figure 4A 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, from 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, followed by the cross-trial gain mean and cross-trial gain variance. 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. Importantly, the cross-trial gain mean is relatively flat, and does not resemble the inferred stimulus drive or exhibit clear stimulus-locked temporal structure. This suggests that trialspecific gain fluctuations largely average out across repeats, consistent with the interpretation that the gain process captures random trial-to-trial variability rather than stimulus-driven activity. In contrast, the cross-trial gain variance shows a general reduction during the stimulus presentation compared to the inter-trial interval, indicating that gain variability is quenched during stimulus presentation. These observations indicate that the CMP model meaningfully partitions response variability across stimulus conditions. Figure 4B shows the estimated EPL gain covariance, corresponding to optimal hyperparameters $\rho_g = 2.9$, $\ell_g = 0.93$, and $q = 0.95$.

To further examine the temporal structure of inferred gain variability, we quantified cross-trial gain variance before and after stimulus onset. Following the approach of Churchland et al. [39], we compared gain variability in two matched 400-ms windows: a pre-stimulus window ending at stimulus onset and a stimulus-period window beginning 100 ms after stimulus onset. For each neuron and stimulus condition, we computed the cross-trial variance of the inferred gain at each time bin and averaged this quantity within each window. We then compared pre- and post-stimulus gain variability across neuron-stimulus pairs (Figure 4C). Across the 654 V1 neurons, post-stimulus gain variance was significantly lower than pre-stimulus gain variance (Figure 4C; one-sided paired Wilcoxon signed-rank test, $p \leq 10^{-15}$). This reduction is consistent with gain variability quenching following stimulus onset [39, 40].

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 Methods and Materials for details). The bottom row of Figure

Figure 3. Simulation-based validation of CMP inference.

(A) From top to bottom: true and inferred stimulus-driven components for stimulus 1 and stimulus 3; true and inferred gain processes for two representative trials from each stimulus condition; corresponding true and inferred firing rates; and observed spike counts. **(B)** Comparison of true and inferred gain hyperparameters ℓ_g, ρ_g, q and stimulus-drive hyperparameters $\ell_f^{(1:3)}$, $\rho_f^{(1:3)}$.

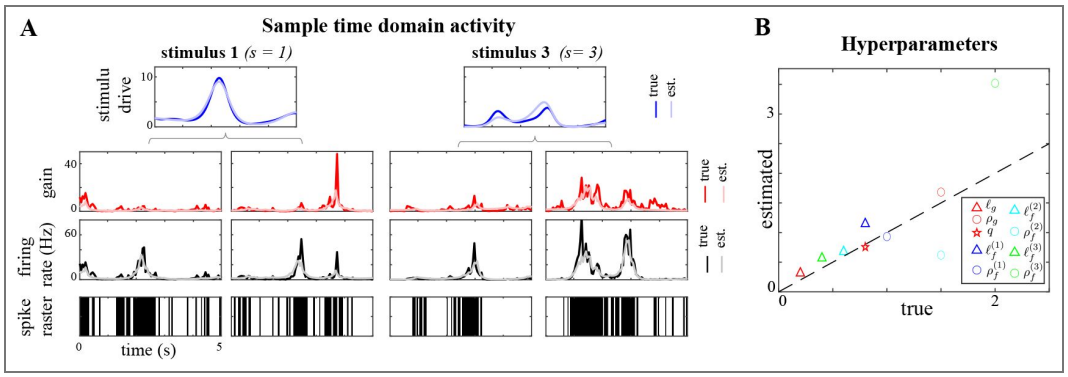
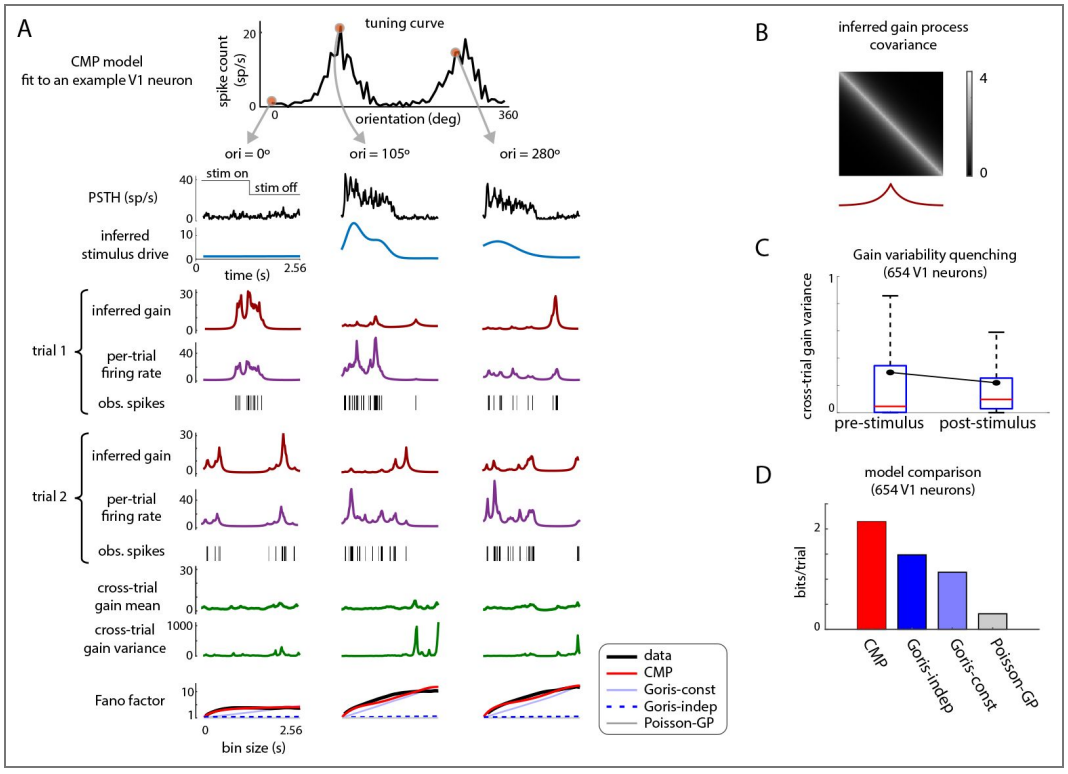


Figure 4. Performance comparisons of different models on neural population data recorded from macaque primary visual cortex under 72 drifting sinusoidal gratings (data from Graf et al. [36]).

(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), cross-trial gain mean, cross-trial gain variance 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)** Comparison of pre-stimulus and post-stimulus cross-trial gain variance across V1 neurons. Boxes show the distribution across neurons, and black dots indicate the mean. **(D)** Average test 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.



4A 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 Poisson-GP model corresponds to the CMP model without a gain component, meaning that it uses an inhomogeneous Poisson observation model with a stimulus-specific, GP-smoothed time-varying firing rate. This differs from the baseline Poisson model, which estimates the time-dependent firing rate using a stimulus-conditioned PSTH with a Gamma prior for regularization in low-repeat conditions (see Methods and Materials for details of all model variants). 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. For both Goris variants shown here, the stimulus component was set to the GP-smoothed firing-rate estimate obtained from the Poisson-GP model, so that differences across these baselines reflect the assumed temporal structure of the gain process rather than differences in stimulus-drive estimation. 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 (Figure 4D). The CMP model achieved the highest predictive performance across all neurons. We also use these comparisons to separate the contribution of the GP-smoothed stimulus drive from the contribution of the stochastic gain process: the Poisson-GP model tests the effect of smoothing the stimulus-driven firing rate alone, whereas the CMP model additionally includes continuous-time gain fluctuations. In Figure S1, we further compared the CMP model to three variants with different gain kernels: CMP-RBF (where the EPL kernel in Equation 6 is replaced by the standard RBF kernel [33]), 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 $s = 1, \ldots, S$ stimulus conditions, denoted by $\left\{\rho_g^{(s)}, \ell_g^{(s)}, q^{(s)}\right\}$. Figure S2A shows the inferred activity under both models, and Figure S2B 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 assessed the consistency of CMP inference with respect to temporal discretization by comparing inference results across three sufficiently fine grid resolutions: $dt = 0.02$ s, $dt = 0.05$ s, and $dt = 0.1$ s. As shown in Figure S3A, the estimated gain hyperparameters remain stable across these resolutions, indicating that the inference procedure is robust to moderate changes in grid size once the temporal discretization is fine enough to approximate the underlying continuous-time model. Figure S3B shows corresponding test log-likelihoods across models and grid resolutions, confirming that CMP consistently outperforms alternative models over this range.

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 Goris et al. [1]). The recordings were obtained using drifting sinusoidal gratings of the preferred size and speed, varying either in $S = 12$ spatial frequency (ranging from 0 to 10 cycles/deg) or in $S = 16$ 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.

Figure 5A 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 models across the visual hierarchy. To examine performance at the single-neuron level, Figure 5B 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. Figure 5C 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 Figure S4, 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-style models with and without this component (Figure S5). All Goris-style results reported above use versions with a GP prior on the stimulus drive, in which the stimulus-dependent firing rates were set to the smooth firing-rate estimates obtained from the Poisson-GP model, providing a stronger GP-smoothed stimulus-drive estimate for these baselines. In the versions without the GP prior, the stimulus-drive parameters were instead estimated directly under the corresponding Goris-style likelihood, making these models closer to the original modulated Poisson formulation [1] in their treatment of the stimulus drive. Figure S5 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 the Poisson-GP model, despite lacking a stochastic gain component, outperforms both variants of the Goris-style model when the GP prior on the stimulus drive is removed (Figure S5A–C). These results indicate that the GP prior on the stimulus drive provides an important improvement in model fit, while the additional continuous-time gain component in CMP further improves the model's ability to capture overdispersion and trial-to-trial variability (Figure S5D).

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 Figure 6. 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. Figure 6D presents the median gain covariance and its corresponding autocorrelation function for each area.

The gain power-law exponent (Figure 6A) 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 (Figure 6B) 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 (Figure 6D) 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 (Figure 6C) 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: $\bar{\rho}_f = \frac{1}{s}\sum_{s=1}^{S} \rho_f^{(s)}$, as shown in Figure 6E. 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

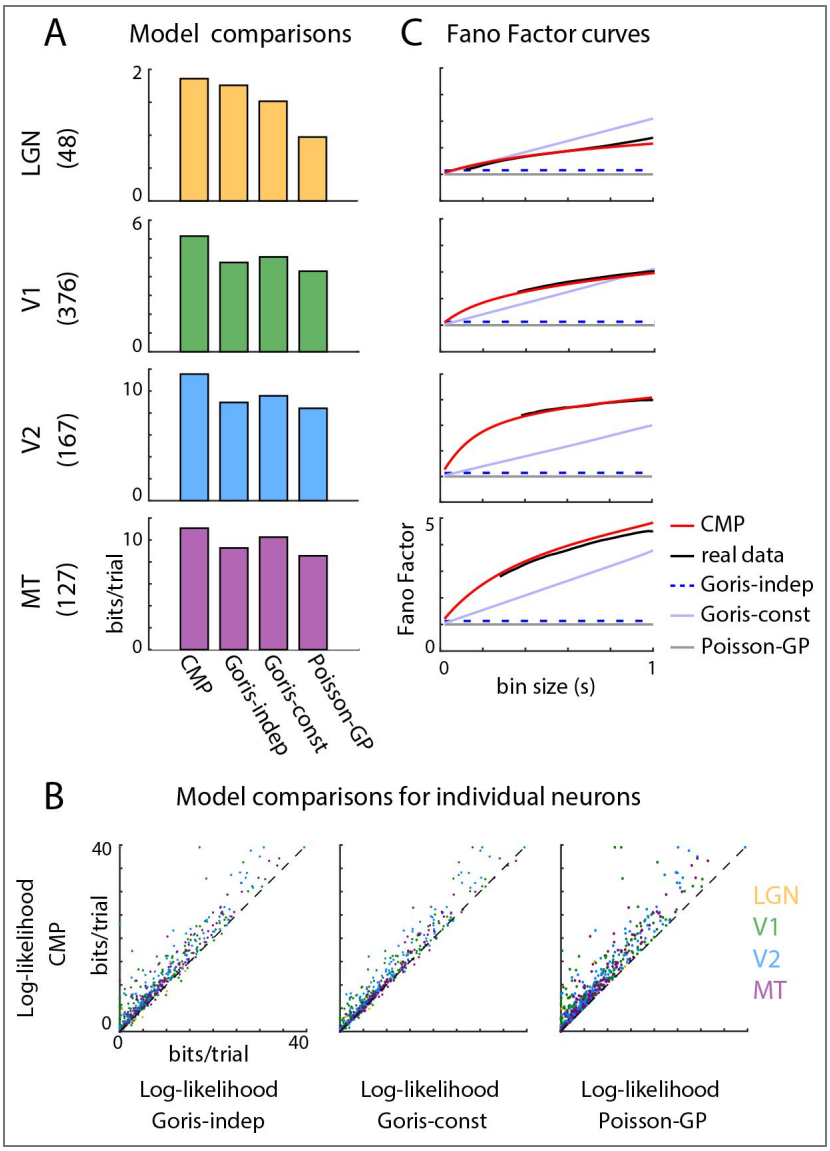


Figure 5. Performance comparisons of different models on neuronal population data recorded from different areas along the visual hierarchy—the lateral geniculate nucleus (LGN), V1, V2 and MT—under drifting sinusoidal gratings of preferred size and speed, varying in spatial frequency (12 frequencies from 0 to 10 cycles/deg) or in drift direction (16 directions) (data from Goris et al. [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.

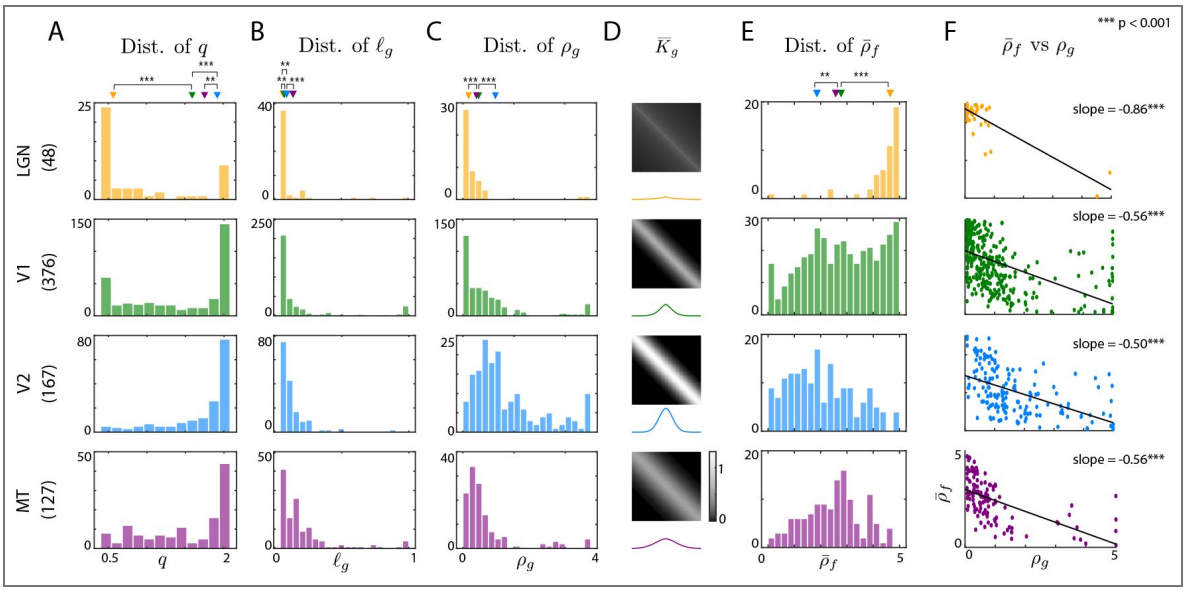


Figure 6. Comparison of inferred CMP hyperparameters across visual areas: the lateral geniculate nucleus (LGN), V1, V2, and MT (data from [1]).

