www.sciencemag.org/content/349/6244/184/suppl/DC1

Supplementary Materials for

Single-trial spike trains in parietal cortex reveal discrete steps during decision-making

Kenneth W. Latimer, Jacob L. Yates, Miriam L. R. Meister, Alexander C. Huk, Jonathan W. Pillow*

*Corresponding author. E-mail: pillow@princeton.edu

Published 10 July 2015, *Science* **349**, 184 (2015)
DOI: 10.1126/science.aaa4056

This PDF file includes:

Supplementary Materials for

Single-trial Spike Trains in Parietal Cortex Reveal Discrete Steps During Decision-making

Kenneth W. Latimer, Jacob L. Yates, Miriam L. R. Meister, Alexander C. Huk, & Jonathan W. Pillow*

*To whom correspondence should be addressed; E-mail: pillow@princeton.edu.

Contents

1 Materials and Methods - Behavior and electrophysiology

The data shown here have been reported previously (*14*). Here we briefly summarize the task and data collection. Single units were recorded while a monkey performed a moving-dot direction-discrimination task. The dot motion was displayed for random times, uniformly distributed from 500 to 1000 ms. The motion was directed towards one of the two choice targets: a target placed inside the cell's response field (in-RF target), and one placed diametrically opposite (out-RF target). 500 ms after the motion stimulus ended, a go-signal was provided and the monkey made a saccade to one of two choice targets, indicating its choice of motion direction. Dot coherence (the strength of the motion stimulus) was varied across trials. Dot coherence was drawn from conventional values of 0, 3.2, 6.4, 12.8, 25.6, and 51.2%. Coherence levels for each trial were selected uniformly random and motion direction was sampled independently with a 50% chance of an in-RF direction. To simplify our analysis, we collapsed the coherences into 5 levels: zero=0%, positive/negative high =$\{25.6, 51.2\}\%$, and positive/negative low =$\{3.2, 6.4, 12.8\}\%$ (positive values indicate motion towards the in-RF target and negative values towards the out-RF target). The original study included trials in which the choice targets were displayed during the entire trial and trials in which the targets were flashed briefly before motion onset. We chose to include only trials for which the targets were displayed throughout the trial.

In the original study (*14*), 80 spatially selective LIP cells were recorded from 2 adult, male rhesus macaques (*M. mulatta*; 14 from monkey J, 66 from monkey P). The full population included cells with a range of response styles, including cells selective primarily for motor response. For this study, we wished to include only cells with choice selectivity during the motion epoch. We used a d' analysis to quantify choice selectivity in the spike counts of the cells during the period 200-700 ms after dot motion onset (before the go signal was given on any trial). The d' value measures choice selectivity for a single cell as

$$d' = \frac{\mu_{in} - \mu_{out}}{\sqrt{\frac{1}{2}\left(\sigma_{in}^2 + \sigma_{out}^2\right)}} \tag{1}$$

where μ_{in} and μ_{out} are the mean spike counts on the in-RF and out-RF trials respectively. The variance of the spike counts on the in-RF and out-RF trials are σ_{in}^2 and σ_{out}^2. We selected the top 50% of cells (40 cells, 6 from monkey J and 34 from monkey P). The d' of the 40 selected cells ranged from 0.397 to 1.661 with mean 0.819 and standard deviation 0.359.

For our statistical analyses, we used the spike trains for each trial starting ~200ms after motion onset (assuming a 200 ms latency of the decision-related activity to appear in LIP (*15*)) until 200 ms after motion offset. Therefore, the spike trains we used for fitting were 500-1000 ms long, uniformly distributed on this interval. We repeated the analyses with earlier start times and achieved similar results (Sec. 3.4).

The mean number of trials per cell was 385 (std=148.4, range=[96 750]). The mean number of spikes observed on each trial was 12.4 (std=11.7, range=[0 116]). This set of cells was qualitatively similar to those that have received focus in other studies, and exhibit response profiles (shown in Figures S13- S15) similar to those present in the well-studied Roitman & Shadlen dataset (Fig. S22).

We did not attempt to elaborate our models with an urgency signal, as introduced by (*15*). Our LIP responses did not exhibit an urgency signal, originally defined as the residuals between the data and an unbiased diffusion model, and in practice characterized as the upward deflection of responses for 0% coherence trials not conditioned on choice. The latter is flat in our dataset (e.g., Figure 3A).

2 Materials and Methods - Modeling

Here we define the two models of LIP spike train responses during decision-making. The models define an observed spike train as a Poisson process with rate determined by an unobserved (latent) noise process on each trial, yielding a doubly stochastic model. We fit both models to each cell independently.

We use Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution of model parameters given the observed spike trains. These methods provide samples from the posterior distribution over the model parameters by alternately sampling the model parameters and the latent variables for every trial. The resulting samples of the model parameters approximate the posterior distribution of parameters given the data, marginalized over the latent variables. This allows us to include uncertainty in our parameter estimates, and to avoid any approximations that would be required to fit these non-linear, non-Gaussian models with deterministic algorithms such as expectation-maximization. We fit the models to 90% of the trials for each neuron (selecting 90% from each coherence group), and held out 10% for computing cross-validation statistics such as predictive log-likelihood. The results of this analysis were consistent with the results obtained from DIC and Bayes Factors, but much more costly to compute because they required 10 folds of fitting and validation, and so we have not included them here.

We ran the MCMC algorithms for a total of $60\,000$ iterations and discarded the first $10\,000$ samples (the "burn in" period), to ensure that the Markov Chain had converged to its asymptotic distribution, the true posterior. With the remaining samples, we only took every 5th sample (a procedure known as "thinning" the chain) in order to reduce autocorrelation (or increase independence) between samples. Thus, we effectively obtained $10\,000$ samples from the posterior distribution of model parameters. We used these samples to represent the posterior distribution for performing model comparison analyses (Figs. 3B, S6-S7). For all other analyses, we used the posterior mean (mean of the $10\,000$ samples) as a point estimate of the model parameters.

We implemented the sampling algorithms on a GPU using a combination of Matlab and CUDA. All sampling and analyses were performed on single desktop computer equipped with an Nvidia GTX Titan GPU and an Intel i7-4930K CPU (6 cores, 3.40 GHz). For the ramping (diffusion-to-bound) sampler running on a 500 trial dataset, the MCMC required 0.35 s to generate a sample. The stepping model sampler required 0.03 s per iteration. These times can be compared with our original CPU-only Matlab implementation of the ramping MCMC algorithm, which required a prohibitively slow 17.6 s per sample. Our use of both C/CUDA and a modern GPU produced an implementation suitable for running on a

single desktop.