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 K_g across the neuronal population (top) and gain auto-correlation (bottom). **(E)** Distribution of the average stimulus-drive variance $\left(\bar{\rho}_f = \frac{1}{s}\sum_{s=1}^{S}\rho_f^{(s)}\right)$ across neurons. **(F)** Scatter plots of $\bar{\rho}_f$ versus ρ_g across neurons, with linear regression fits (black) and slopes annotated. $^{***}p < 0.001$.

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, Figure 6F shows scatter 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 [35, 41–43].

Discussion

Advances in high-density electrophysiological recording techniques [44, 45] now allow large-scale measurements of neuronal activity across multiple brain areas and experimental conditions. Analyses of sensory responses to repeated stimuli consistently reveal substantial trial-to-trial variability in neural spiking, arising from diverse sources such as synaptic transmission noise, fluctuations in excitation–inhibition balance, noisy peripheral inputs, recurrent dynamics, adaptation, and synaptic plasticity [6–11]. Capturing this variability is essential for understanding neural coding [12–14]. Traditional Poisson spiking models [18, 19] provide a foundational framework but predict spike count variance equal to the mean. This assumption is often violated in many brain regions where neurons exhibit overdispersion, meaning variability that exceeds the Poisson prediction [1]. Although numerous models have been proposed to account for this [23–29, 34, 35, 46, 47], few provide a detailed, time-resolved characterization 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 [37], 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 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. In our application, the EPL kernel consistently outperformed standard alternatives such as the RBF and Matérn kernels. In addition, we developed a scalable variational inference method that jointly estimates the latent stimulus-driven and stochastic gain components, along with their hyperparameters, directly from spiking data. We validated this approach using synthetic experiments. When applied to recordings from macaque V1 [36], 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 [35, 41–43]. 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 [48]. They also align with broader theories in systems neuroscience, such as efficient coding [49] and probabilistic inference [50], 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, 36, 51]. The CMP model preserves the interpretability of classical variability-partitioning approaches while extending them from spike-count models to a continuous-time spike-train framework. This enables time-resolved inference of stimulus-driven and stimulus-independent components, analytic characterization of Fano factor scaling with bin size, and direct estimation of the temporal covariance structure of gain fluctuations. While we found that the 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. While the EPL kernel provides a flexible and tractable way to capture a wide range of temporal correlation structures in the stochastic gain process, its generative interpretation depends on the parameter q. For $q = 1$, a straightforward stochastic generator is a simple low-pass filtered Gaussian white noise process, equivalent to an Ornstein–Uhlenbeck (OU) process [52, 53]. For $q < 1$ there is no known closed-form stochastic differential equation that exactly produces the EPL covariance, but such long-memory dynamics can be closely approximated by superpositions of OU processes with a spectrum of time constants or by fractional stochastic processes such as fractional Brownian motion [54]. Conversely, when $q > 1$, correlations decay more rapidly than exponentially, corresponding to short-memory dynamics with diminished low-frequency power; these can often be approximated by higher-order linear filters or by driving OU processes with temporally correlated noise. Together, these interpretations highlight how the parameter q shapes the temporal structure of trial-to-trial variability.

An important direction for future work is to move beyond descriptive modeling and explore the biological interpretability of inferred hyperparameters. For instance, differences in q or ℓ_g across the visual hierarchy could reflect distinct circuit or cellular mechanisms, such as changes in synaptic integration timescales, recurrent dynamics, or bursting behavior. This perspective could help explain phenomena such as the disproportionately slow gain fluctuations we observe in the LGN (e.g., $q \approx 0.5$), potentially linking them to specific biophysical or network-level processes. Establishing such connections would deepen our understanding of how structured variability arises and how it shapes sensory computation.

Mechanistic interpretations of structured variability may also benefit from parallels with other biological systems. For example, variability in bacterial flagellar motor switching and single-cell gene expression can be explained by a slow extrinsic process modulating a faster Poisson process, producing heavy-tailed dwell times or variance–mean scaling beyond Poisson expectations [55, 56]. This two-timescale structure closely parallels the CMP framework and suggests that future work incorporating explicit modulatory dynamics could offer deeper mechanistic insight into neural variability and its functional significance.

The present CMP framework captures Poisson and super-Poisson variability through a conditionally Poisson observation model with stochastic gain modulation. As a result, the model does not generate sub-Poisson spike count variability in its current form. Sub-Poisson variability has been observed in some neural systems and may arise from mechanisms such as refractoriness, spike-history dependence, regularized spiking, or other non-Poisson renewal structure [57–59]. Extending CMP to capture such effects would require modifying the observation model, for example by incorporating spike-history terms, renewal-process likelihoods, or

alternative count distributions that allow variance to fall below the mean [13, 14, 29]. Such extensions could make it possible to partition both super-Poisson and sub-Poisson components of neural variability in continuous time and to examine how these components vary with behavioral state, attention, or arousal [60, 61].

Beyond these conceptual extensions, there are also several methodological directions for future research. First, extending CMP to high-dimensional population activity may allow joint modeling of shared variability across neurons. A natural direction would be to combine the CMP framework with ideas from Gaussian Process Factor Analysis [35] to introduce low-dimensional shared gain activity across neurons. Such an extension would allow CMP to capture correlated variability and population-wide modulatory dynamics in simultaneously recorded neural ensembles. Second, adapting the model to alternative link functions [26] could broaden its applicability to different statistical structures. Third, developing inference algorithms compatible with calcium imaging data [43] would extend its utility to optical recordings. Fourth, integrating behavioral covariates [62] or state variables such as attention [60] or arousal [61] could improve inference of nonstationary trial structure. Finally, while our focus here was on neural spike trains, the general formulation of the CMP model could be applied to other domains involving overdispersed count data [63–65].

To facilitate reproducibility and dissemination, we have released our implementation of the CMP model on GitHub [66]. In summary, the 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, while opening the door to deeper mechanistic interpretations of variability in neural systems.

Materials and methods

CMP model inference

We start by recalling the general CMP forward model:

$$y(t) \mid s \sim \mathrm{PoissProc}\left(f_s(t) \cdot g(t)\right),$$

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

$$\log g(t) \sim \mathcal{GP}\left(0, K_g\right), K_g\left(t_1, t_2\right) = \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_1 - t_2}{\ell_g}\right|^q\right).$$

In our inference framework, we consider a setting with multiple repeats (n = 1, ..., N) for a given stimulus condition, with data potentially collected across different stimulus conditions (s = 1, ..., S). We perform inference on a uniform grid of time points (t = 1, ..., 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 $y_{t,s,n}$ denote the spike count for the t^{th} time bin, s^{th} stimulus condition, and n^{th} repeat. Further, let $\mathbf{f}_s = [f_{1,s} \ldots f_{T,s}]^\top$ and $\mathbf{g}_{s,n} = [g_{1,s,n} \ldots g_{T,s,n}]^\top$ denote the corresponding stimulus-driven and stochastic gain processes. Further, let $\mathbf{f}_s = \log \mathbf{f}_s$ and $\mathbf{g}_{s,n} = \log \mathbf{g}_{s,n}$ represent the log-transformed stimulus-driven and gain components, for notational brevity. Note that, following the CMP generative model, the latent variables $\mathbf{f} = \mathbf{f}_{1:S}$ and $\mathbf{g} = \mathbf{g}_{1:S,1:N}$, along with the hyperparameters $\theta = \left\{\rho_g, \ell_g, q, \rho_f^{(s)}, \ell_f^{(s)}, s = 1, \ldots, S\right\}$, are all unknown quantities that need to be estimated given the spiking observations $\mathbf{y} = y_{1:T,1:S,1:N}$.

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 [35, 67]. Motivated by this, we adopt a frequency-domain representation of the stimulus-driven components (**f**) in our inference procedure. Accordingly, we relate the time-domain components to their corresponding Fourier-domain components:

$$\mathbf{f_s} = \mathbf{B}^{(s)}\mathbf{w_s}$$

where $\mathbf{B}^{(s)} \in \mathbb{R}^{T\times M_s}$ (noting that typically $M_s \ll T$ due to pruning of unnecessary Fourier coefficients that do not substantially contribute to explaining variability in $\mathbf{f}_s$ [35, 67]) represents the standard real orthonormal inverse discrete Fourier transform matrix. Then, with the diagonalization of the RBF kernel [35, 67], the Fourier-transformed stimulus-driven components $\mathbf{w}_s = [w_{1,s}, w_{2,s}, \dots, w_{M_s,s}]'$ are distributed as independent Gaussian random variables with the following parameters:

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

where $\tilde{k}_m^{(s)} = \sqrt{2\pi}\rho_f^{(s)}\ell_f^{(s)}\exp\left(-\frac{1}{2}f_m^{(s)^2}\ell_f^{(s)^2}\right)$, with $f_m^{(s)}$ denoting the m^{th} Fourier frequency [35, 67]. Thus, we use the frequency-domain stimulus-driven components $\mathbf{w} = \mathbf{w_{1:S}}$ in place of their time-domain counterparts $\mathbf{f}$, reducing computational complexity by avoiding matrix inversions and enabling pruning ($M_s \ll T$). 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) [38], 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 [34]:

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

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 $s = 1, \dots, S$ and trials $n = 1, \dots, N$:

$$\begin{aligned}q(\mathbf{w},\mathbf{g}) &= \prod_{s=1}^{S}\left(q(\mathbf{w_s})\prod_{n=1}^{N}q(\mathbf{g_{s,n}})\right)\\ q(\tilde{\mathbf{g}}_{\mathbf{s,n}}) &\sim \mathcal{N}(\boldsymbol{\mu}_{s,n}, \boldsymbol{\Sigma}_{s,n}) \text{ and } q(\mathbf{w_s}) \sim \mathcal{N}(\boldsymbol{\nu}_s, \boldsymbol{\Omega}_s),\end{aligned}$$

where $\boldsymbol{\Sigma}_{s,n} \in \mathbb{R}^{T\times T}$ and $\boldsymbol{\Omega}_s \in \mathbb{R}^{M_s\times M_s}$ are full covariance matrices. Note that, by the property of linear transformations of Gaussian random variables [68], this in turn implies that the variational distributions over the time-domain stimulus-driven components are also Gaussian:

$$q(\tilde{\mathbf{f}}_\mathbf{s}) = q(\mathbf{B}^{(s)}\mathbf{w_s}) \sim \mathcal{N}(\mathbf{B}^{(s)}\boldsymbol{\nu}_s, \mathbf{B}^{(s)}\boldsymbol{\Omega}_s\mathbf{B}^{(s)'})$$

Algorithm 1.

Joint inference of variational parameters and hyperparameters

Inputs: Spiking observations $y_{1:T,1:S,1:N}$
Outputs: Estimated hyperparameters $\theta = \{\rho_g, \ell_g, q, \rho_f^{(1:S)}, \ell_f^{(1:S)}\}$, and variational parameters $\boldsymbol{\nu}_{1:S}$, $\boldsymbol{\Omega}_{1:S}$, $\boldsymbol{\mu}_{1:s,1:N}$, and $\boldsymbol{\Sigma}_{1:s,1:N}$
Initialization: $\delta = 1$, $\theta_g = \{\rho_g = 0.1, \ell_g = 0.1, q = 1\}$, $\theta_f^{(1:S)} = \{\rho_f^{(1:S)} = 0.1, \ell_f^{(1:S)} = 1\}$
while $\delta > 10^{-5}$ **do**
  $\theta^{\text{prev}} = \theta$
  Fix $\theta_g = \{\rho_g, \ell_g, q\}$
  for $s = 1, \ldots, S$ **do**
    Optimize $\theta_f^{(s)} = \{\rho_f^{(s)}, \ell_f^{(s)}\}$ by maximizing the ELBO in Equation 12, using Algorithm 2 to derive the conditional variational parameters
  Fix $\theta_f^{(1:S)} = \{\rho_f^{(1:S)}, \ell_f^{(1:S)}\}$
  Optimize $\theta_g = \{\rho_g, \ell_g, q\}$ by maximizing the ELBO in Equation 12, using Algorithm 2 to derive the conditional variational parameters
  $\delta = ||\theta - \theta^{\text{prev}}||/||\theta||$
for $s = 1, \ldots, S$ **do**
  Estimate the final settings of $\boldsymbol{\nu}_s$, $\boldsymbol{\Omega}_s$, $\boldsymbol{\mu}_{s,1:N}$, and $\boldsymbol{\Sigma}_{s,1:N}$ based on the final estimates of the hyperparameters $\theta^{(s)} = \{\theta_g, \theta_f^{(s)}\}$ using Algorithm 2

Accordingly, the ELBO can be derived as:

$$
\begin{aligned}
\mathscr{L}(q) =& \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{w}, \mathbf{g} \mid \theta)] - \mathbb{E}_q[\log q(\mathbf{w}, \mathbf{g})] \\
=& -\sum_{t,s,n} dt \exp\left(\left(\mathbf{B}^{(s)}\boldsymbol{\nu}_s\right)_t + \left(\mathbf{B}^{(s)}\boldsymbol{\Omega}_s\mathbf{B}^{(s)\top}\right)_{t,t}/2 \right. \\
& \left. + (\boldsymbol{\mu}_{s,n})_t + (\boldsymbol{\Sigma}_{s,n})_{t,t}/2\right) \\
& + \sum_{t,s,n} y_{t,s,n}\left(\left(\mathbf{B}^{(s)}\boldsymbol{\nu}_s\right)_t + (\boldsymbol{\mu}_{s,n})_t\right) - \frac{SN}{2}\log|K_g| \\
& - \frac{1}{2}\sum_{s,n} \operatorname{trace}\left(K_g^{-1}\left(\boldsymbol{\mu}_{s,n}\boldsymbol{\mu}_{s,n}^{\top} + \boldsymbol{\Sigma}_{s,n}\right)\right) \\
& - \frac{1}{2}\sum_{s,m}\log\tilde{k}_m^{(s)} - \frac{1}{2}\sum_{s,m}\left(\tilde{k}_m^{(s)}\right)^{-1}\left(\nu_{m,s}^2 + (\boldsymbol{\Omega}_s)_{m,m}\right) \\
& + \sum_{s,n}\frac{1}{2}\log|\boldsymbol{\Sigma}_{s,n}| + \sum_s \frac{1}{2}\log|\boldsymbol{\Omega}_s| \\
& + \frac{1}{2}\sum_s M_s + \text{ const.}
\end{aligned}
\tag{12}
$$

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 $\theta_f^{(s)} = \left\{\rho_f^{(s)}, \ell_f^{(s)}\right\}$ for s = 1, ..., S, and the gain hyperparameters $\theta_g = \{\rho_g, \ell_g, q\}$, until convergence. Each optimization maximizes the ELBO in Equation 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:, $\rho_g, \rho_f^{(1:S)} \in [0.05, 6]$, $\ell_g, \ell_f^{(1:S)} \in [0.05, 10]$ and $q \in [0.4, 2]$.

Algorithm 2.

Estimation of variational parameters given hyperparameters

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

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

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 $\tilde{k}^{(s)}_{1:Ms}$ are known), as summarized in Algorithm 2. Note that the ELBO in Equation 12 is independent across stimulus conditions s = 1, ..., S, when conditioned on θ. Therefore, we focus on estimating the variational parameters corresponding to the s^{th} stimulus: $\boldsymbol{\nu}_s$, $\boldsymbol{\Omega}_s$, $\boldsymbol{\mu}_{s,1:N}$, and $\boldsymbol{\Sigma}_{s,1:N}$. We maximize the ELBO to find the optimal settings of these variational parameters, alternating between updating the stimulus-drive variational parameters ($\boldsymbol{\nu}_s$, $\boldsymbol{\Omega}_s$) and the gain variational parameters ($\boldsymbol{\mu}_{s,1:N}$, $\boldsymbol{\Sigma}_{s,1:N}$) until convergence. Each maximization step uses a Newton–Raphson procedure.