For computational convenience, we define the models in discrete time using bins of length Δ_t ($\Delta_t = 10$ ms bins here). However, we simulate spike trains at a 0.2 ms resolution, assuming a homogeneous spike rate within 10 ms bins. Here we provide some notation for the models. We denote the entire set of spike trains (for the single neuron being modeled) as $\mathbf{y}$, and the spike counts at time bin t in trial j as $y_{j,t}$, with trials numbered 1 to N. Some parameters depend on the stimulus coherence in the trial. We consider stimulus coherences as categorical, rather than assuming a functional form for the stimulus-dependent parameters. The total number of categories is denoted C (here, $C = 5$). The coherence for trial j is $c(j)$. Some model parameters depend on stimulus coherence, and a subscript (for example, p_c) denotes coherence dependence. Our analysis method allows trials to be of varying length: the length of trial j in number of discrete bins is denoted T_j.

2.1 Ramping (diffusion-to-bound) model

The spike rate in the ramping model follows a diffusion-to-bound process. The parameters of the model are $\Theta = \left\{\beta_{1:C}, x_0, \omega^2, \gamma\right\}$. The β and ω^2 parameters are the drift and diffusion terms (respectively) for the drift-diffusion process. The diffusion process starts at x_0, and the bound height is determined by γ. The drift-diffusion process $x_{j,1:T_j}$ determines the spike rate for trial j. The full model can be described:

$$x_{j,1} = x_0 + \epsilon_{j,0} \tag{2}$$

$$x_{j,t+1} = x_{j,t} + \beta_{c(j)} + \epsilon_{j,t} \tag{3}$$

$$\epsilon_{j,t} \sim \mathcal{N}(0, \omega^2) \tag{4}$$

$$\tau_j = \begin{cases} \inf_t x_{j,t} \geq 1 & : \text{ if there exists } x_{j,1:T_j} \geq 1 \\ \infty & : \text{ otherwise} \end{cases} \tag{5}$$

$$y_{j,t} | t < \tau_j \sim \text{Poisson}\left(\log(1 + \exp(\gamma x_t))\Delta_t\right) \tag{6}$$

$$y_{j,t} | t \geq \tau_j \sim \text{Poisson}\left(\log(1 + \exp(\gamma))\Delta_t\right) \tag{7}$$

The drift term β is the only coherence-dependent parameter (representing the strength of evidence in the stimulus). Spike rates are kept positive with the soft-rectification function $\log(1 + \exp(\gamma x_t))$.

In our parameterization, the latent diffusion process, $x_{j,1:T_j}$, does not stop at the bound, but the spike rate is held constant after the bound crossing time τ_j. This is equivalent to a model that stops the diffusion at bound hitting time, because spike rate is constant after this time in either case. Additionally, the bound-hitting time occurs when $x_{j,t}$ crosses a constant bound at 1 - the bound height in terms of spike rate changes with the parameter γ. The transfer function makes bound in spikes per second equal to

$$\log(1 + \exp(\gamma)) \approx \gamma. \tag{8}$$

Therefore, the bound height is given by γ. Our choice in parameterization not only simplifies model inference, but it makes the noise in the integration process independent of each neuron's firing rate *a priori*.

2.1.1 Prior distributions: ramping model

The prior distributions for the model parameters take the following form

$$x_0 \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{9}$$
$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \tag{10}$$
$$\omega^2 \sim \text{Inv-Gamma}(\alpha_\omega, \beta_\omega) \tag{11}$$
$$\gamma \sim \text{Gamma}(\alpha_\gamma, \beta_\gamma) \tag{12}$$

We chose the following values for the priors

1. $\mu_0 = 0$, $\sigma_0 = 10$
2. $\mu_\beta = 0$, $\sigma_\beta = 0.1$
3. $\alpha_\omega = 0.02$, $\beta_\omega = 0.02$
4. $\alpha_\gamma = 2$, $\beta_\gamma = 0.05$

The prior over the maximum firing rate, γ, has mean 40 spikes/s with a standard deviation of 28.3, which covers the range of firing rates we expect to encounter in LIP. The typical values for the starting point of the diffusion process, x_0, should lie between 0 and 1, because the diffusion process runs to a bound of 1. The broad Gaussian prior we chose for x_0 is nearly uniform over this region. The diffusion slope, β, corresponds to motion evidence (dot coherence), which was drawn for each trial from a distribution with mean 0%. We chose the standard deviation for the β prior by considering the range of realistic time-to-bound distributions: assuming $x_0 = 0$, then if $\beta > 0.1$ the mean time-to-bound would be under 100 ms (extremely fast). Therefore, we chose the standard deviation for the prior on β to be 0.1, which places most of the prior probability mass on reasonable bound-hit times, without being too constricting.

In the MCMC section, we use the prior parameter symbols instead of these specific values so that it is clear where the priors are placed.

2.1.2 MCMC: ramping model

The sampler consists of two primary steps: (1) sampling the latent states given the previous value of the model parameters, and (2) sampling the model parameters given the newly sampled latent states. This gives a representation of the joint posterior of model parameters and latent states. By ignoring the latent state samples, we obtain an estimate of the posterior distribution over model parameters given the data alone, marginalizing over latent parameters. We sample the latent states in each step because this results in a more efficient chain in which the model parameters are easily sampled using a fixed value of the latent states, especially since the data are divided into independent trials (*23, 24*). We initialize the sampler by setting the bound height, γ, to the average spike rate in the final bin of all in-RF choice trials. The initial diffusion value x_0 was set to 0.1. The remaining parameters were set to the mode of the prior distribution.

Sampling the diffusion paths

We obtain the sth sample of the latent state for all trials $\mathbf{x}^{(s)}$ conditioned on the previous sample of the parameters, $\Theta^{(s-1)}$, and the observed spikes $\mathbf{y}$. The latent state of each trial is independent of all other trials given the model parameters and the data. Therefore, we outline the sampler for a single trial, and we drop the subscript denoting trial number for simplicity of notation.

The posterior distribution over the latent states does not have a closed form. Instead, we decompose the posterior using the Markovian structure of $\mathbf{x}$:

$$p(\mathbf{x}|\mathbf{y}, \Theta, \tau) = p(x_T|\Theta, y_{1:T}, \tau) \prod_{t=1}^{T-1} p(x_t|x_{t+1}, \Theta, y_{1:t}, \tau) \tag{13}$$

Using Bayes theorem, we compute each of the right-hand side terms as

$$p(x_t|x_{t+1}, \Theta, y_{1:t}, \tau) \propto p(x_{t+1}|x_t, \Theta, \tau) p(x_t|y_{1:t}, \Theta, \tau) \tag{14}$$

Therefore, if we can compute $p(x_t|y_{1:t}, \Theta, \tau)$ for $t = 1$ to T and $p(\tau|\Theta, y_{1:t})$, we can sample x_T from $p(x_T|\Theta, y_{1:T})$ and then work our way backwards, sampling x_t from $p(x_t|x_{t+1}, \Theta, y_{1:t}, \tau)$ for $t = \tau$ to 1 in order to obtain a sample from the complete posterior distribution, $p(\mathbf{x}, \tau|\mathbf{y}, \Theta)$.

We use a particle filter to approximate the distributions $p(x_t|y_{1:t}, \tau \geq t, \Theta^{(s-1)})$ for $t = 1$ to T (*25*). We use a set of M particles (we set $M = 200$) to approximate a series of distributions for times $t = 1$ to T. In our algorithm, a set of particles approximates the distribution of x_t for paths that have not crossed the bound by time t. At time t, particle k has position $\hat{x}_t^{(k)}$ and $w_t^{(k)}$ which form the distribution

$$p(x_t|y_{1:t}, \tau \geq t, \Theta^{(s-1)}) \approx \sum_{k=1}^{M} w_t^{(k)} \delta(x_t - \hat{x}_t^{(k)}) \tag{15}$$

where δ denotes the Dirac delta function. The weights must sum to 1 at each time ($\sum_{k=1}^{M} w_t^{(k)} = 1$). Additionally, we augment the particle filter by tracking the distribution of the bound hit time, τ, relative to time t: $P(\tau < t|y_{1:t}, \Theta^{(s-1)})$ and $P(\tau = t|y_{1:t}, \Theta^{(s-1)})$, and $P(\tau > t|y_{1:t}, \Theta^{(s-1)})$. This formulation allows us to deal with the bound without the need to track each particle's history, which could result in a high percentage of degenerate particles. Once we obtain these distributions over x_t and $\tau \geq t$ from time $t = 1$ to T, we sample the values of $x_t^{(s)}$ and $\tau^{(s)}$ by working backwards from time $t = T$ to 1.

Initially, we set $\hat{x}_0^{(k)} = x_0$ and $w_0^{(k)} = \frac{1}{M}$ and $P(\tau > 0) = 1$. Particles are propagated through time using a sequential importance resampling (SIR) algorithm. Particle positions at time $t+1$ are randomly sampled

$$\hat{x}_{t+1}^{(k)} \sim \pi(\hat{x}_{t+1}|\hat{x}_t^{(k)}, y_{t+1}, \Theta^{(s-1)}) \tag{16}$$

The particle weights are updated as

$$w_{t+1}^{(k)} \propto w_t^{(k)} \frac{p(y_{t+1}|\hat{x}_{t+1}^{(k)}, \Theta^{(s-1)}) p(\hat{x}_{t+1}^{(k)}|\hat{x}_t^{(k)}, \Theta^{(s-1)})}{\pi(\hat{x}_{t+1}^{(k)}|\hat{x}_t^{(k)}, y_t, \Theta^{(s-1)})} \tag{17}$$

Because we want the particles to track the distribution of x under the bound, we set the proposal to a truncated Gaussian with mean and variance given by the drift-diffusion model

$$\pi(\hat{x}_{t+1}|\hat{x}_t^{(k)}, y_{t+1}, \Theta^{(s-1)}) \propto \mathbf{1}_{(-\infty,1)}(\hat{x}_{t+1}) \mathcal{N}(\hat{x}_{t+1}; \hat{x}_t + \beta^{(s-1)}, \omega^{2,(s-1)}). \tag{18}$$

The numerator terms are (from the model definition)

$$p(\hat{x}_{t+1}^{(k)}|\hat{x}_t^{(k)}, \Theta^{(s-1)}) = \mathcal{N}(\hat{x}_{t+1}^{(k)}; \hat{x}_t^{(k)} + \beta^{(s-1)}, \omega^{2,(s-1)}), \tag{19}$$

$$p(y_{t+1}|\hat{x}_{t+1}^{(k)}, \Theta^{(s-1)}) = \text{Poisson}\left(y_{t+1}; \log(1 + \exp(\gamma^{(s-1)}\hat{x}_{t+1}^{(k)}))\Delta_t\right) \tag{20}$$

The bound-time, τ, is tracked through time. We assume that $P(\tau > 0) = 1$. The bound-hit time probabilities are propagated through time, from $t = 1$ to T, with the following updates

$$P(\tau < t|y_{1:t}, \Theta^{(s-1)}) \propto p(y_t|\tau < t, \Theta^{(s-1)})P(\tau < t|y_{1:t-1}, \Theta^{(s-1)}) \tag{21}$$

$$P(\tau = t|y_{1:t}, \Theta^{(s-1)}) \propto p(y_t|\tau = t, \Theta^{(s-1)})P(\tau = t|y_{1:t-1}, \Theta^{(s-1)}) \tag{22}$$

$$P(\tau > t|y_{1:t}, \Theta^{(s-1)}) \propto p(y_t|\tau > t, y_{1:t-1}, \Theta^{(s-1)})P(\tau > t|y_{1:t-1}, \Theta^{(s-1)}) \tag{23}$$

and normalizing the probabilities so that

$$P(\tau > t|y_{1:t}, \Theta^{(s-1)}) + P(\tau = t|y_{1:t}, \Theta^{(s-1)}) + P(\tau < t|y_{1:t}, \Theta^{(s-1)}) = 1. \tag{24}$$

The particle distributions provide the updated probability distribution over τ relative to time t.

$$\begin{aligned} P(\tau = t|y_{1:t-1}, \Theta^{(s-1)}) \approx\ & P(\tau > t-1|y_{1:t-1}, \Theta^{(s-1)}) \\ & \cdot \sum_{k=1}^{M} w_{t-1}^{(k)}(1 - \Phi(1; \hat{x}_t^{(k)} + \beta^{(s-1)}, \omega^{2,(s-1)}))) \end{aligned} \tag{25}$$

$$p(y_t|\tau > t, y_{1:t-1}, \Theta^{(s-1)}) \approx \sum_{k=1}^{M} w_{t-1}^{*(k)} p(y_t|x_t = \hat{x}_t^{(k)}) \tag{26}$$

$$\text{where } w_{t+1}^{*(k)} = w_t^{(k)} \frac{p(\hat{x}_{t+1}^{(k)}|\hat{x}_t^{(k)}, \Theta^{(s-1)})}{\pi(\hat{x}_{t+1}^{(k)}|\hat{x}_t^{(k)}, y_t, \Theta^{(s-1)})} \tag{27}$$

$$\begin{aligned} P(\tau > t|y_{1:t-1}, \Theta^{(s-1)}) \approx\ & P(\tau > t-1|y_{1:t-1}, \Theta^{(s-1)}) \\ & \cdot \left(\sum_{k=1}^{M} w_{t-1}^{(k)} \Phi(1; \hat{x}_t^{(k)} + \beta^{(s-1)}, \omega^{2,(s-1)})\right) \end{aligned} \tag{28}$$

$\Phi(x; \mu, \sigma^2)$ denotes the normal cumulative density function with mean μ and variance σ^2. The weights, $w_{t+1}^{*(k)}$ indicate the probability of particles $\hat{x}_{t+1}^{(k)}$ given only $y_{1:t}$ (without observing y_{t+1}).