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

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

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

$$\nabla_{\boldsymbol{\Omega}_s}\mathscr{L} = -\tfrac{1}{2}\sum_n \mathbf{B}^{(s)\top}\widehat{\boldsymbol{\Lambda}}_{s,n}\mathbf{B}^{(s)} + \tfrac{1}{2}\left(\boldsymbol{\Omega}_s\right)^{-1} - \tfrac{1}{2}\left(\widetilde{\mathbf{K}}^{(s)}\right)^{-1}$$

and setting this derivative to zero results in:

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

Accordingly, we update the stimulus-drive variational mean $\boldsymbol{\nu}_s$ using Newton–Raphson optimization with the gradient and Hessian in Equation 13, and subsequently update the stimulus-drive variational covariance $\boldsymbol{\Omega}_s$ using the Hessian as in Equation 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 $\boldsymbol{\mu}_{s,n}$ given by:

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

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

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

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,

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

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

Observing that the integral in Equation 16 is intractable, we used importance sampling [69] to approximate it. We employed the variational distribution $q\left(\tilde{\mathbf{g}}_{\mathbf{s,n}}\right)$ as the proposal distribution,

$$p\left(\mathbf{y}_{s,n} \mid \mathbf{B}_s\right) = \mathbb{E}_{q(\mathbf{g}_{s,n})}\left[p\left(\mathbf{y}_{s,n} \mid \mathbf{g}_{s,n}, s\right) \frac{p(\mathbf{g}_{s,n})}{q(\mathbf{g}_{s,n})}\right]$$
$$\approx \frac{1}{I} \sum_{i=1}^{I} p\left(\mathbf{y}_{s,n} \mid \mathbf{g}_{s,n}^{(i)}, s\right) \frac{p\left(\mathbf{g}_{s,n}^{(i)}\right)}{q\left(\mathbf{g}_{s,n}^{(i)}\right)}$$

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

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

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

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

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

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

and the variance is:

$$\begin{aligned}\text{var}\left[y_\Delta \mid s\right] = & e^{\rho_g/2}\int_0^\Delta f_s(t)dt \\ & + e^{\rho_g}\int_0^\Delta\int_0^\Delta f_s\left(t_1\right) f_s\left(t_2\right)\left(e^{K_g(t_1,t_2)}-1\right)dt_1dt_2\end{aligned}$$

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

$$FF(\Delta) = 1 + \frac{e^{\rho_g/2}\int_0^\Delta\int_0^\Delta f_s(t_1)f_s(t_2)\left(e^{K_g(t_1,t_2)}-1\right)dt_1dt_2}{\int_0^\Delta f_s(t)dt}$$

The details of the derivation are provided in the Appendix 2. 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 duration Δ, shift it systematically through the entire 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:

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

and, to compensate for sparsity, we further assumed a Gamma prior on the firing rate $\lambda_{t,s} \sim$ Gamma (α, β), where α and β are hyperparameters. Since the Gamma distribution is conjugate to the Poisson distribution [70], the posterior distribution is also Gamma-distributed:

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

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

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

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 $\lambda_{t,s}$ 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:

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

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 model in Goris et al. [1] considered spike counts within a single time bin. For spiking observations from the s^{th} stimulus presentation with $t = 1, \ldots, T$ time bins, we assumed that the stimulus-independent gain is independent across time bins with shared variance $\sigma_{g_s}^2$, i.e.,

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

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

$$\begin{aligned} p\left(y_{1:T,s,1:N} \mid r_s, s_s\right) &= \int_0^\infty p\left(y_{1:T,s,1:N}, \lambda_{1:T,s,1:N} \mid r_s, s_s\right) d\lambda_{1:T,s,1:N} \\ &= \prod_{t,n=1}^{T,N} \text{NegativeBinomial}\left(y_{t,s,n} \mid r_s, \frac{1}{1+s_{t,s}}\right) \end{aligned} \tag{17}$$

where $s_{t,s} = \sigma_{g_s}^2 f_{t,s} dt$. For the Goris-indep baseline used in the main model comparisons, we set the stimulus component to the smooth firing-rate estimate obtained from the Poisson-GP model. Specifically, we set $s_{t,s} = \sigma_{g_s}^2 f_{t,s} dt = f_{t,s} dt / r_s = \hat{\lambda}_{t,s} / r_s$, where $\hat{\lambda}_{t,s}$ denotes the Poisson-GP estimate of the stimulus-dependent firing rate. For $s = 1, \ldots, S$, we searched for the values of r_s that minimize the negative log-likelihood of the spiking observations $y_{1:T,s,1:N}$. Details of Equation 17, the Fano factor, and the version without a smoothness prior on the stimulus drive are provided in the Appendix 3.

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:

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

where $g_{s,n} \sim \text{Gamma}(r_s, s_s)$ with shape parameter $r_s = \frac{1}{\sigma_{g_s}^2}$ and scale parameter $s_s = \sigma_{g_s}^2$. Further, we assumed that the total spike count across all time bins in a trial, $Y_{s,n} = \sum_t y_{t,s,n}$, also shares the same gain variance $\sigma_{g_s}^2$, i.e, $Y_{s,n} \sim \text{Poisson}\left(\tilde{g}_{s,n} f_s dt T\right)$ with $\tilde{g}_{s,n} \sim \text{Gamma}\left(r_s, s_s\right)$. For the Gorisconst baseline used in the main model comparisons, we estimated $\sigma_{g_s}^2$ by fitting $Y_{s,n}$ across training data $n = 1, \ldots, N$, as in Goris et al. [1], and set the stimulus encoding component $f_{t,s} dt = \hat{\lambda}_{t,s}$, where $\hat{\lambda}_{t,s}$ denotes the smooth firing-rate estimate obtained from the Poisson-GP model. Details of the mean-variance relationship, log-likelihood derivation, and the version without a smoothness prior on the stimulus drive are provided in the Appendix 3.

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 [36]. 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.

Supplementary figures

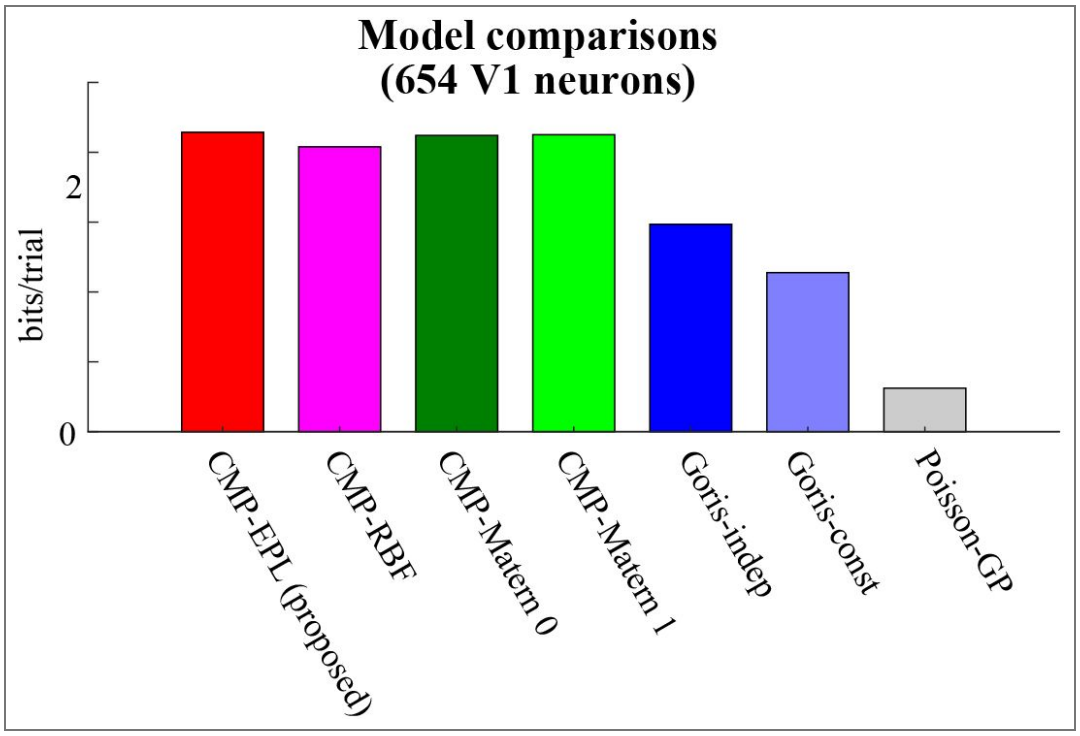


Figure S1. Comparison of the average test data log-likelihood improvement of each model relative to the baseline Poisson model, across 654 V1 neurons and all 72 grating orientations in recordings from the macaque primary visual cortex (data from [36]). From left to right: the proposed CMP model, the CMP-RBF model, the CMP-Matérn 0 model, the CMP-Matérn 1 model, the Goris-independent gain model, the Goris-constant gain model, and the Poisson-GP model. These results demonstrate that the CMP model systematically outperforms all other models in fitting held-out data.

Figure S2. Comparison of the performance of the proposed CMP framework with an alternative model in which, instead of sharing a common set of gain hyperparameters ρ_g, ℓ_g, q across stimuli (as in the CMP framework), each stimulus condition has its own independent gain hyperparameters $\rho_g^{(s)}, \ell_g^{(s)}, q^{(s)}$.

Results are shown for a sample neuron from the macaque primary visual cortex dataset (data from [36]). **(A)** Comparison of the inferred time-domain activity. From top to bottom: the inferred stimulus-driven components at three selected orientation settings (0°, 125°, 295°); inferred gain processes for two trials per orientation; inferred firing rates from the proposed CMP model (solid lines) and the alternative model with independent gain hyperparameters per stimulus (dashed lines); and the observed spike raster. **(B)** Comparison of the log-likelihood improvement relative to the baseline Poisson model for the proposed CMP model (left) versus the alternative independent-gain model (right). These results show that the performance of both models is very similar, corroborating that the proposed CMP modeling framework robustly captures trial-to-trial variability across different stimuli using a shared set of hyperparameters.

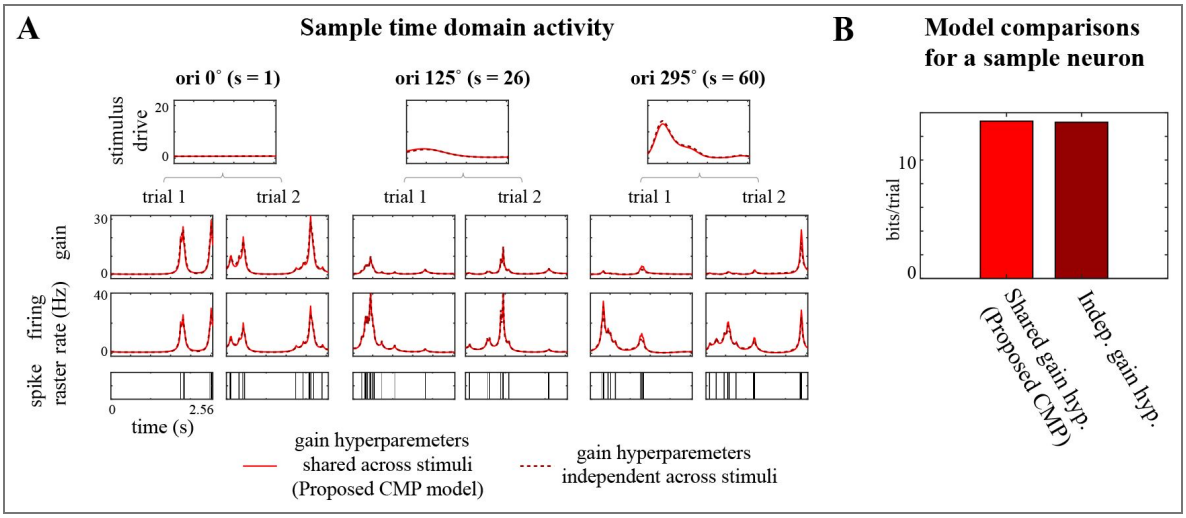
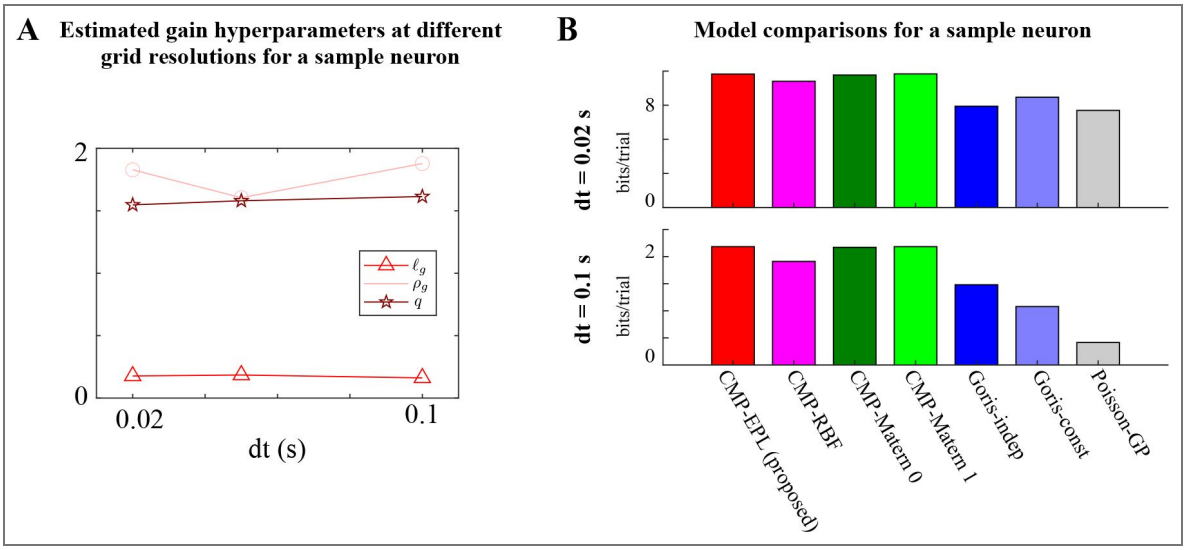


Figure S3. Comparison of CMP model inference at different grid sizes *dt* for a sample neuron from the macaque primary visual cortex dataset (data from [36]).

(A) Comparison of the inferred gain hyperparameters ℓ_g, ρ_g, q across three sufficiently fine grid resolutions: dt = 0.02 s, dt = 0.05 s, and dt = 0.1 s. These results indicate that the inference procedure is robust to moderate changes in grid size once the temporal discretization is fine enough to approximate the underlying continuous-time model. **(B)** Comparison of the test data log-likelihood gain relative to the baseline Poisson model at two different grid resolutions (dt = 0.02 s and dt = 0.1 s). From left to right: the proposed CMP model, CMP-RBF model, CMP-Matérn 0 model, CMP-Matérn 1 model, Goris-independent gain model, Goris-constant gain model, and the Poisson-GP model. These results show that the CMP consistently outperforms alternative models over this range.