After running the particle filter from $t = 1$ to T, we are ready to sample the latent trajectory $x_{1:T}^{(s)}$ and $\tau^{(s)}$. We accomplish this by working backwards from time T to sample the value for $\tau^{(s)}$. Once $\tau^{(s)}$ is sampled, we continue sampling backwards in time to establish the latent trajectory $x_{1:\tau}^{(s)}$.

With probability $P(\tau \leq T|y_{1:T})$ (calculated by the forward-pass) we take $\tau^{(s)} \leq T$. Otherwise, let $\tau^{(s)} = \infty$, to signify that the diffusion process did not reach the bound on this trial. If we instead have that $\tau^{(s)} \leq T$, then $\tau^{(s)}$ is sampled by working backwards from $t = T - 1$, then $t = T - 2$, and so on until an exact value for $\tau^{(s)}$ is found. We work backwards setting $\tau^{(s)} \leq t$ with probability

$$P(\tau \leq t|\tau \leq t+1, y_{1:t}, \Theta^{(s-1)}) = \frac{P(\tau \leq t|y_{1:t}, \Theta^{(s-1)})}{P(\tau \leq t|y_{1:t}, \Theta^{(s-1)}) + P(\tau = t+1|y_{1:t}, \Theta^{(s-1)})} \tag{29}$$

$$P(\tau = t+1|y_{1:t}, \Theta^{(s-1)}) \approx P(\tau > t|y_{1:t}, \Theta^{(s-1)}) \sum_{k=1}^{M} w_t^{(k)} (1 - \Phi(1; \hat{x}_t^{(k)} + \beta^{(s-1)}, \omega^{(s-1)})) \tag{30}$$

Otherwise, we set $\tau^{(s)} = t+1$.

If $\tau^{(s)} > T$, we sample a value for $x_T^{(s)}$ from the particle set at time T using the probability distribution

$$p(x_T|y_{1:T}, \tau^{(s)} > T, \Theta^{(s-1)}) \approx \sum_{k=1}^{M} w_T^{(k)} \delta(x_T - \hat{x}_T^{(k)}). \tag{31}$$

and then work backwards in time sampling $x_{1:T-1}^{(s)}$ (sampling first $x_{T-1}^{(s)}$, then $x_{T-2}^{(s)}$, and so on) as described below.

If instead $\tau^{(s)} < T$, we set the value of $x_{\tau-1}^{(s)}$ to one of the particles $\hat{x}_t^{(k)}$ where $t = \tau^{(s)} - 1$ by sampling from the distribution

$$p(x_t|\tau^{(s)}, y_{1:T}, \Theta^{(s-1)}) \propto p(\tau^{(s)} = t+1|x_t, \Theta^{(s-1)}) p(x_t|y_{1:t}, \Theta^{(s-1)}) \tag{32}$$

$$\approx \sum_{k=1}^{M} w_t^{(k)} \delta(x_t - \hat{x}_t^{(k)}) \left(1 - \Phi(1; \hat{x}_t^{(k)} + \beta^{(s-1)}, \omega^{2,(s-1)})\right). \tag{33}$$

We can then sample the remaining trajectory, $x_{1:\tau-2}^{(s)}$, by sampling backwards through time.

Backwards sampling again requires the particles. The value of $x_{t-1}^{(s)}$ given $x_t^{(s)}$ is sampled using the approximated distribution

$$p(x_{t-1}|x_t^{(s)}, y_{1:t-1}, \Theta^{(s-1)}, \tau > t-1) \propto p(x_t^{(s)}|x_{t-1}, \Theta^{(s-1)}) p(x_{1:t-1}|y_{1:t-1}, \Theta^{(s-1)}, \tau > t-1) \tag{34}$$

$$\approx \sum_{k=1}^{M} \delta(x_{t-1} - \hat{x}_{t-1}^{(k)}) \mathcal{N}(x_t^{(s)}; \hat{x}_{t-1}^{(k)} + \beta^{(s-1)}, \omega^{2,(s-1)}) w_{t-1}^{(k)} \tag{35}$$

After we have sampled all the way backwards to $x_1^{(s)}$ and if $\tau^{(s)} \leq T$, we sample $x_{\tau^{(s)}}^{(s)}$ from the truncated normal distribution

$$p(x_\tau|x_{\tau-1}^{(s)}, \Theta^{(s-1)}) \propto \mathbf{1}_{[1,\infty)} \mathcal{N}(x_\tau; x_{\tau-1}^{(s)} + \beta, \omega^2) \tag{36}$$

where $\mathbf{1}$ is the indicator function

$$\mathbf{1}_{[1,\infty)}(x) = \begin{cases} 1 & \text{if } x \in [1, \infty) \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

After $\tau^{(s)}$, the observations (spikes) no longer depend on the new values of the latent state. This independence also means that we can actually drop $x_{\tau^{(s)}+1:T}^{(s)}$ from the model parameter sampling step (everything about the spike rate is given by $x_{1:\tau^{(s)}}^{(s)}$). Dropping these terms increases sampler efficiency.

Sampling the ramping model parameters

With the sth sample of the latent states, we sample a new set of the model parameters. We first sample the parameters $x_0, \beta_{1:C}$, and ω^2. We apply Bayes' rule to the posterior

$$p(x_0, \beta_{1:C}, \omega^2|\mathbf{x}^{(s)}, \mathbf{y}) \propto p(\mathbf{x}^{(s)}|x_0, \beta_{1:C}, \omega^2) p(x_0, \beta_{1:C}, \omega^2) \tag{38}$$

in order to compute the distributions. The parameters $x_0, \beta_{1:C}$, and ω^2 conditioned on $\mathbf{x}^{(s)}$ are independent of the observations and γ. The model defines the latent paths as a simple linear-Gaussian process, and therefore the parameters can be sampled exactly using Gibbs steps. The model definition of $p(\mathbf{x}^{(s)}|x_0, \beta_{1:C}, \omega^2)$ states that the differences $(x_{j,t} - x_{j,t-1})$ are normally distributed with mean $\beta_{c(j)}$ and variance ω^2.