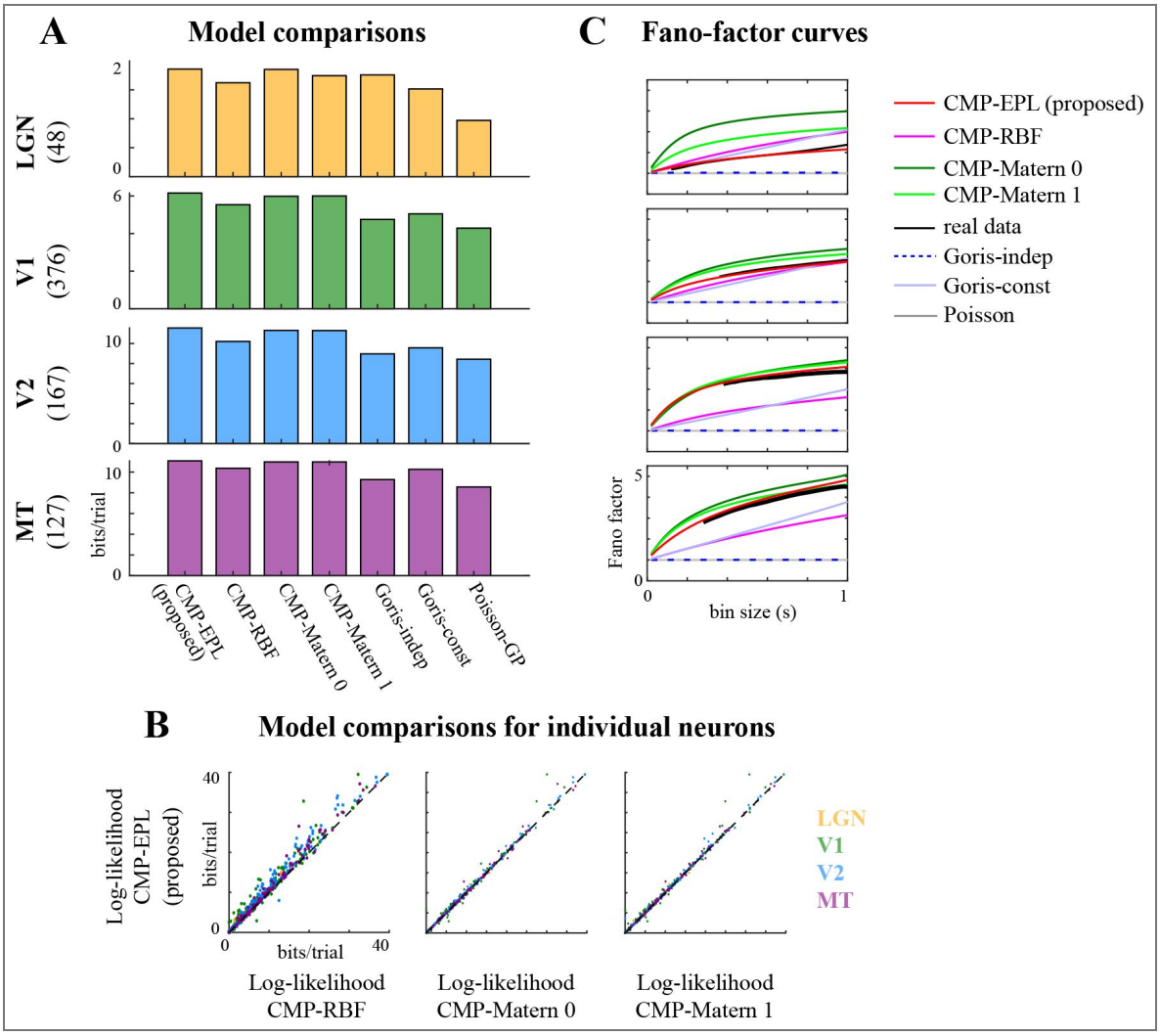


Figure S4. Performance comparisons of different models on neuronal population data recorded from various areas along the visual hierarchy: the lateral geniculate nucleus (LGN), V1, V2, and MT, in response to drifting sinusoidal gratings of the preferred size and speed, varying either in spatial frequency (12 spatial frequencies, ranging from 0 to 10 cycles/deg) or in drift direction (16 equally spaced directions) (data from [5]).

(A) Log-likelihood improvement of each model relative to the baseline Poisson model, averaged across held-out data for each population. Rows from top to bottom: LGN, V1, V2, and MT. The number of neurons in each population is indicated in parentheses. From left to right: proposed CMP model, CMP-RBF model, CMP-Matérn 0 model, CMP-Matérn 1 model, Goris-independent gain model, Goris-constant gain model, and the Poisson-GP model. **(B)** Comparison of the log-likelihood of the proposed CMP model with the CMP-RBF model (left), the CMP-Matérn 0 model (middle), and the CMP-Matérn 1 model (right) for individual neurons. Colors indicate brain areas: LGN (yellow), V1 (green), V2 (blue), and MT (purple). **(C)** Comparison of the inferred Fano factor as a function of bin size, averaged across each population (rows from top to bottom: LGN, V1, V2, and MT). Curves show the real data (black), the proposed CMP model (red), CMP-RBF model (magenta), CMP-Matérn 0 model (dark green), CMP-Matérn 1 model (light green), Goris-independent gain model (dark blue dashed lines), Goris-constant gain model (light blue), and the baseline Poisson model (gray).

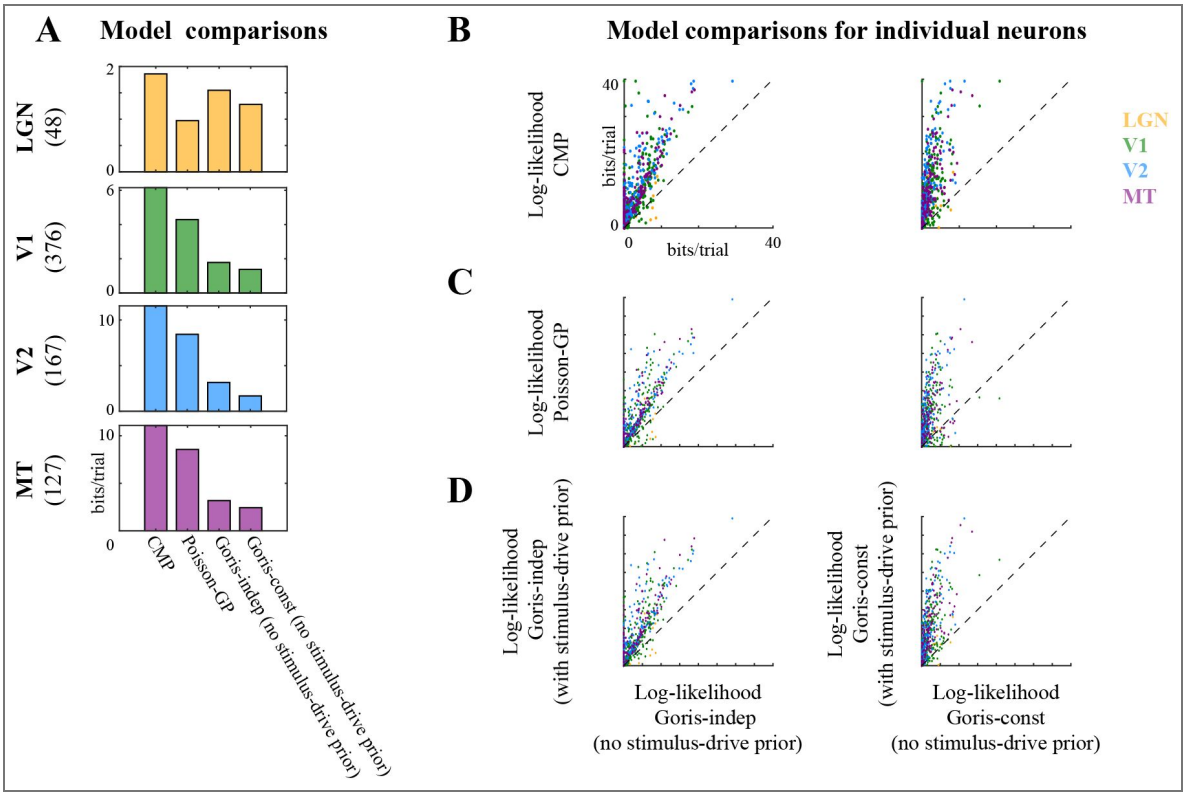


Figure S5. Comparison of the performance of the Goris model with and without a smoothness prior on the stimulus drive (data from [5]).

(A) Log-likelihood improvement of each model relative to the baseline Poisson model, averaged across held-out data for each population. Rows from top to bottom: LGN, V1, V2, and MT, with the number of neurons in each population indicated in parentheses. From left to right: the proposed CMP model, the Poisson-GP model, the Goris-independent gain model without a GP prior on the stimulus drive, and the Goris-constant gain model without a GP prior on the stimulus drive. **(B)** Comparison of the log-likelihood of the proposed CMP model versus the Goris-independent gain model without a GP prior on the stimulus drive (left), and versus the Goris-constant gain model without a GP prior on the stimulus drive (right), for individual neurons. **(C)** Comparison of the log-likelihood of the Poisson-GP model versus the Goris-independent gain model without a GP prior (left), and versus the Goris-constant gain model without a GP prior (right), for individual neurons. **(D)** Comparison of the log-likelihood of the Goris-independent gain model with a GP prior on the stimulus drive versus without a GP prior (left), and the Goris-constant gain model with a GP prior versus without a GP prior (right), for individual neurons. Colors indicate the areas: LGN (yellow), V1 (green), V2 (blue), and MT (purple). These results show that including a smoothness prior on the stimulus drive generally improves the model's ability to fit held-out data, though the proposed CMP framework still achieves systematically higher performance across all areas. This highlights the importance of jointly modeling smooth stimulus tuning and structured gain fluctuations for accurately capturing neural variability.

Data availability

No new datasets were generated for this study. The analyses are based on previously published neural recordings that can be obtained from the original authors upon reasonable request. The modeling and analysis code for the Continuous Modulated Poisson (CMP) framework is available at *https://github.com/Anuththara-Rupasinghe/CMP*.

Acknowledgements

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).

Appendix 1. Proof of the validity of the EPL covariance function

In this section, we provide the formal proof that the proposed Exponetiated Power Law (EPL) kernel is a valid covariance function.

Theorem 1.

(Validity of the EPL Kernel). *The Exponetiated Power Law (EPL) kernel given by:*

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

where $\rho_g > 0$ *is the marginal variance,* $\ell_g > 0$ *is the length scale, and* $0 < q \leq 2$ *is the power law exponent, is a valid covariance function.*

Proof. We begin by recalling the definition of a valid covariance function [71].

Definition 1.

(Covariance function). *A (real-valued) function*

$$K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

is called a covariance kernel *(or* positive semidefinite kernel*) if, for every finite set of points* $t_1, \ldots, t_n \in \mathbb{R}$ *and all real weights* $a_1, \ldots, a_n \in \mathbb{R}$*, the following Gram sum is non-negative:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K\left(t_i, t_j\right) \geq 0$$

We now invoke the following result from the Schoenberg's theorem [72].

Lemma 1.

(Schoenberg (1938)). *The function* $\exp\left(-|x|^q\right)$ *is positive definite on* $\mathbb{R}$ *for all* $0 < q \leq 2$. *Proof of Lemma 1*. The proof is provided in [72].

From Lemma 1 and Definition 1, it follows that for any finite set $t_1, \ldots, t_n \in \mathbb{R}$ and weights $a_1, \ldots, a_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \exp\left(-\left|\tilde{t}_i - \tilde{t}_j\right|^q\right) \geq 0$$

if $0 < q \leq 2$.

Next, consider the change of variables $t_i = 2^{1/q}\ell_g \cdot \tilde{t}_i$, which implies:

$$\left|\tilde{t}_i - \tilde{t}_j\right|^q = \left|\frac{t_i}{(2)^{1/q}\ell_g} - \frac{t_j}{(2)^{1/q}\ell_g}\right|^q = \frac{1}{2}\left|\frac{t_i - t_j}{\ell_g}\right|^q$$

Substituting this into the above expression, we obtain:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \exp\left(-\frac{1}{2}\left|\frac{t_i - t_j}{\ell_g}\right|^q\right) \geq 0$$

Finally, since $\rho_g > 0$, multiplying by ρ_g preserves non-negativity:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_i - t_j}{\ell_g}\right|^q\right) \geq 0$$

Hence, for any finite set $t_1, \ldots, t_n \in \mathbb{R}$ and any real weights $a_1, \ldots, a_n \in \mathbb{R}$, we have:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_g\left(t_i, t_j\right) \geq 0$$

which establishes that the EPL kernel in Equation 1 is positive semidefinite, and thus a valid covariance function.

Appendix 2. Partitioning variability with CMP

In this section, we provide the detailed derivations of the mean–variance relationship under the CMP model. 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 (Δ), and thus to analytically characterize how overdispersion grows with time bin size.

Recall the CMP forward model:

$$\begin{aligned} y(t) \mid s &\sim \text{PoissProc}(\lambda(t)), \\ \lambda(t) &= f_s(t) \cdot g(t), \\ \tilde{g}(t) &:= \log g(t) \sim \mathcal{GP}\left(0, K_g\right), \\ K_g\left(t_1, t_2\right) &= \rho_g \exp\left(-\frac{1}{2}\left|\frac{t_1 - t_2}{\ell_g}\right|^q\right). \end{aligned}$$

Under this model, the mean spike count in a time bin of size Δ, conditioned on the stimulus s, can be derived as:

$$\begin{aligned} \text{mean}\left[y_\Delta \mid s\right] &= \mathbb{E}[y_\Delta \mid s \\ &= \mathbb{E}\left[\int_0^\Delta y(t)dt \mid s\right] \\ &\overset{(a)}{=} \int_0^\Delta \mathbb{E}[y(t) \mid s]dt \\ &\overset{(b)}{=} \int_0^\Delta \mathbb{E}[\lambda(t) \mid s]dt \\ &= \int_0^\Delta \mathbb{E}\left[f_s(t) \cdot g(t) \mid s\right]dt \\ &= \int_0^\Delta f_s(t)\mathbb{E}[g(t)]dt \\ &= \int_0^\Delta f_s(t)\mathbb{E}[\exp(\tilde{g}(t))]dt \\ &= \int_0^\Delta f_s(t)\exp\left(\rho_g/2\right)dt \\ &= \exp\left(\rho_g/2\right)\int_0^\Delta f_s(t)dt \end{aligned} \tag{2}$$

where (*a*) follows from Fubini's Theorem [73], (*b*) uses the fact that the mean of a Poisson process equals its rate function [3], and (*c*) applies the moment-generating function of a Gaussian (i.e., the mean of a log-normal distribution) [74].