We sample $\beta_{1:C}^{(s)}$ and $x_0^{(s)}$ from independent Gaussian distributions given the previous diffusion variance term $\omega^{2,(s-1)}$.

$$\beta_c^{(s)}|\mathbf{x}^{(s)}, \mathbf{y}, \omega^{2,(s-1)} \sim \mathcal{N}(B \cdot A^{-1}, A^{-1}) \tag{39}$$

$$A = \frac{1}{\sigma_\beta^2} + \frac{1}{\omega^{2,(s-1)}} \sum_{j \in \{i:c(i)=c\}} ((T_j \wedge \tau_j^{(s)}) - 1) \tag{40}$$

$$B = \frac{\mu_\beta}{\sigma_\beta^2} + \frac{1}{\omega^{2,(s-1)}} \sum_{j \in \{i:c(i)=c\}} \sum_{t=2}^{T_j \wedge \tau_j^{(s)}} (x_{j,t}^{(s)} - x_{j,t-1}^{(s)}) \tag{41}$$

$$x_0^{(s)}|\mathbf{x}^{(s)}, \mathbf{y}, \omega^{2,(s-1)} \sim \mathcal{N}(D \cdot C^{-1}, C^{-1}) \tag{42}$$

$$C = \frac{1}{\sigma_0^2} + \frac{N}{\omega^{2,(s-1)}} \tag{43}$$

$$D = \frac{\mu_0}{\sigma_0^2} + \frac{1}{\omega^{2,(s-1)}} \sum_{j=1}^{N} x_{j,1}^{(s)} \tag{44}$$

$$T_j \wedge \tau_j^{(s)} = \min(T_j, \tau_j^{(s)}) \tag{45}$$

The term $T_j \wedge \tau_j^{(s)}$ signifies the bound crossing time (the "effective" length of the latent trajectory $x_{j,1:T_j}^{(s)}$). The spike rate is no longer dependent on the diffusion process once the process has crossed the bound (for $t \geq \tau_j^{(s)}$, $y_{j,t}$ is independent of the value of $x_{j,t}^{(s)}$). We therefore only need to consider the values of $x_{j,t}$ for $t \leq \tau_j$ in order to sample from the posterior (replacing $T_j \wedge \tau_j^{(s)}$ with T_j results in a correct, but slower, sampler).

The next step is to sample ω^2 given the newly generated samples of $\beta_{1:C}$ and x_0.

$$\omega^{2,(s)}|\beta_{1:C}^{(s)}, x_0^{(s)}, \mathbf{x}^{(s)}, \mathbf{y} \sim \text{Inv-Gamma}(E, F) \tag{46}$$

$$E = \alpha_\omega + \frac{1}{2} \sum_{j=1}^{N} (T_j \wedge \tau_j^{(s)}) \tag{47}$$

$$\begin{aligned} F = \beta_\omega + \frac{1}{2} \sum_{j=1}^{N} \Bigg[& \left(x_{j,1}^{(s)} - x_0^{(s)} \right)^2 \\ & + \sum_{t=2}^{T_j \wedge \tau_j} \left(x_{j,t}^{(s)} - (x_{j,t-1}^{(s)} + \beta_{c(j)}^{(s)}) \right)^2 \Bigg] \end{aligned} \tag{48}$$

β_c depends only on trials of stimulus coherence c, while x_0 and ω^2 are coupled to all trials.

Even though the bound height parameter, γ, is independent of all other parameters given the latent

states, we cannot sample γ with a closed-form Gibbs step. Instead, we generate samples via a Metropolis-Hastings (MH) step. The MH algorithm samples $\gamma^{(s)}$ with the following steps

1 Sample $\gamma^* \sim q(\gamma|\gamma^{(s-1)})$ where q is an arbitrary proposal distribution.

2 Sample $u \sim U([0,1])$.

3 $\gamma^{(s)} = \begin{cases} \gamma^* & , \quad u < \alpha \\ \gamma^{(s-1)} & , \quad \text{otherwise} \end{cases}$ where $\alpha = \min\left(1, \frac{p(\gamma^*|\mathbf{y},\mathbf{x}^{(s)})q(\gamma^{(s-1)}|\gamma^*)}{p(\gamma^{(s-1)}|\mathbf{y},\mathbf{x}^{(s)})q(\gamma^*|\gamma^{(s-1)})}\right)$

We use a Langevin step for the proposal distribution (*26*)

$$q(\gamma^*|\gamma^{(s-1)}, \mathbf{y}, \mathbf{x}^{(s)}) = \mathcal{N}\left(\gamma^*; \gamma^{(s-1)} + \epsilon^2 \frac{1}{2}\mathbf{G}^{-1}(\gamma^{(s-1)})\frac{d}{d\gamma}\mathcal{L}(\gamma^{(s-1)}), \epsilon^2\mathbf{G}^{-1}(\gamma^{(s-1)})\right) \tag{49}$$

$$\begin{aligned} \mathcal{L}(\gamma) &= \log p(\mathbf{y}|\mathbf{x}^{(s)}, \gamma) + \log p(\gamma) \\ &= \sum_{j=1}^{N}\sum_{t=1}^{T_j} \log p(y_{j,t}|x_{j,t}^{(s)}, \gamma) + \log p(\gamma) \end{aligned} \tag{50}$$

$$\mathbf{G}(\gamma) = -\mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[\frac{d^2}{d\gamma^2}\mathcal{L}(\gamma)\right] \tag{51}$$

Our proposal uses the Fisher information plus the Hessian of the log prior, $\mathbf{G}(\gamma)$, to condition the step, as suggested by (*27*). The result of this conditioning made selecting an effective value for ϵ simple. We set ϵ to a small initial value (0.1) and slowly raised it during the burn-in period to a larger value of 1.

Labeling the soft-rectifying function

$$h(x, \gamma) = \log(1 + \exp(x\gamma)) \tag{52}$$

the derivative of the log likelihood is

$$\frac{d}{d\gamma}\mathcal{L}(\gamma) = \frac{d}{d\gamma}\left[\log p(\mathbf{y}|\mathbf{x}^{(s)}, \gamma) + \log p(\gamma)\right] \tag{53}$$

$$\begin{aligned} &= \frac{d}{d\gamma}\left[\sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(-h(x_{j,t}^{(s)}, \gamma)\Delta_t + y_{j,t}\log(h(x_{j,t}^{(s)}, \gamma))\right)\right. \\ &\qquad \left. +(\alpha_\gamma - 1)\log(\gamma) - \gamma\beta_\gamma + const\right] \end{aligned} \tag{54}$$

$$= \sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(h'(x_{j,t}^{(s)}, \gamma)\left(-\Delta_t + y_{j,t}\frac{1}{h(x_{j,t}^{(s)}, \gamma)}\right)\right) + \frac{\alpha_\gamma - 1}{\gamma} - \beta_\gamma \tag{55}$$

where

$$h'(x, \gamma) = \frac{d}{d\gamma}h(x, \gamma) = \frac{1}{1 + \exp(-x\gamma)}x \tag{56}$$