Using similar steps, the second moment can be derived as:

$$\begin{aligned}
\mathbb{E}\left[y_\Delta^2 \mid s\right] &= \mathbb{E}\left[\left(\int_0^\Delta y(t)dt\right)^2 \mid s\right] \\
&= \mathbb{E}\left[\iint y(t_1)y(t_2)\,dt_1dt_2 \mid s\right] \\
&= \iint \mathbb{E}\left[y(t_1)y(t_2) \mid s\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \mathbb{E}\left[y(t_1)y(t_2) \mid s\right] dt_1dt_2 + \iint_{t_1\neq t_2} \mathbb{E}\left[y(t_1)y(t_2) \mid s\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \mathbb{E}\left[y(t_1)^2 \mid s\right] dt_1dt_2 + \iint_{t_1\neq t_2} \mathbb{E}\left[y(t_1)y(t_2) \mid s\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \mathbb{E}\left[\lambda(t_1)+\lambda(t_1)^2 \mid s\right] dt_1dt_2 + \iint_{t_1\neq t_2} \mathbb{E}\left[\lambda(t_1)\lambda(t_2) \mid s\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \left(f_s(t_1)\mathbb{E}\left[g(t_1)\right] + f_s(t_1)^2\mathbb{E}\left[g(t_1)^2\right]\right) dt_1dt_2 + \iint_{t_1\neq t_2} f_s(t_1)f_s(t_2)\mathbb{E}\left[g(t_1)g(t_2)\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \left(f_s(t_1)\mathbb{E}\left[\exp\left(\tilde{g}(t_1)\right)\right] + f_s(t_1)^2\mathbb{E}\left[\exp\left(2\tilde{g}(t_1)\right)\right]\right) dt_1dt_2 \\
&\quad + \iint_{t_1\neq t_2} f_s(t_1)f_s(t_2)\mathbb{E}\left[\exp\left(\tilde{g}(t_1)+\tilde{g}(t_2)\right)\right] dt_1dt_2 \\
&= \iint_{t_1=t_2} \left(f_s(t_1)\exp\left(\rho_g/2\right) + f_s(t_1)^2\exp\left(2\rho_g\right)\right) dt_1dt_2
\end{aligned}$$

$$\begin{aligned}
&+ \iint_{t_1\neq t_2} f_s\left(t_1\right) f_s\left(t_2\right) \exp\left(\rho_g + K_g\left(t_1,t_2\right)\right) dt_1dt_2 \\
&= \exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt \\
&+ \exp\left(\rho_g\right) \iint f_s\left(t_1\right) f_s\left(t_2\right) \exp\left(K_g\left(t_1,t_2\right)\right) dt_1dt_2
\end{aligned} \tag{3}$$

Thus, using Equation 2 and Equation 3 the variance can be formulated as:

$$\begin{aligned}
\mathrm{var}\left[y_\Delta \mid s\right] &= \mathbb{E}\left[y_\Delta^2 \mid s\right] - \left(\mathbb{E}\left[y_\Delta \mid s\right]\right)^2 \\
&= \exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt + \exp\left(\rho_g\right) \\
&\quad \iint f_s\left(t_1\right) f_s\left(t_2\right) \exp\left(K_g\left(t_1,t_2\right)\right) dt_1dt_2 \\
&\quad - \left(\exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt\right) \\
&= \exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt + \exp\left(\rho_g\right) \\
&\quad \iint f_s\left(t_1\right) f_s\left(t_2\right) \left(\exp\left(K_g\left(t_1,t_2\right)\right) - 1\right) dt_1dt_2
\end{aligned} \tag{4}$$

From these expressions, we can compute Fano factor as a function of bin size, which is simply the ratio of variance (Equation 4) to mean (Equation 2):

$$\begin{aligned}
FF(\Delta) &= \frac{\mathrm{var}\left[y_\Delta \mid s\right]}{\mathbb{E}\left[y_\Delta \mid s\right]} \\
&= \frac{\exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt + \exp\left(\rho_g\right) \iint f_s\left(t_1\right) f_s\left(t_2\right) \left(\exp\left(K_g\left(t_1,t_2\right)\right) - 1\right) dt_1dt_2}{\exp\left(\rho_g/2\right) \int_0^\Delta f_s(t)dt} \\
&= 1 + \frac{\exp\left(\rho_g/2\right) \int_0^\Delta \int_0^\Delta f_s\left(t_1\right) f_s\left(t_2\right) \left(\exp\left(K_g\left(t_1,t_2\right)\right) - 1\right) dt_1dt_2}{\int_0^\Delta f_s(t)dt}.
\end{aligned}$$

In the special case where the stimulus component is non-fluctuating, i.e., $f_s(t) = f$, these expressions simplify to:

$$\begin{aligned}\mathbb{E}\left[y_\Delta \mid s\right] &= \exp\left(\rho_g/2\right)\int_0^\Delta f_s(t)dt \\ &= \exp\left(\rho_g/2\right)\bar{f}\Delta \\ &= \Delta\bar{\lambda},\end{aligned}$$

$$\begin{aligned}\mathrm{var}[y_\Delta \mid s] &= \exp\left(\rho_g/2\right)\int_0^\Delta f_s(t)dt + \exp\left(\rho_g\right)\iint f_s(t_1)f_s(t_2)\left(\exp\left(K_g(t_1,t_2)\right)-1\right)dt_1dt_2 \\ &= \exp\left(\rho_g/2\right)\bar{f}\Delta + \exp\left(\rho_g\right)\bar{f}^2\left(\iint\left(\exp\left(K_g(t_1,t_2)\right)\right)dt_1dt_2 - \Delta^2\right) \\ &= \Delta\bar{\lambda} + \left(\Delta\bar{\lambda}\right)^2 \times \left(\frac{1}{\Delta^2}\iint\left(\exp\left(K_g(t_1,t_2)\right)\right)dt_1dt_2 - 1\right) \\ &= \Delta\bar{\lambda} + \left(\Delta\bar{\lambda}\right)^2\left(\kappa(\Delta)-1\right),\end{aligned}$$

and:

$$FF(\Delta) = 1 + \Delta\lambda(\kappa(\Delta)-1)$$

where we define $\lambda = f\exp\left(\rho_g/2\right)$ and $\kappa(\Delta) = \frac{1}{\Delta^2}\int_0^\Delta\int_0^\Delta e^{K_g(t_1,t_2)}dt_1dt_2$.

Appendix 3. Details about the Goris model variants

In this appendix, we provide technical details for two extensions of the modulated Poisson framework introduced by Goris et al. [1], adapted for spike-train data with multiple time bins per trial. Specifically, we derive the likelihoods, means, and variances for the Goris-indep model, which assumes time-bin–independent modulatory gains, and the Goris-const model, which assumes a constant gain across time bins within a trial. These derivations clarify the statistical properties of each model and describe how we compute log-likelihoods for model comparison. All Goris-style baselines reported in the main model comparisons use a smooth stimulus-drive estimate obtained from the Poisson-GP model, so that comparisons among these baselines focus on the assumed temporal structure of the gain process. We also evaluate versions without this smoothness prior, in which the stimulus-drive parameters are estimated directly under the corresponding Goris-style likelihood. Details for both versions are provided below.

3.1. Goris-indep model

Here we provide further details on the Goris-independent gain model, including the derivation of the likelihood, as well as the mean and variance under this model. The Goris-indep model is an extension of the modulated Poisson model in Goris et al. [1], adapted for spike train data with multiple time bins per trial. Given spiking observations of the s^{th} stimulus presentation across t = 1, ..., T time bins, we assume that the modulatory gain process is independent across time bins, with shared variance $\sigma^2_{g_s}$, i.e.,

$$\begin{aligned}y_{t,s,n} &\sim \mathrm{Poisson}\left(\lambda_{t,s,n}\right), \\ \lambda_{t,s,n} &= g_{t,s,n}f_{t,s}dt,\end{aligned}$$

where $g_{t,s,n}$ ~ Gamma(r_s, s_s), with shape parameter $r_s = \frac{1}{\sigma^2_{g_s}}$ and scale parameter $s_s = \sigma^2_{g_s}$. This implies $\lambda_{t,s,n}$ ~ Gamma (r_s, $s_{t,s}$), where $s_{t,s} = \sigma^2_{g_s}f_{t,s}dt$.

Appendix 3.1.1 Likelihood derivation

Following the results in [1], marginalizing over the gain variable yields independent Negative Binomial distributions:

$$
\begin{aligned}
&p(y_{1:T,s,1:N}|r_s, s_s) \\
&= \int_0^\infty p(y_{1:T,s,1:N}, \lambda_{1:T,s,1:N}|r_s, s_s))d\lambda_{1:T,s,1:N} \\
&= \int_0^\infty \prod_{t,n=1}^{T,N} \text{Poisson}\left(y_{t,s,n}|\lambda_{t,s,n}\right) \times \text{Gamma}\left(\lambda_{t,s,n}|r_s, s_{t,s}\right) d\lambda_{t,s,n} \\
&= \prod_{t,n=1}^{T,N} \int_0^\infty \exp\left(-\lambda_{t,s,n}\right) \frac{\lambda_{t,s,n}^{y_{t,s,n}}}{y_{t,s,n}!} \times \lambda_{t,s,n}^{r_s-1} \frac{\exp\left(-\frac{\lambda_{t,s,n}}{s_{t,s}}\right)}{s_{t,s}^{r_s}\Gamma(r_s)} d\lambda_{t,s,n} \\
&= \prod_{t,n=1}^{T,N} \frac{1}{s_{t,s}^{r_s}} \frac{1}{y_{t,s,n}!\Gamma(r_s)} \int_0^\infty \lambda_{t,s,n}^{r_s+y_{t,s,n}-1} \exp\left(-\lambda_{t,s,n}\left(1+\frac{1}{s_{t,s}}\right)\right) d\lambda_{t,s,n} \\
&= \prod_{t,n=1}^{T,N} \frac{1}{s_{t,s}^{r_s}} \frac{1}{y_{t,s,n}!\Gamma(r_s)} \Gamma(r_s + y_{t,s,n}) \left(\frac{s_{t,s}}{1+s_{t,s}}\right)^{y_{t,s,n}+r_s} \int_0^\infty \text{Gamma}\left(\lambda_{t,s,n}\middle| r_s + y_{t,s,n}, \frac{s_{t,s}}{1+s_{t,s}}\right) d\lambda_{t,s,n} \\
&= \prod_{t,n=1}^{T,N} \frac{\Gamma(r_s + y_{t,s,n})}{y_{t,s,n}!\Gamma(r_s)} \left(\frac{1}{1+s_{t,s}}\right)^{r_s} \left(\frac{s_{t,s}}{1+s_{t,s}}\right)^{y_{t,s,n}} \\
&= \prod_{t,n=1}^{T,N} \frac{(r_s + y_{t,s,n} - 1)!}{y_{t,s,n}!(r_s-1)!} \left(\frac{1}{1+s_{t,s}}\right)^{r_s} \left(\frac{s_{t,s}}{1+s_{t,s}}\right)^{y_{t,s,n}} \\
&= \prod_{t,n=1}^{T,N} \text{NegativeBinomial}\left(y_{t,s,n}\middle| r_s, \frac{1}{1+s_{t,s}}\right).
\end{aligned}
\tag{5}
$$

For the Goris-indep model used in the main model comparisons, we set the stimulus component to the smooth firing-rate estimate obtained from the Poisson-GP model. Specifically, we set:

$$s_{t,s} = \sigma_{g_s}^2 f_{t,s} dt = f_{t,s} dt / r_s = \hat{\lambda}_{t,s} / r_s$$

where $\hat{\lambda}_{t,s}$ denotes the Poisson-GP estimate of the stimulus-dependent firing rate. We then search for the value of r_s that minimizes the negative log-likelihood of the spiking observations $y_{1:T,s,1:N}$ for each stimulus condition. For the comparison without a smoothness prior on the stimulus drive, we instead estimate both r_s and $s_{1:T,s}$ directly by minimizing the negative log-likelihood under the above Negative Binomial mixture model, as in Goris et al. [1].

Appendix 3.1.2 Mean-variance derivations

Based on the above derivation, we see that:

$$p\left(y_{t,s,n} \mid r_s, s_{t,s}\right) \sim \text{NegativeBinomial}\left(r_s, \frac{1}{1+s_{t,s}}\right)$$

and $y_{1:t,s,n}$ are independent across $n = 1, \ldots, N$ when conditioned on r_s and $s_{t,s}$. Thus, from the mean and variance of the Negative Binomial distribution, we have:

$$
\begin{aligned}
\text{mean}\left[y_{t,s,1:N} \mid f_{t,s}, dt\right] &= \mathbb{E}\left[y_{t,s,n} \mid f_{t,s}, dt\right] = r_s s_{t,s} = f_{t,s} dt \\
\text{var}\left[y_{t,s,1:N} \mid f_{t,s}, dt\right] &= r_s s_{t,s}\left(1+s_{t,s}\right) = f_{t,s} dt\left(1+\sigma_{g_s}^2 f_{t,s} dt\right).
\end{aligned}
$$

Next, we extend this to derive the mean and variance of spike counts over a larger time window $\Delta = R\,dt$, by summing across the relevant bins $t_1, \ldots, t_R$:

$$
\begin{aligned}
\text{mean}\left(y_{\Delta,s,1:N} \mid \mathbf{f_s}, dt\right) = \mathbb{E}\left[y_{\Delta,s,n} \mid \mathbf{f_s}, dt\right] &= \mathbb{E}\left[\sum_{r'=1}^{R} y_{t_{r'},s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r'=1}^{R} \mathbb{E}\left[y_{t_{r'},s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r'=1}^{R} r_s s_{t_{r'},s} \\
&= \textstyle\sum_{r'=1}^{R} f_{t_{r'},s} dt
\end{aligned}
$$

and:

$$\begin{aligned}\operatorname{var}\left[y_{\Delta,s,1:N} \mid \mathbf{f_s}, dt\right] &= \operatorname{var}\left[\sum_{r'=1}^{R} y_{t_{r',s,n}} \mid \mathbf{f_s}, dt\right] \\ &= \sum_{r'=1}^{R} \operatorname{var}\left[y_{t_{r',s,n}} \mid \mathbf{f_s}, dt\right] \\ &= \sum_{r'=1}^{R} r_s s_{t_{r',s}}\left(1 + s_{t_{r',s}}\right) \\ &= \sum_{r'=1}^{R} f_{t_{r',s}} dt\left(1 + \sigma_{g_s}^2 f_{t_{r',s}} dt\right)\end{aligned}$$

3.2. Goris-const model

In this section, we provide further details on the Goris-constant gain model, including the derivation of the mean and variance, as well as the log-likelihood under its modeling assumptions. The Goris-const model is the second variant we consider for extending the Goris model in [1] to multiple time bins. Here, we assume that the modulatory gain process is constant across time bins:

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

where $g_{s,n} \sim \text{Gamma}(r_s, s_s)$, with shape parameter $r_s = \frac{1}{\sigma_{g_s}^2}$ and scale parameter $s_s = \sigma_{g_s}^2$. Furthermore, we assume that the total spike count across all time bins in a trial, $Y_{s,n} = \sum_t y_{t,s,n}$, shares the same gain variance $\sigma_{g_s}^2$, i.e.,

$$Y_{s,n} \sim \text{Poisson}\left(\tilde{g}_{s,n} f_s T\right), \tag{6}$$

with $\tilde{g}_{s,n} \sim \text{Gamma}\left(r_s, s_s\right)$. For the Goris-const model used in the main model comparisons, we estimate $\sigma_{g_s}^2$ by fitting the total spike counts $Y_{s,n}$ across training trials, as in Goris et al. [1], and set the stimulus encoding component $f_{t,s} dt = \hat{\lambda}_{t,s}$, where $\hat{\lambda}_{t,s}$ denotes the smooth firing-rate estimate obtained from the Poisson-GP model. For the comparison without a smoothness prior on the stimulus drive, we instead estimate $f_{t,s} dt$ using the same method as in the baseline Poisson model.

3.2.1. Mean-variance derivations

Note that from the moments of the Gamma distribution:

$$\mathbb{E}\left[g_{s,n}\right] = 1, \quad \mathbb{E}\left[g_{s,n}^2\right] = \sigma_{g_s}^2 + 1. \tag{7}$$

Accordingly, we derive the first moment, second moment, and variance at bin size dt as:

$$\begin{aligned}\text{mean}\left[y_{t,s,1:N} | f_{t,s}, dt\right] &= \mathbb{E}\left[y_{t,s,n} | f_{t,s}, dt\right] = \mathbb{E}\left[\mathbb{E}\left[y_{t,s,n} | f_{t,s}, dt, g_{s,n}\right]\right] = \mathbb{E}\left[g_{s,n} f_{t,s} dt\right] = f_{t,s} dt \mathbb{E}\left[g_{s,n}\right] \\ &= f_{t,s} dt, \\ \mathbb{E}\left[y_{t,s,n}^2 | f_{t,s}, dt\right] &= \mathbb{E}\left[\mathbb{E}\left[y_{t,s,n}^2 | f_{t,s}, dt, g_{s,n}\right]\right] = \mathbb{E}\left[g_{s,n} f_{t,s} dt + g_{s,n}^2 f_{t,s}^2 dt^2\right] \\ &= f_{t,s} dt + f_{t,s}^2 dt^2 (\sigma_{g_s}^2 + 1), \\ \mathbb{E}\left[y_{t_1,s,n} y_{t_2,s,n} | \mathbf{f_s}, dt\right] &= \mathbb{E}\left[\mathbb{E}\left[y_{t_1,s,n} y_{t_2,s,n} | \mathbf{f_s}, dt, g_{s,n}\right]\right] = \mathbb{E}\left[g_{s,n}^2 f_{t_1,s} f_{t_2,s} dt^2\right] = f_{t_1,s} f_{t_2,s} dt^2 (\sigma_{G_s}^2 + 1), \\ \text{var}\left[y_{t,s,1:N} | f_{t,s}, dt\right] &= \mathbb{E}\left[y_{t,s,n}^2 | f_{t,s}, dt\right] - \mathbb{E}\left[y_{t,s,n} | f_{t,s}, dt\right]^2 = f_{t,s} dt (f_{t,s} dt \sigma_{g_s}^2 + 1).\end{aligned}$$

We then extend these results to a larger bin size $\Delta = R\,dt$, by integrating over the differential statistics of the corresponding time bins $t_1, \cdots, t_R$ at bin-size dt:

$$\begin{aligned}
\text{mean}\ (y_{\Delta,s,1:N} \mid \mathbf{f_s}, dt) = \mathbb{E}\left[y_{\Delta,s,n} \mid \mathbf{f_s}, dt\right] &= \mathbb{E}\left[\sum_{r=1}^{R} y_{t_r,s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r=1}^{R} \mathbb{E}\left[y_{t_r,s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r=1}^{R} f_{t_r,s} dt, \\
\mathbb{E}\left[y^2_{\Delta,s,n} \mid \mathbf{f_s}, dt\right] &= \mathbb{E}\left[\left(\sum_{r=1}^{R} y_{t_r,s,n}\right)^2 \mid \mathbf{f_s}, dt\right] \\
&= \mathbb{E}\left[\sum_{r=1}^{R}\sum_{r'=1}^{R} y_{t_r,s,n} y_{t_{r'},s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r=1}^{R} \mathbb{E}\left[y^2_{t_r,s,n} \mid \mathbf{f_s}, dt\right] + \sum_{r \neq r'} \mathbb{E}\left[y_{t_r,s,n} y_{t_{r'},s,n} \mid \mathbf{f_s}, dt\right] \\
&= \sum_{r=1}^{R} \left(f_{t_r,s} dt + f^2_{t_r,s} dt^2 \left(\sigma^2_{g_s} + 1\right)\right) + \sum_{r \neq r'} \left(f_{t_r,s} f_{t_{r'},s} dt^2 \left(\sigma^2_{g_s} + 1\right)\right) \\
&= \sum_{r=1}^{R} f_{t_r,s} dt + \sum_{r=1}^{R}\sum_{r'=1}^{R} \left(f_{t_r,s} f_{t_{r'},s} dt^2 \left(\sigma^2_{g_s} + 1\right)\right),
\end{aligned}$$

and therefore, the variance becomes:

$$\begin{aligned}
\text{var}\left[y_{\Delta,s,1:N} \mid \mathbf{f_s}, dt\right] &= \mathbb{E}\left[y^2_{\Delta,s,n} \mid \mathbf{f_s}, dt\right] - \mathbb{E}[y_{\Delta,s,n} \mid \mathbf{f_s}, dt]^2 \\
&= \sum_{r=1}^{R} f_{t_r,s} dt + \sum_{r=1}^{R}\sum_{r'=1}^{R} \left(f_{t_r,s} f_{t_{r'},s} dt^2 \sigma^2_G\right) \\
&= \sum_{r=1}^{R} f_{t_r,s} dt \left(1 + \sigma^2_{g_s} \sum_{r'=1}^{R} f_{t_{r'},s} dt\right).
\end{aligned}$$

3.2.2. Log-likelihood derivation

We calculate the test log-likelihood of this model using a Monte Carlo integration procedure [75]:

$$\log(p(\mathbf{y} \mid \mathbf{f})) = \sum_{k,s} \log\left(p\left(\mathbf{y}_{s,n} \mid \mathbf{f}_s\right)\right), \quad \text{where,} \tag{8}$$

$$\begin{aligned}
p\left(\mathbf{y}_{s,n} \mid \mathbf{f}_s\right) &= \int p\left(\mathbf{y}_{s,n}, g_{s,n} \mid \mathbf{f}_s\right) dg_{s,n} = \int p\left(\mathbf{y}_{s,n} \mid g_{s,n}, \mathbf{f}_s\right) p\left(g_{s,n}\right) dg_{s,n} \\
&= \mathbb{E}_{g_{s,n}}\left[p\left(\mathbf{y}_{s,n} \mid g_{s,n}, \mathbf{f}_s\right)\right] \\
&\approx \frac{1}{I} \sum_{i=1}^{I} p\left(\mathbf{y}_{s,n} \mid g^{(i)}_{s,n}, \mathbf{f}_s\right), g^{(i)}_{s,n} \sim \text{Gamma}\left(r_s, s_s\right).
\end{aligned}$$

Defining:

$$\begin{aligned}
L^{(i)}_{s,n} &= \log\left(p\left(\mathbf{y}_{s,n} \mid g^{(i)}_{s,n}, \mathbf{f}_s\right)\right) \\
&= \log\left(\prod_{t=1}^{T} \frac{\exp\left(-g^{(i)}_{s,n} f_{t,s} dt\right)\left(g^{(i)}_{s,n} f_{t,s} dt\right)^{y_{t,s,n}}}{y_{t,s,n}!}\right) \\
&= \sum_{t=1}^{T} \left(-g^{(i)}_{s,n} f_{t,s} dt + y_{t,s,n} \log\left(g^{(i)}_{s,n} f_{t,s} dt\right) - \log\left(y_{t,s,n}!\right)\right),
\end{aligned}$$

we get:

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

where $L_{s,n}^{\max} = \max_i L_{s,n}^{(i)}$. We set I = 1000 and use this formulation to derive the log-likelihood of the spiking observations in the n^{th} test trial of the s^{th} stimulus, and finally calculate the joint log-likelihood of all test data following Equation 8.

Additional information

Funding

Funder	Grant reference number	Author
NIH T32 institutional training grant	T32MH065214	Anuththara Rupasinghe
National Science Foundation (NSF)	NSF CAREER award 2340338	Adam S Charles
Simons Foundation (SF)	Simons Collaboration on the Global Brain (SCGB AWD543027)	Jonathan W Pillow
HHS | NIH | BRAIN Initiative (The BRAIN Initiative)	NIH BRAIN initiative R01DA056404	Jonathan W Pillow
HHS | NIH | BRAIN Initiative (The BRAIN Initiative)	NIH BRAIN initiative R01EB02694	Jonathan W Pillow
HHS | NIH | National Eye Institute (NEI)	R01EY033064	Jonathan W Pillow
HHS | NIH | BRAIN Initiative (The BRAIN Initiative)	U19 NIH-NINDS BRAIN Initiative Award (5U19NS123716)	Jonathan W Pillow

Author ORCID iDs

Anuththara Rupasinghe: https://orcid.org/0000-0003-2143-8709
Adam S Charles: https://orcid.org/0000-0002-9045-3489
Jonathan W Pillow: https://orcid.org/0000-0002-3638-8831

References

[1] **Goris R L T,** Movshon J A, Simoncelli E P (2014) Partitioning neuronal variability. *Nature Neuro-science* **17**:858-865 https://doi.org/10.1038/nn.3711 | PubMed

[2] **Tomko George J,** Crapper Donald R (1974) Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Research* **79**:405-418 https://doi.org/10.1016/0006-8993(74)90438-7 | PubMed

[3] **Henry GH,** Bishop PO, Tupper RM, Dreher B (1973) Orientation specificity and response variability of cells in the striate cortex. *Vision Research* **13**:1771-1779 https://doi.org/10.1016/0042-6989(73)90094-1 | PubMed

[4] **Carandini M** (2004) Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS biology* **2**:e264 https://doi.org/10.1371/journal.pbio.0020264 | PubMed

[5] **Dinstein Ilan,** Heeger David J, Behrmann Marlene (2015) Neural variability: friend or foe?. *Trends in Cognitive Sciences* **19**:322-328 https://doi.org/10.1016/j.tics.2015.04.005 | PubMed

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

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

[8] **Schneeweis David M,** Schnapf Julie L (1999) The photovoltage of macaque cone photoreceptors: Adaptation, noise, and kinetics. *Journal of Neuroscience* **19**:1203-1216 https://doi.org/10.1523/JNEUROSCI.19-04-01203.1999 | PubMed

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

[10] **Clifford Colin WG,** Webster Michael A, Stanley Garrett B, Stocker Alan A, Kohn Adam, Sharpee Tatyana O, Schwartz Odelia (2007) Visual adaptation: Neural, psychological and computational aspects. *Vision Research* **47**:3125-3131 https://doi.org/10.1016/j.visres.2007.08.023 | PubMed

[11] **Feldman Daniel E** (2009) Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience* **32**:33-55 https://doi.org/10.1146/annurev.neuro.051508.135516 | PubMed

[12] **Perkel Donald H,** Gerstein George L, Moore George P (1967) Neuronal spike trains and stochastic point processes: I. the single spike train. *Biophysical Journal* **7**:391-418 https://doi.org/10.1016/S0006-3495(67)86596-2 | PubMed

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

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

[15] **Pillow JW** (2009) Time-rescaling methods for the estimation and assessment of non-poisson neural encoding models. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (Eds). *Advances in Neural Information Processing Systems 22* MIT Press. pp. 1473-1481

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

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

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

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

[20] **Swindale NV,** Mitchell DE (1994) Comparison of receptive field properties of neurons in area 17 of normal and bilaterally amblyopic cats. *Experimental Brain Research* **99**:399-410 https://doi.org/10.1007/BF00228976 | PubMed

[21] **Baddeley Roland,** Abbott LF, Booth Michael CA, Sengpiel Frank, Freeman Tobe, Wakeman Edward A, Rolls Edmund T (1997) 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**:1775-1783 https://doi.org/10.1098/rspb.1997.0246 | PubMed

[22] **Lánsky P,** Vaillant J (2000) Stochastic model of the overdispersion in the place cell discharge. *Biosystems* **58**:27-32 https://doi.org/10.1016/s0303-2647(00)00103-9 | PubMed

[23] **Pillow Jonathan,** Scott James (2012) Fully bayesian inference for neural models with negative-binomial spiking. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ (Eds). *Advances in Neural Information Processing Systems* **25** Curran Associates, Inc.

[24] **Gao Yuanjun,** Busing Lars, Shenoy Krishna V, Cunningham John P (2015) High-dimensional neural spike train analysis with generalized count linear dynamical systems. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (Eds). *Advances in Neural Information Processing Systems* **28** Curran Associates,

Inc.
https://proceedings.neurips.cc/paper_files/paper/2015/file/9996535e07258a7bbfd8b132435c5962-Paper.pdf

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

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

[27] **Barbieri R,** Quirk MC, Frank LM, Wilson MA, Brown EN (2001) Construction and analysis of non-poisson stimulus-response models of neural spiking activity. *Journal of Neuroscience Methods* **105**:25-37 https://doi.org/10.1016/s0165-0270(00)00344-7 | PubMed

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

[29] **Aghamohammadi Cina,** Chandrasekaran Chandramouli, Engel Tatiana A (2024) A doubly stochastic renewal framework for partitioning spiking variability. *bioRxiv* https://doi.org/10.1101/2024.02.21.581457

[30] **Charles A,** Pillow J (2018) Additive continuous-time joint partitioning of neural variability. In: 2018 Conference on Cognitive Computational Neuroscience. https://doi.org/10.32470/ccn.2018.1087-0

[31] **Goris Robbe LT,** Ziemba Corey M, Movshon J Anthony, Simoncelli Eero P (2018) Slow gain fluctuations limit benefits of temporal integration in visual cortex. *Journal of Vision* **18**:8-8 https://doi.org/10.1167/18.8.8 | PubMed

[32] **Hénaff Olivier J,** Boundy-Singer Zoe M, Meding Kristof, Ziemba Corey M, Goris Robbe L T (2020) Representation of visual uncertainty through neural gain variability. *Nat. Commun* **11**:2513 https://doi.org/10.1038/s41467-020-15533-0 | PubMed

[33] **Rasmussen Carl Edward,** Williams Christopher KI (2006) *Gaussian processes for machine learning. Adaptive computation and machine learning* MIT Press.

[34] **Zhao Yuan,** Park Il Memming (2017) Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains. *Neural Computation* **29**:1293-1316 https://doi.org/10.1162/NECO_a_00953 | PubMed

[35] **Keeley Stephen,** Aoi Mikio, Yu Yiyi, Smith Spencer, Pillow Jonathan W (2020) Identifying signal and noise structure in neural population activity with gaussian process factor models. In: Advances in Neural Information Processing Systems 33. pp. 13795-13805

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

[37] **Cox DR** (1955) Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)* **17**:129-157 https://doi.org/10.1111/j.2517-6161.1955.tb00188.x

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

[39] **Churchland Mark M,** Yu Byron M, Cunningham John P, Sugrue Leo P, Cohen Marlene R, Corrado Greg S, Newsome William T, Clark Andrew M, Hosseini Paymon, Scott Benjamin B, *et al.* (2010) Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci* **13**:369-378 https://doi.org/10.1038/nn.2501 | PubMed

[40] **Goris Robbe L T,** Coen-Cagli Ruben, Miller Kenneth D, Priebe Nicholas J, Lengyel Máté (2024) Response sub-additivity and variability quenching in visual cortex. *Nat. Rev. Neurosci* **25**:237-252 https://doi.org/10.1038/s41583-024-00795-0 | PubMed

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

[42] **Bartolo Ramon,** Saunders Richard C, Mitz Andrew R, Averbeck Bruno B (2020) Information-limiting correlations in large neural populations. *Journal of Neuroscience* **40**:1668-1678 https://doi.org/10.1523/JNEUROSCI.2072-19.2019 | PubMed

[43] **Rupasinghe Anuththara,** Francis Nikolas, Liu Ji, Bowen Zac, Kanold Patrick O, Babadi Behtash (2021) Direct extraction of signal and noise correlations from two-photon calcium imaging of ensemble neuronal activity. *eLife* **10** https://doi.org/10.7554/eLife.68046 | PubMed

[44] **Jun James J,** Steinmetz Nick, Siegle Joshua H, Denman Daniel J, Bauza Marius, Barbarits Brian, Lee Albert K, Anastassiou Costas, Andrei Alexandru, Aydin Cagatay, *et al.* (2017) Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**:232-236 https://doi.org/10.1038/nature24636 | PubMed

[45] **Du Jiangang,** Blanche Timothy J, Harrison Reid R, Lester Henry A, Sotiris C (2011) Masmanidis. Multiplexed, high density electrophysiology with nanofabricated neural probes. *PLOS One* **6**:1-11 https://doi.org/10.1371/journal.pone.0026204 | PubMed

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

[47] **Sokoloski Sacha,** Aschner Amir, Coen-Cagli Ruben (2021) Modelling the neural code in large populations of correlated neurons. *eLife* **10** https://doi.org/10.7554/eLife.64615 | PubMed

[48] **Faisal A Aldo,** Selen Luc PJ, Wolpert Daniel M (2008) Noise in the nervous system. *Nature Reviews Neuroscience* **9**:292-303 https://doi.org/10.1038/nrn2258 | PubMed

[49] **Simoncelli Eero P,** Olshausen Bruno A (2001) Natural image statistics and neural representation. *Annual review of neuroscience* **24**:1193-1216 https://doi.org/10.1146/annurev.neuro.24.1.1193 | PubMed

[50] **Orbán Gergő,** Berkes Pietro, Fiser József, Lengyel Máté (2016) Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**:530-543 https://doi.org/10.1016/j.neuron.2016.09.038 | PubMed

[51] **Churchland Mark M,** Yu Byron M, Ryu Stephen I, Santhanam Gopal, Shenoy Krishna V (2006) Neural variability in premotor cortex provides a signature of motor preparation. *J Neurosci* **26**:3697-3712 https://doi.org/10.1523/jneurosci.3762-05.2006 | PubMed

[52] **Gardiner CW** (2004) *Handbook of stochastic methods for physics, chemistry and the natural sciences, volume 13 of Springer Series in Synergetics* (third) Berlin: Springer-Verlag.