The Fisher information combined with the prior Hessian can be calculated as

$$\mathbf{G}(\gamma) = -\mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[\frac{d}{d\gamma}\left(\sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(h'(x_{j,t}^{(s)}, \gamma^{(s-1)})\left(-\Delta_t + y_{j,t}\frac{1}{h(x_{j,t}^{(s)}, \gamma)}\right)\right) + \frac{\alpha_\gamma - 1}{\gamma} - \beta_\gamma\right)\right] \tag{57}$$

$$= -\mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[\sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(h''(x_{j,t}^{(s)},\gamma)\left(-\Delta_t + y_{j,t}\frac{1}{h(x_{j,t}^{(s)},\gamma)}\right) - y_{j,t}\left(\frac{h'(x_{j,t}^{(s)},\gamma)}{h(x_{j,t}^{(s)},\gamma)}\right)^2\right) - \frac{\alpha_\gamma - 1}{(\gamma)^2}\right] \tag{58}$$

$$= -\sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(h''(x_{j,t}^{(s)},\gamma)\left(-\Delta_t + \mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[y_{j,t}\right]\frac{1}{h(x_{j,t}^{(s)},\gamma)}\right) - \mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[y_{j,t}\right]\left(\frac{h'(x_{j,t}^{(s)},\gamma)}{h(x_{j,t}^{(s)},\gamma)}\right)^2\right) + \frac{\alpha_\gamma - 1}{(\gamma)^2} \tag{59}$$

Noting that $\mathbb{E}_{\mathbf{y}|\gamma,\mathbf{x}^{(s)}}\left[y_{j,t}\right] = h(x_{j,t}^{(s)},\gamma)\Delta_t$

$$\mathbf{G}(\gamma) = \frac{\alpha_\gamma - 1}{(\gamma^2)} + \sum_{j=1}^{N}\sum_{t=1}^{T_j}\left(\Delta_t\frac{\left(h'(x_{j,t}^{(s)},\gamma)\right)^2}{h(x_{j,t}^{(s)},\gamma)}\right) \tag{60}$$

2.2 Discrete stepping model

The stepping model allows the firing rate to occupy three discrete states: an initial state and two "decision" states, one for each of the two possible choices in the task. Each state is associated with constant firing rate ($\alpha_{0:2}$), and these firing rates are constant across all trials (*28*). We define state transitions (steps) as instantaneous events for simplicity. We only allow a single transition between states during a trial (i.e., "up" or "down"), although the model does not force a step to occur on every trial.

Let z_j denote the step time (in discrete bins) for trial j and d_j denote the state stepped to. If z_j is greater than the trial length, then no step occurs during the trial. We refer to the initial state as state 0.

$$z_j \sim \text{Negative Binomial}(p_{c(j)}, r) \tag{61}$$

$$P(d_j = 1) = \phi_{c(j)} \tag{62}$$

$$P(d_j = 2) = 1 - \phi_{c(j)} \tag{63}$$

$$y_{j,t}|t \leq z_j \sim \text{Poisson}(\alpha_0\Delta_t) \tag{64}$$

$$y_{j,t}|t > z_j \sim \text{Poisson}(\alpha_{d_j}\Delta_t) \tag{65}$$

The negative binomial distribution is a discrete-time analogue of a gamma distribution. The distribution can be interpreted as the number of coin flips needed to get r heads,where $1-p$ is the probability of a head (although we allow r to take continuous values). If the negative binomial parameter r is set at 1, then this model becomes a more commonly used Hidden Markov model (HMM), where step times follow an exponential distribution, with added restriction that the model cannot step out of states 1 or 2. We chose not to restrict state transitions to be Markovian so that the stepping model could exhibit similar hit-time distributions as the ramping model. Additionally, we note that trial lengths are finite, but z_j can take arbitrarily large values. Values larger than the trial length are interpreted to mean that no step took place on the trial.

The model parameters we estimate are $\Psi = \{\alpha_{0:2}, r, \phi_{1:C}, p_{1:C}\}$.

2.2.1 Prior distributions: stepping model

The prior distributions for the model parameters take the following form

$$\alpha_0 \sim \text{Gamma}\,(\alpha_\alpha, \beta_\alpha) \tag{66}$$

$$\alpha_1 \sim \text{Gamma}\,(\alpha_\alpha, \beta_\alpha) \tag{67}$$

$$p(\alpha_2|\alpha_1) \propto \mathbf{1}(\alpha_2 > \alpha_1)\text{Gamma}\,(\alpha_\alpha, \beta_\alpha) \tag{68}$$

$$p_c \sim \text{Beta}(\alpha_p, \beta_p) \tag{69}$$

$$\phi_c \sim \text{Beta}(\alpha_\phi, \beta_\phi) \tag{70}$$

$$r \sim \text{Gamma}(\alpha_r, \beta_r) \tag{71}$$

We enforced $\alpha_2 > \alpha_1$ with the truncated gamma prior, because we wanted state 2 to always represent the up state and state 1 the down state (without this restriction, the model is unidentifiable because the state labels can be swapped). Otherwise, the prior distributions are independent.

We chose the following parameters

1. $\alpha_\alpha = 1$, $\beta_\alpha = 0.01$
2. $\alpha_p = 1$, $\beta_p = 1$
3. $\alpha_\phi = 1$, $\beta_\phi = 1$
4. $\alpha_r = 2$, $\beta_r = 1$

The distribution over firing rates is broadly tuned, with both a mean and standard deviation of 100 spikes/s. The beta distributions over ϕ and p are uniform over the range of values $[0, 1]$. We chose the prior on r to be peaked at 1, where the model becomes a HMM.

In the MCMC section, we use the prior parameter symbols instead of these specific values so that it is clear where the priors are placed.

2.2.2 MCMC: stepping model

As with the ramping model, the sampler consists of two main steps: (1) sampling the latent stepping process given the parameters and (2) sampling the parameters given the new latent states. We initialized the chain by setting the rate parameters based on average spike rates: α_0 was set to the average spike rate over all trials in the first bin, α_2 was set to the average spike rate in the final bin of all in-RF choice trials, and α_1 was set to $\frac{1}{2}\alpha_2$. The initial values for all the ϕ and p parameters was 0.5 (the mean of the prior). The final parameter, r, was initially set to the mode of the prior distribution.