[53] **Uhlenbeck GE,** Ornstein LS (1930) On the theory of the brownian motion. *Phys. Rev* **36**:823-841 https://doi.org/10.1103/PhysRev.36.823

[54] **Mandelbrot Benoit B,** Van Ness John W (1968) Fractional brownian motions, fractional noises and applications. *SIAM Review* **10**:422-437 https://doi.org/10.1137/1010093

[55] **Tu Yuhai,** Grinstein G (2005) How white noise generates power-law switching in bacterial flagellar motors. *Phys. Rev. Lett* **94**:208101 https://doi.org/10.1103/PhysRevLett.94.208101 | PubMed

[56] **Sassi Alberto Stefano,** Garcia-Alcala Mayra, Aldana Maximino, Tu Yuhai (2022) Protein concentration fluctuations in the high expression regime: Taylor's law and its mechanistic origin. *Phys. Rev. X* **12**:011051 https://doi.org/10.1103/PhysRevX.12.011051 | PubMed

[57] **Berry M J,** Meister M (1998) Refractoriness and neural precision. *J. Neurosci* **18**:2200-2211 https://doi.org/10.1523/jneurosci.18-06-02200.1998 | PubMed

[58] **Maimon Gaby,** Assad John A (2009) Beyond poisson: increased spike-time regularity across primate parietal cortex. *Neuron* **62**:426-440 https://doi.org/10.1016/j.neuron.2009.03.021 | PubMed

[59] **Ferrari Ulisse,** Deny Stéphane, Marre Olivier, Mora Thierry (2018) A simple model for low variability in neural spike trains. *Neural Computation* **30**:3009-3036 https://doi.org/10.1162/neco_a_01125 | PubMed

[60] **Cohen Marlene R,** John H R Maunsell (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* **12**:1594-1600 https://doi.org/10.1038/nn.2439 | PubMed

[61] **McGinley Matthew J,** Vinck Martin, Reimer Jacob, Batista-Brito Renata, Zagha Edward, Cadwell Cathryn R, Tolias Andreas S, Cardin Jessica A, McCormick David A (2015) Waking state: Rapid variations modulate neural and behavioral responses. *Neuron* **87**:1143-1161 https://doi.org/10.1016/j.neuron.2015.09.012 | PubMed

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

[63] **Hilbe Joseph M** (2011) *Negative Binomial Regression* (2) Cambridge University Press.

[64] **Cameron A Colin,** Trivedi Pravin K (2013) *Regression Analysis of Count Data* (2) Cambridge University Press. https://doi.org/10.1017/cbo9781139013567

[65] **Lindsey J k** (1995) *Modelling Frequency and Count Data* Oxford University Press. https://doi.org/10.1093/oso/9780198523314.001.0001 | https://doi.org/10.1093/oso/9780198523314.001.0001

[66] **Anuththara Rupasinghe** (2024) Continous partitioning of neuronal variability: Matlab codes. GitHub. https://github.com/Anuththara-Rupasinghe/CMP

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

[68] **Tong Yung Liang** (2012) *The multivariate normal distribution* Springer Science & Business Media.

[69] **Tokdar Surya T,** Kass Robert E (2010) Importance sampling: a review. *WIREs Computational Statistics* **2**:54-60 https://doi.org/10.1002/wics.56

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

[71] **Rasmussen Carl Edward,** Williams Christopher KI (2006) *Gaussian processes for machine learning* MIT Press.

[72] **Schoenberg IJ** (1938) Metric spaces and positive definite functions. *Transactions of the American Mathematical Society* **44**:522-536 https://doi.org/10.1090/s0002-9947-1938-1501980-0

[73] **Folland GB** (2013) *Real Analysis: Modern Techniques and Their Applications* Wiley.

[74] **Casella George,** Berger Roger L (2002) *Statistical Inference* (2nd) Pacific Grove, CA: Duxbury Press.

[75] **Robert Christian P,** Casella George, Casella George (1999) *Monte Carlo statistical methods* **2** Springer.

Peer reviews

Reviewer #1 (Public review):

Summary:

In this manuscript, Rupasinghe and co-authors introduce a new statistical model for spiking neurons. Building on earlier work, they propose to model spikes as arising from a Poisson process whereby the firing rate is the product of stimulus drive and a stimulus-independent gain signal. The critical innovation of this work is that the gain signal is modeled in continuous time. Earlier explorations of this statistical construction treated the gain-signal as

constant within a trial. This innovation is elegant and important. It makes the model richer, more plausible, and more broadly applicable. The authors show that the model parameters are recoverable from realistic amounts of data and then apply the framework to previously studied datasets. They show that the new model outperforms earlier models and alternative candidates in capturing spiking data across four visual areas of the macaque monkey. Analysis of the model parameters replicates some earlier findings and uncovers several new insights. The model and fitting methods can be broadly applied to partition different types of signals and noise from spiking data and are likely to be widely adopted in the systems neuroscience community.

Strengths:

(1) Through clever use of advanced statistical techniques, the authors manage to infer critical information from single trial single cell data.

(2) The question of which aspect of a spike train is signal and which is noise is omnipresent in neuroscience. By improving our ability to characterize the distinct factors that shape spiking activity, this work makes a fundamental contribution to the literature.

Weaknesses:

(1) The work is entirely focused on single cell data. While this is a great starting point, expanding the approach to spiking activity in neural populations is an important future goal. The discussion lays out a roadmap towards this goal.

Comments on revised version.

I thank the authors for their sincere engagement with the reviews. They have addressed all issues I had raised. I found the first version of the manuscript already impressive. The revised version is a bit clearer about the exact relationship to some prior work and now documents additional new findings that validate the successful partitioning of signal and noise and directly connect stimulus-induced variability quenching to the stabilization of the latent gain signal. This makes it a really great paper.

https://doi.org/10.7554/eLife.109719.2.sa2

Reviewer #2 (Public review):

Summary:

Neurons have varied responses to external stimuli that cannot be explained by naive Poisson models. Previous work has quantified and partitioned higher-than-Poisson variability in the brain into different components. The authors improve on these methods to infer how both the stimulus drive and internal gain dynamics impact neuronal variability continuously in time. The clean and well-reasoned model is rigorously developed and then applied to neural data across the visual hierarchy. This lends new insights into how variability is partitioned, agreeing with and extending previous work on how that variability changes from early visual areas (LGN, V1) through to higher, motion-sensitive areas (area MT). Another key contribution is that this partitioning can be fully addressed as a continuous-time process, which allows for dissection of how the timescale of fluctuations in these two components changes across the brain's processing arc.

Strengths:

(1) The model is cleanly derived and thoroughly documented, including useable code shared in a GitHub repo. This makes the method immediately portable to other neural systems.

(2) The figures and writing are clear and understandable and all pieces of the derivations are included in the main text and supplementary information.

(3) Comparisons to other models, particularly the one from Goris et al., 2014 shows how this Continuous Modulated Poisson (CMP) model outperforms previous work.

(4) New insights about how variability partitioning changes across the visual stream from LGN to MT are revealed, including how the gain fluctuates on longer timescales in higher visual areas. Another key result about the anticorrelation between the variance in stimulus drive and gain fluctuations comports with theories about how neurons maintain efficient, reliable encoding.

(5) In addition to the results reported here, this work will serve as an excellent tutorial for students and postdocs first delving into the sources of variability in the brain.

Weaknesses:

(1) The work builds off previous studies of the partitioning of variability in the brain, but provides important new extensions as noted above. Sub-poisson variability cannot be addressed in the current framework, but ideas for extensions are included in the Discussion.

Comments on revised version.

The revisions have thoroughly addressed my previous comments and concerns and the paper's clarity and scope have improved.

https://doi.org/10.7554/eLife.109719.2.sa1

Author response:

The following is the authors' response to the original reviews.

In the revised manuscript, we have expanded the real-data analyses, clarified the relationship between CMP and prior modulated Poisson models, and added discussion of model limitations and future extensions. In summary, the major changes include:

(1) We revised the Introduction, Results, Methods, and Goris-model appendix to clarify the relationship between CMP and prior modulated Poisson models. In particular, we now emphasize that the key distinction is CMP's continuous-time stochastic gain process.

(2) We moved the simulation-based recoverability analysis from Appendix 3 into the main Results section (Figure 3 in the revised manuscript), making the validation of the inference procedure more visible to readers.

(3) We added new analyses of the inferred gain process. Specifically, we now show the cross-trial gain mean and cross-trial gain variance in Figure 4A to assess whether gain captures stimulus-locked structure, and we added an analysis of pre- versus post-stimulus cross-trial gain variability in Figure 4C to test for gain-variability quenching during stimulus presentation.

(4) We clarified the definitions and implementation of the Baseline Poisson, Poisson-GP, and Goris-style comparison models, including the role of the smoothness prior on the stimulus drive.

(5) We expanded the Discussion to describe future extensions to population recordings, including a GPFA-inspired extension with low-dimensional shared gain activity across neurons.

(6) We added a Discussion paragraph clarifying that the current CMP model captures Poisson and super-Poisson variability, but not sub-Poisson variability, and outlined possible extensions using spike-history terms, renewal-process likelihoods, or alternative count distributions.

> ***eLife Assessment***
>
> *This work of fundamental significance introduces a novel statistical model of spiking activity that incorporates continuous−time gain modulation. The authors provide exceptional evidence that the model outperforms earlier approaches and alternative candidates in capturing spiking responses across multiple visual areas in the macaque. Beyond its methodological contribution, the study offers new insights into how stimulus−driven variability and internally generated gain fluctuations evolve over time and between brain areas. The framework is likely to find broad application beyond the datasets examined here.*

We sincerely thank the Senior Editor, Reviewing Editor, and both reviewers for their careful evaluation and constructive feedback. We are encouraged by the positive assessment of the work and by the recognition of its methodological and conceptual contributions. We especially appreciate the acknowledgement that the continuous-time formulation provides a useful framework for modeling gain modulation in spiking activity, improves upon earlier approaches in capturing responses across multiple visual areas, and offers new insights into how stimulus-driven variability and internally generated gain fluctuations evolve over time and across brain regions.

In the revised manuscript, we have addressed the reviewers' comments by clarifying the relationship between CMP and prior modulated Poisson models, strengthening the presentation of the simulation-based recoverability analysis, adding new validation analyses of the inferred gain process, and expanding the Discussion of model scope, limitations, and future directions. In particular, we now more clearly distinguish the continuous-time gain process in CMP from Goris-style models with constant or piecewise-constant gain, move the simulation recoverability analysis into the main Results, examine trial-averaged inferred gain and gain-variability quenching, clarify the definitions of the baseline and comparison models, and discuss extensions to population recordings and sub-Poisson variability.

We believe these revisions improve the clarity, rigour, and scope of the manuscript. Below, we address each reviewer comment in turn and describe the corresponding changes made in the revised manuscript.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *In this manuscript, Rupasinghe and co−authors introduce a new statistical model for spiking neurons. Building on earlier work, they propose to model spikes as arising from a Poisson process whereby the firing rate is the product of stimulus drive and astimulus−independent gain signal. The critical innovation of this work is that the gain signal is modeled in continuous time. Earlier explorations of this statistical construction treated the gain−signal as constant within a trial. This innovation is elegant and important. It makes the model richer, more plausible, and more broadly applicable. The authors show that the model parameters are recoverable from realistic amounts of data and then apply the framework to previously studied datasets. They show that the new model outperforms earlier models and alternative candidates in capturing spiking data across four visual areas of the macaque monkey. Analysis of the model parameters*

> *replicates some earlier findings and uncovers several new insights. The model and fitting methods can be broadly applied to partition different types of signals and noise from spiking data and are likely to be widely adopted in the systems neuroscience community.*
>
> *Strengths:*
>
> *(1) Through clever use of advanced statistical techniques, the authors manage to infer critical information from single−trial single−cell data.*
>
> *(2) The question of which aspect of a spike train is signal and which is noise is omnipresent in neuroscience. By improving our ability to characterize the distinct factors that shape spiking activity, this work makes a fundamental contribution to the literature.*

We sincerely thank the reviewer for the thoughtful and detailed evaluation of our manuscript. We are pleased that the continuous-time formulation and its methodological contributions were viewed as elegant, important, and broadly applicable. We also appreciate the reviewer's recognition that the framework provides a useful way to separate stimulus-driven and modulatory components of neural variability from single-trial, single-cell data. The reviewer's comments helped us improve the precision of our framing, clarify the relationship between CMP and prior modulated Poisson models, and strengthen the validation of the inferred gain process. Below, we respond to each point in turn and describe the revisions made in the manuscript.

> *Weaknesses:*
>
> *Overall, I find the work impressive and important. I have a couple of questions and suggestions.*
>
> *(1) The work is entirely focused on single−cell data. While this is a great starting point, expanding the approach to spiking activity in neural populations is an importantfuture goal.*

We thank the reviewer for this important suggestion. We agree that extending the CMP framework to population recordings is a natural and important direction for future work. In the present study, we focus on single-neuron responses to establish the continuous-time model, validate the inference, and characterize how stimulus-driven activity and stochastic gain fluctuations can be separated at the level of individual cells. However, the same modeling principles could be extended to simultaneously recorded neural populations by introducing shared latent structure across neurons. For example, one natural direction would be to combine CMP with ideas from Gaussian Process Factor Analysis [Keeley et al., 2020], using low-dimensional shared gain activity to capture population-wide fluctuations, while retaining neuron-specific stimulus-driven components. Such an extension would allow the model to capture correlated variability and shared modulatory dynamics across neural ensembles. In the revised manuscript, we have expanded the Discussion to describe this possible future extension to population recordings.

To address this comment, we expanded the Discussion (Page 14: lines 473-478) to describe a possible GPFA-inspired extension of CMP to population recordings.

> *(2) Line 49−53: These statements seem incorrect to me. The modulated Poisson model , as introduced in Goris et al (2014), is a process model that can perfectly be used to generate spike trains (within a trial, spiking emerges from a Poisson process, which canbe homogeneous or inhomogeneous). Moreover, the model contains a parameter thatrepresents the duration of the counting window (delta t). The dependency of over−dispersion on the size of the time bins for real neurons is shown in Figure 1b (inset plot) of that paper (and shown to resemble the model prediction). This time− dependency was further explored by the same authors in Goris et al (2018 − Journal ofVision) and also in*

> *Henaff et al (2020 − Nature Communications). I suggest that the authors rephrase this argument (here and at some later points in the paper). They could just say that the Goris model makes the simplistic and implausible assumption that, within a given trial, gain does not fluctuate. This is clearly an important limitation and the key difference with the continuous model introduced here.*

We sincerely thank the reviewer for identifying this lack of clarity in our original description. We agree that our original description was not sufficiently precise. The modulated Poisson model introduced by Goris et al. (2014) is indeed a generative process model and can be used to generate spike trains, with spiking arising from a Poisson process that may be homogeneous or inhomogeneous within a trial. We apologize for implying otherwise.

Our intended point was that, in the original formulation, the modulatory gain is represented as a scalar random variable associated with a counting window or trial, and therefore does not explicitly model gain as a continuously time-varying process within a trial. Thus, the key limitation addressed by CMP is not the use of a Poisson process, but the assumption that gain is constant or piecewise constant over the relevant interval.

In the revised manuscript, we have rephrased the Introduction to clarify this distinction. We now describe the Goris model more accurately as a modulated Poisson framework in which gain is constant over the counting window, and we emphasize that CMP extends this framework by replacing this assumption with a continuous-time stochastic gain process. We have also added discussion of related time-dependent analyses and extensions [Goris et al., 2018, Hénaff et al., 2020], as thoughtfully suggested by the reviewer.

In addition, we revised the Results and Methods to clarify how the Goris-style baselines were implemented in our comparisons. Specifically, all Goris-style results reported in the main model comparisons use versions with a smoothness prior on the stimulus drive, where the stimulus-dependent firing rates are set to the smooth firing-rate estimates obtained from the Poisson-GP model. This ensures that the comparisons focus on different assumptions about the temporal structure of the gain process, rather than differences in stimulus-drive estimation. We also clarified the comparison to Goris-style variants without this smoothness prior, in which the stimulus-drive parameters are estimated directly under the corresponding Goris-style likelihood (Figure 5 - figure supplement 2). These results show that the smoothness prior on the stimulus drive substantially improves model performance. Finally, we revised the Figure 1 caption and the Goris-model appendix to make these distinctions explicit.