Sampling the step times

Our goal is to obtain the sth sample given the $(s-1)$th sample of the model parameters and the spike times:

$$(\mathbf{z}, \mathbf{d})^{(s)} \sim P(\mathbf{z}, \mathbf{d}|\Psi^{(s-1)}, \mathbf{y}). \tag{72}$$

With the model parameters fixed, the trials are conditionally independent. Unfortunately, there is unfortunately no simple analytic form for this distribution. However, in this simple stepping case, we can sample the step times numerically by computing the distribution on a finite grid of time points. We truncated the grid at 1500 time bins, which is 15 times longer than the longest trial and much longer than we would ever expect to set a step time. Therefore, the truncated distribution is an extremely close approximation to the true distribution. We note that z_j can be greater than T_j, which is interpreted to mean no step occurred during the trial, and this does not affect the later sampling steps. For $z_j = 1$ to 1500 and $d_j = 1$ to 2, we calculate

$$\begin{aligned} p(z_j, d_j | y_{j,1:T_j}, \Psi^{(s-1)}) &\propto p(y_{j,1:T_j} | z_j, d_j, \alpha^{(s-1)}) p(z_j | r^{(s-1)}, p_{c(j)}^{(s-1)}) p(d_j | \phi_{c(j)}^{(s-1)}) \\ &= p(z_j | r^{(s-1)}, p_{c(j)}^{(s-1)}) p(d_j | \phi_{c(j)}^{(s-1)}) \prod_{t=1}^{T_j} p(y_{j,t} | z_j, d_j, \alpha^{(s-1)}) \end{aligned} \tag{73}$$

The distributions on the right side are calculated using the model definition for the Poisson observation $p(y_{j,1:t}|z_j, d_j, \alpha^{(s-1)})$, the negative binomially distributed step time $p(z_j|r^{(s-1)}, s^{(s-1)})$, and the Bernoulli state choice $p(d_j|\phi^{(s-1)})$. Once all values are calculated, the joint distribution can be normalized and sampled directly.

Sampling the stepping parameters

With the sth sample of the latent states, we sample a new setting for the model parameters. This step is broken into two parts: first we sample $\alpha_{0:2}$, p, and ϕ. Then with the new values of $p_{1:C}$ and z, we sample the final parameter: r.

$\alpha_{0:2}$, p, and ϕ are all sampled independently with Gibbs' steps.

$$\begin{aligned} \phi_c^{(s)} | \mathbf{d}^{(s)} &\sim \text{Beta}\left(\alpha_\phi + D_c, \beta_\phi + N - D_c\right) \\ D_c &= \sum_{j \in \{i : c(i) = c\}} \mathbf{1}(d_j^{(s)} = 2) \end{aligned} \tag{74}$$

$$p_c^{(s)} | \mathbf{z}^{(s)}, r^{(s-1)} \sim \text{Beta}\left(\alpha_p + \sum_{j \in \{i:c(i)=c\}} z_j^{(s)}, \beta_p + r^{(s-1)} \sum_{j \in \{i:c(i)=c\}} 1\right) \tag{75}$$

$$\alpha_0^{(s)} | \mathbf{z}^{(s)}, \mathbf{d}^{(s)}, \mathbf{y} \sim \text{Gamma}\left(\alpha_\alpha + \sum_{j=1}^{N} \sum_{t=1}^{z_j^{(s)}} y_{j,t}, \beta_\alpha + \sum_{j=1}^{N} z_j^{(s)}\right) \tag{76}$$

$$\alpha_1^{(s)} | \mathbf{z}^{(s)}, \mathbf{d}^{(s)}, \mathbf{y} \sim \text{Gamma}\left(\alpha_1, \beta_1\right) \tag{77}$$

$$\alpha_2^{(s)} | \alpha_1^{(s)} \mathbf{z}^{(s)}, \mathbf{d}^{(s)}, \mathbf{y} \sim \mathbf{1}(\alpha_2^{(s)} > \alpha_1^{(s)}) \text{Gamma}\left(\alpha_2, \beta_2\right) \tag{78}$$

$$\alpha_i = \alpha_\alpha + \sum_{j=1}^{N} \sum_{t=z_j^{(s)}+1}^{T_j} \mathbf{1}(d_j^{(s)} = i) y_{j,t} \tag{79}$$

$$\beta_i = \beta_\alpha + \sum_{j=1}^{N} \mathbf{1}(d_j^{(s)} = i)(T_j - z_j^{(s)}) \tag{80}$$

The truncated gamma distribution on $\alpha_2^{(s)}$ enforces that $\alpha_2^{(s)} > \alpha_1^{(s)}$ holds (see model prior section).

No simple closed form distribution exists for the posterior over the negative binomial parameter r. We obtained samples using a Metropolis adjusted Langevin algorithm that uses the following proposals (see the ramping sampler for a brief description of Metropolis-Hastings proposals):

$$q(r^*|r^{(s-1)}, \mathbf{y}, \mathbf{z}^{(s)}, p^{(s)}_{1:C}) = \mathcal{N}\left(r^*; r^{(s-1)} + \epsilon^2 \frac{1}{2}\frac{d}{dr}\mathcal{L}(r^{(s-1)}), \epsilon^2\right) \tag{81}$$

$$\begin{aligned}\mathcal{L}(r^{(s-1)}) &= \log p(\mathbf{z}^{(s)}|r^{(s-1)}, p^{(s)}_{1:C}) + \log p(r^{(s-1)}) \\ &= \sum_{j=1}^{N} \log p(z^{(s)}_j|r^{(s-1)}, p^{(s)}_{c(j)}) + \log p(r^{(s-1)})\end{aligned} \tag{82}$$

We applied a simple automatic procedure to select a value for ϵ. We initialized ϵ to a small value (0.05), and during the burn-in period, we would raise or lower (by multiplicative factors of 1.25 or 0.75) until the sampler stabilized to an acceptance rate in the range of $30 - 70\%$.

The derivative can be calculated

$$\mathcal{L}(r^{(s-1)}) = \sum_{j=1}^{N} \log p(z^{(s)}_j|r^{(s-1)}, p^{(s)}_{c(j)}) + \log p(r) \tag{83}$$

$$\begin{aligned}&= \sum_{j=1}^{N} \left[\log\Gamma(z^{(s)}_j + r^{(s-1)}) - \log\Gamma(z^{(s)}_j + 1) - \log\Gamma(r^{(s-1)}) - z^{(s)}_j \log(p^{(s)}_{c(j)}) + r^{(s-1)}\log(1 - p^{(s)}_{c(j)})\right] \\ &\qquad + (\alpha_r - 1)\log(r^{(s-1)}) - r^{(s-1)}\beta_r + const\end{aligned} \tag{84}$$

$$\frac{d}{dr}\mathcal{L}(r^{(s-1)}) = \sum_{j=1}^{N} \left[\psi(z^{(s)}_j + r^{(s-1)}) - \psi(r^{(s-1)}) + \log(1 - p^{(s)}_{c(j)})\right] + \frac{\alpha_r - 1}{r^{(s-1)}} - \beta_r \tag{85}$$

where Γ and ψ are the gamma and digamma functions respectively. We point out that this sampling step uses the entire set of coherence-dependent p parameters.