To address this comment, we revised the Introduction (Pages 2-3: Lines 49-77), Results (Page 9: Lines 263-266, 273-277, Page 11: Lines 319-326), Methods (Page 21), Figure 1 caption, and Goris-model appendix to clarify that CMP extends the Goris framework by modeling gain as a continuously time-varying process within trials.

> *(3) Line 54−55: I think the first part of the claim is a bit misleading. There is nothing in the Goris model that would inherently limit it to homogeneous Poisson processes, as seems to be implied by this description. The model is built on the assumption thatspike generation within a trial arises from a Poisson process. This may very well be an inhomogeneous Poisson process (i.e., a stimulus−dependent time−varying firing rate). Homogeneous and inhomogeneous Poisson processes both give rise to Poisson distributed spike counts (and thus a mixture of Poisson distributions across trials in the Goris model). I suggest the authors clarify this description a bit. Note that the two model variants illustrated in Figure 1b and c were also explored in Henaff et al (2020 − Nature Communications).*

We thank the reviewer for this helpful clarification. We agree that the Goris model is not limited to homogeneous Poisson spiking and can incorporate a stimulus-dependent, time-

varying firing rate within trials. We did not intend to imply otherwise, and we have revised the relevant text to avoid this misunderstanding.

Our intended point was that, in formulating continuous-time extensions of the modulated Poisson framework, we explicitly model the time-varying stimulus drive using a smoothness prior, as in the CMP framework, and then consider different assumptions about the temporal structure of the gain process, including constant gain and independently resampled gain across time bins. This highlights the distinction between piecewise-constant gain assumptions and the fully continuous gain process introduced in CMP.

In the revised manuscript, we have clarified this distinction in the Introduction, Results, and Methods. We now state that the Goris-style variants use stimulus-dependent, time-varying Poisson firing rates, and that the main difference between these variants and CMP lies in the temporal structure assumed for the gain process. We have also acknowledged related variants explored in Goris et al. [2018] and Hénaff et al. [2020], and clarified that our continuous-time formulations of the Goris model differs by imposing a smoothness prior on the stimulus drive. This allows us to estimate a regularized time-varying stimulus component while comparing different assumptions about gain dynamics, ensuring that the comparison focuses on the temporal structure of the gain process rather than differences in stimulus-drive estimation. We also highlight in Figure 5 - figure supplement 2 that even for the Goris-style models, versions that use a smoothness prior on the stimulus drive outperform versions that do not, which are closer to the original modulated Poisson formulation.

To address this comment, we revised the Introduction (Pages 2-3: Lines 49-77), Results (Page 9: Lines 263-266, 273-277, Page 11: Lines 319-326), Methods (Page 21) to clarify that the Goris-style variants allow stimulus-dependent time-varying firing rates and differ from CMP primarily in their assumptions about gain dynamics. We also added citations to related time-dependent extensions of the modulated Poisson framework.

> *(4) The extension to the continuous case is very elegant!*

We thank the reviewer for the positive comment and are pleased that the continuous-time formulation was viewed as elegant.

> *(5) I find the result shown in Appendix 3 critically important. The recoverability of the model for realistic amounts of data is foundational for the rest of the paper. I wouldconsider including this analysis in the main results section. Not all readers may check Appendix 3, but they should know about this result.*

We thank the reviewer for emphasizing the importance of this result. We agree that demonstrating parameter recoverability is foundational to the paper and should be visible to readers in the main Results section. In the revised manuscript, we have moved the simulation-based validation from Appendix 3 into the main Results. This section now describes the synthetic CMP dataset, the inference procedure used to estimate the latent stimulus-drive and gain processes, and the comparison between true and inferred GP hyperparameters. These results show that the proposed inference framework can accurately recover the ground-truth stimulus drives, gain processes, and hyperparameters from realistic amounts of simulated data.

To address this comment, we moved the simulation-based recoverability analysis from Appendix 3 into the main Results section (Page 6: Lines 200-211 and Figure 3).

> *(6) Figure 3: I am wondering whether the inferred gain is capturing some response fluctuations that originate from the cell's phase−selectivity. Could the authors compute the trial−averaged inferred gain (ideally, aligned to stimulus−phase at the start of the trial if this experimental parameter varied across repeats)? If they have successfully partitioned the response variance, the trial−averaged gain should have no systematic*

> *temporal structure. If it has a sinusoidal modulation, it may partially capture stimulus−drive. This could be an interesting test to run on all model fits to further validate that the partitioning into a signal and noise component succeeded as intended.*

We thank the reviewer for this insightful suggestion. We agree that verifying that the inferred gain does not capture stimulus-driven structure is an important validation of the model. In the revised manuscript, we have added the trial-averaged inferred gain to Figure 4A for the example neuron. This analysis shows that the trial-averaged inferred gain is relatively flat and neither resembles the inferred stimulus drive nor exhibits clear stimulus-locked temporal structure. This suggests that trial-specific gain fluctuations largely average out across repeats, consistent with the interpretation that the gain process captures random trial-to-trial variability rather than stimulus-driven activity.

We also note that a direct comparison of this inferred gain trace across methods is not possible for the Goris-style baselines, because these models do not infer a continuous trial-specific gain process. Instead, they marginalize over scalar or time-bin-independent gain variables when computing likelihoods and Fano factor curves. Thus, the trial-averaged gain diagnostic is specific to the CMP model, where the posterior over the continuous-time gain process is explicitly inferred.

To address this comment, we added the trial-averaged inferred gain to Figure 4A and clarified that it does not show a clear stimulus-locked temporal structure (Page 7: Lines 229-236).

> *(7) One common observation that is currently not explored is the quenching of neuronal response variability following stimulus onset (Churchland et al 2010 − NatureNeuroscience), which was suggested to reflect a quenching of gain variability in Goris et al (2024 − Nature Reviews Neuroscience). Building on the previous suggestion, the authors could compute the temporal evolution of cross−trial gain variability from the inferred gain traces. Do they recognize a reduction in gain variability following stimulus onset? If so, it would be worthwhile to show this.*

We sincerely thank the reviewer for this valuable suggestion. We agree that examining whether gain variability decreases following stimulus onset provides an important test of the inferred gain process. In the revised manuscript, we have added an analysis of the temporal evolution of cross-trial gain variability before and after stimulus onset.

First, in Figure 4A, we now show the cross-trial variance of the inferred gain for the example neuron. This trace shows larger gain variability during the stimulus-off period and a reduction following stimulus onset, suggesting that the inferred gain captures a stimulus-related quenching of trial-to-trial variability. To quantify this effect across the population, we also added a pre- versus post-stimulus comparison in Figure 4C. Following the approach of Churchland et al. [2010], we compared gain variability in two matched 400-ms windows: a pre-stimulus window ending at stimulus onset and a stimulus-period window beginning 100 ms after stimulus onset. For each neuron and stimulus condition, we computed the cross-trial variance of the inferred gain at each time bin, averaged this quantity within each window, and then compared the pre- and post-stimulus values across neuron-stimulus pairs.

This analysis revealed a significant reduction in inferred gain variability following stimulus onset (one-sided paired Wilcoxon signed-rank test, $p \leq 10^{-15}$), consistent with gain variability quenching [Churchland et al., 2010, Goris et al., 2024]. We now report this result in the main text and illustrate it in Figure 4A and Figure 4C. This provides additional evidence that the inferred CMP gain captures meaningful trial-to-trial variability and its temporal modulation around stimulus presentation.

To address this comment, we added the cross-trial gain variance trace to Figure 4A and a

population-level pre- versus post-stimulus gain-variability quenching analysis (Page 7 and 8: Lines 239-248) in Figure 4C.

> *(8) Line 543−565: I want to make sure I understand the Baseline Poisson model and Poisson−GP correctly. For the baseline model, I had imagined that the authors would simply use the stimulus−conditioned PSTH as an estimate of the time−dependent firing rate, coupled with an inhomogeneous Poisson process assumption. But they additionally assume a Gamma prior on the firing rate to compensate for the sparsenessof the data (sometimes only 5 repeats per condition). The Poisson−GP includesexactly the same model components, but now the time−dependent firing rate is modeled by a Gaussian process. Doing this massively improves the goodness−of−fit (Fig 4A). Do I understand this correctly?*

We thank the reviewer for this careful reading. Yes, this understanding is broadly correct, and we have revised the manuscript to clarify the relationships among the Baseline Poisson, Poisson-GP, and Goris-style models. The Baseline Poisson model estimates a stimulus- and time-dependent firing rate independently for each stimulus condition and time bin, using a Gamma-Poisson formulation to regularize the estimate when the number of repeats is limited. The Poisson-GP model uses the same conditionally Poisson observation model, but replaces these independent time-bin-wise rate estimates with a smooth stimulus-specific Gaussian process model for the log firing rate.

We have also clarified how the Goris-style models were implemented. All Goris-style results reported in the main model comparisons use versions with a GP prior on the stimulus drive. In these versions, the stimulus-dependent firing rates are set to the smooth firing-rate estimates obtained from the PoissonGP model, and the gain parameters are then fit under either the independent-gain or constant-gain assumptions. We used these GP-smoothed versions as stronger baselines. In Figure 5, Figure Supplement 2, we additionally compare these models to Goris-style variants without the GP prior on the stimulus drive, in which the stimulus-drive parameters are estimated directly under the corresponding Goris-style likelihood. This comparison shows that adding a GP smoothness prior to the stimulus drive substantially improves held-out model fit. Together, these analyses clarify that the GP-smoothed stimulus drive improves the Goris-style baselines, while the continuous-time gain process in CMP provides an additional improvement by capturing temporally structured trial-to-trial variability.

To address this comment, we clarified the definitions of the Baseline Poisson, Poisson-GP, and Goris-style models (Pages 8-9: Lines 255-259, 263-266, 273-277), and revised the text (Page 11: Lines 319-326) describing Figure 4 - figure Supplement 2 to make explicit how this existing comparison isolates the effect of the GP prior on the stimulus drive.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *Neurons have varied responses to external stimuli that cannot be explained by naive Poisson models. Previous work has quantified and partitioned higher−than−Poisson variability in the brain into different components. The authors improve on these methods to infer how both the stimulus drive and internal gain dynamics impact neuronal variability continuously in time. The clean and well−reasoned model is rigorously developed and then applied to neural data across the visual hierarchy. This lends new insights into how variability is partitioned, agreeing with and extending previous work on how that variability changes from early visual areas (LGN, V1) through to higher, motion−sensitive areas (area MT). Another key contribution is that this partitioning can be fully addressed as a continuous−time process, which allows for the dissection of how*

> *the timescale of fluctuations in these two components changesacross the brain's processing arc.*
>
> *Strengths:*
>
> *(1) The model is cleanly derived and thoroughly documented, including usable code shared in a GitHub repo. This makes the method immediately portable to other neural systems.*
>
> *(2) This is a clear and well−presented piece of work. The figures and writing are clear and understandable, and all pieces of the derivations are included in the main text and supplementary information.*
>
> *(3) Comparisons to other models, particularly the one from Goris et al., 2014 shows how this Continuous Modulated Poisson (CMP) model outperforms previous work.*
>
> *(4) New insights about how variability partitioning changes across the visual stream from LGN to MT are revealed, including how the gain fluctuates on longer timescales in higher visual areas. Another key result about the anticorrelation between the variance in stimulus drive and gain fluctuations comports with theories about how neurons maintain efficient, reliable encoding.*
>
> *(5) In addition to the results reported here, this work will serve as an excellent tutorial for students and postdocs first delving into the sources of variability in the brain.*

We sincerely thank the reviewer for the thoughtful and positive assessment of our work. We are pleased that the model development, empirical analyses, and presentation were viewed as clear, rigorous, and useful for the broader neuroscience community. We also appreciate the reviewer's recognition that the continuous-time formulation meaningfully extends prior variability-partitioning approaches by allowing stimulus drive and internal gain dynamics to be characterized across temporal scales. The reviewer's comments helped us further clarify the positioning of the work, expand the Discussion of model scope and limitations, and better articulate future extensions. Below, we address the specific suggestions raised by the reviewer and describe the revisions made in the manuscript.

> *Weaknesses:*
>
> *The work is somewhat incremental, building on previous studies of the partitioning of variability in the brain, but it provides important new extensions, as noted above.*

Regarding the comment on incremental contribution, we agree that our framework builds directly on previous variability-partitioning approaches, especially the modulated Poisson framework of Goris et al. However, the main goal of this work is to move this class of models from a count-based formulation to a continuous-time spike-train framework. This extension is important because it allows us to model gain as a temporally structured latent process, characterize how variability depends on the timescale over which spikes are counted, and infer the temporal covariance structure of stimulus-independent fluctuations. In addition, the CMP framework provides analytic expressions for the Fano factor as a function of bin size, introduces the EPL covariance function for slowly decaying gain dynamics, and enables direct comparisons of gain amplitude and timescale across visual areas. In the revised manuscript, we have clarified this positioning and emphasized how CMP extends prior variability-partitioning models while preserving their interpretability.

To address this comment, we revised the Introduction (Pages 3-4: Lines 108-111 and Lines 118-121) and Discussion (Page 13: Lines 425-429) to clarify better how CMP builds on prior variability-partitioning models while extending them to continuous-time spike-train data.

> *The only major gap I would suggest addressing in the Discussion is the observation of sub−Poisson variability in the brain. It seems clear that this model can extend to sub−Poisson variability and its partitioning and perhaps even show how that varies in real time, with an animal's attentional state. That is, of course, beyond the scope of the current work, but could be mentioned in the Discussion.*

We thank the reviewer for this suggestion. We agree that sub-Poisson variability is an important phenomenon observed in neural data. Because the CMP model uses a conditionally Poisson observation model with stochastic gain modulation, it naturally captures Poisson and super-Poisson variability but does not generate sub-Poisson spike count statistics in its current form. In the revised manuscript, we have clarified this limitation in the Discussion and outlined possible extensions that could address sub-Poisson variability, including spike-history terms, renewal-process likelihoods, and alternative count distributions [Truccolo et al., 2005, Paninski et al., 2007, Aghamohammadi et al., 2024]. We also note that such extensions could allow future models to examine how sub-Poisson and super-Poisson components vary with behavioral state, attention, or arousal.

To address this comment, we added a Discussion paragraph describing the current model's limitation for sub-Poisson variability and possible extensions to capture it (Page 14: Lines 460-471).

References

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.

Robbe L. T. Goris, Corey M. Ziemba, J. Anthony Movshon, and Eero P. Simoncelli. Slow gain fluctuations limit benefits of temporal integration in visual cortex. Journal of Vision, 18(8):8–8, 08 2018. ISSN 1534-7362. doi: 10.1167/18.8.8. URL *https://doi.org/10.1167/18.8.8*.

Olivier J Hénaff, Zoe M Boundy-Singer, Kristof Meding, Corey M Ziemba, and Robbe L T Goris. Representation of visual uncertainty through neural gain variability. Nat. Commun., 11(1):2513, May 2020.

Mark M Churchland, Byron M Yu, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, Benjamin B Scott, David C Bradley, Matthew A Smith, Adam Kohn, J Anthony Movshon, Katherine M Armstrong, Tirin Moore, Steve W Chang, Lawrence H Snyder, Stephen G Lisberger, Nicholas J Priebe, Ian M Finn, David Ferster, Stephen I Ryu, Gopal Santhanam, Maneesh Sahani, and Krishna V Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat. Neurosci., 13(3):369–378, March 2010.

Robbe L T Goris, Ruben Coen-Cagli, Kenneth D Miller, Nicholas J Priebe, and Máté Lengyel. Response sub-additivity and variability quenching in visual cortex. Nat. Rev. Neurosci., 25(4):237–252, April 2024.

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.

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*.

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

https://doi.org/10.7554/eLife.109719.2.sa0