2.3 Model comparison

We compared model fits to the data using the Deviance Information Criterion (DIC) (*16*). The metric is defined as

$$DIC = 2\log p(Data|\bar{\Theta}, \mathcal{M}_r) - 4\mathbb{E}_{\Theta|Data,\mathcal{M}_r}\left[\log p(Data|\Theta, \mathcal{M}_r)\right]. \tag{86}$$

where $\bar{\Theta}$ is the posterior mean of the parameters given the data, and $\mathcal{M}_r$ is ramping model. (For the stepping model $\mathcal{M}_s$, replace Θ with Ψ). The expectation term can be estimated using the samples from the MCMC ($\Theta^{(1)}$ to $\Theta^{(S)}$)

$$\mathbb{E}_{\Theta|\mathbf{y},\mathcal{M}_r}\left[\log p(Data|\Theta, \mathcal{M}_r)\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(\mathbf{y}|\Theta^{(s)}, \mathcal{M}_r) \tag{87}$$

Estimating $\log p(\mathbf{y}|\Theta^{(s)}, \mathcal{M}_r)$ for each sample is computationally expensive. For the ramping model, we estimated this value by sampling 300 trajectories for each trial from the distribution $p(\mathbf{x}|\Theta^{(s)})$. For

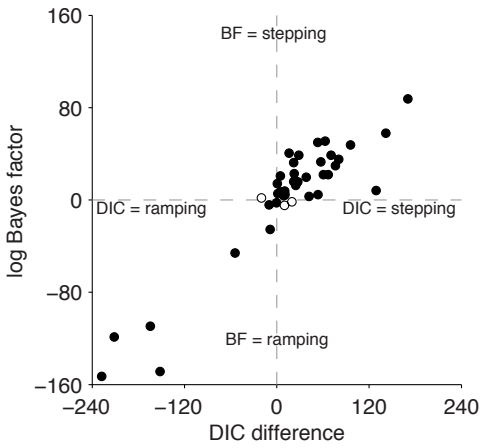

Fig. S1: We compare estimated DIC differences with the log Bayes factor model comparison metric for all 40 cells. These values are highly correlated ($r = 92.7, p < 10^{-8}$) and they provide the same model comparison result for all but 3 cells (open circles), all of which had weak (< 10) model assignments from the log Bayes factor.

the stepping model, we estimated this quantity by truncating the possible step times to a maximum of 1500 and then calculating the probability on a discrete grid over time points.

Higher DIC is evidence against a model. DIC has both a log-likelihood term of the data at the posterior mean (the basic goodness-of-fit term), and a penalty term which integrates over the posterior distribution of the parameters. This term is designed to act as an estimate of the effective number of parameters in the model. In fact, this term converges exactly to the number of parameters in a Gaussian model. More well-known metrics, like the Bayesian information criterion (BIC) are more difficult to apply to latent variable models: how does one include the latent variables in the count of model parameters? The latent spaces in the two models are starkly different: the stepping model's latent state can be described by two values (step time, and state stepped-to), while the state in the ramping model is much more complex. DIC avoids this issue entirely with its estimate of the effective number of parameters. Although DIC has not been widely applied in neuroscience, it has become a widely used model fitness criterion within many other domains (*29*).

One example alternative to DIC is the Bayes factor, which compares marginal likelihoods of the data given the proposed models:

$$BF = \frac{P(Data|\mathcal{M}_r)}{P(Data|\mathcal{M}_s)} \tag{88}$$

This value also integrates over model fit uncertainty, but it also suffers from strong dependency on the choice of prior distribution. DIC has the advantage of working with improper, uninformative priors. However, for our data, we achieved similar results with both DIC and Bayes factors (Figure S1). The marginal likelihoods can be calculated from the output of the MCMC algorithm (*30, 31*). We confirmed that DIC is an accurate metric for model comparison using simulated data (see Section 3.1).

2.4 Spike rate and variances

For each cell, we computed the coherence-sorted spike rates (PSTH) and variances (PSTV) shown in Figs. 3A (left) and 4 by calculating the spike count mean and variance in a 25 ms sliding window (boxcar filter), moved by 5 ms increments. For each trial, we performed this convolution only over the spikes within the analysis window for that trial. We computed population rates and variances by averaging the rates and variances from all 40 cells. Spikes occurring later than 200 ms after motion offset were not included in the PSTH. We multiplied the mean and variance by $frac10.025$ to transform the PSTH into units of spikes per second.

We compared the model predictions of the coherence sorted PSTH/PSTV to the true population spike rate and variance during the interval 205-700 ms after motion onset (100 time points; Fig. 4A). We simulated 1000 spike trains per coherence level per cell using the posterior-mean parameters from each model to obtain the coherence-sorted model predicted rate (MR) and variance (MV). We then calculated the fraction of variance explained in the coherence-dependent PSTH (or PSTV):

$$R^2 = 1 - \frac{\sum_{c=1}^{5}\sum_{t=205}^{700}\left(MR_{c,t} - PSTH_{c,t}\right)^2}{\sum_{c=1}^{5}\sum_{t=205}^{700}\left(\overline{PSTH} - PSTH_{c,t}\right)^2} \tag{89}$$

$$\overline{PSTH} = \frac{1}{5}\frac{1}{100}\sum_{c=1}^{5}\sum_{t=205}^{700} PSTH_{c,t} \tag{90}$$

where $PSTH_{c,t}$ is the PSTH at time t (in milliseconds) for coherence level c. The average spike rate over all time and coherence levels is $\overline{PSTH}$. The sums over t are in increments of 5 ms. An R^2 of 1 indicates that the model PSTH perfectly matches the data, lower values indicate a worse fit.

Credible intervals on the R^2 values included uncertainty in both the measured PSTH/PSTV as well as the model fit uncertainty. We obtained 1000 samples of the data PSTH (PSTV) by randomly drawing a set of trials with replacement for each cell and computing the population PSTH (PSTV) with those trials. Each sample used the number of trials per coherence as were actually observed for each cell. We obtained errors on the model PSTH (PSTV) by using the output from the MCMC. We simulated each model with each of the $10\,000$ parameter samples from the MCMC output. We calculated the R^2 for each of the $10\,000$ simulated PSTHs (PSTVs) against the 1000 bootstrapped data PSTHs (PSTVs), resulting $10\,000\,000 R^2$ values per model (Fig. S2). The 95% credible interval was the 2.5 and 97.5 percentiles of the R^2 values.

2.5 Model-based step decoding

Figures 2,4, and S13- S15 visualize the LIP responses using decoded steps. We estimated step times and direction using a Bayesian decoder with the model fit parameters ($\bar{\Psi}$ = posterior mean of Ψ). For

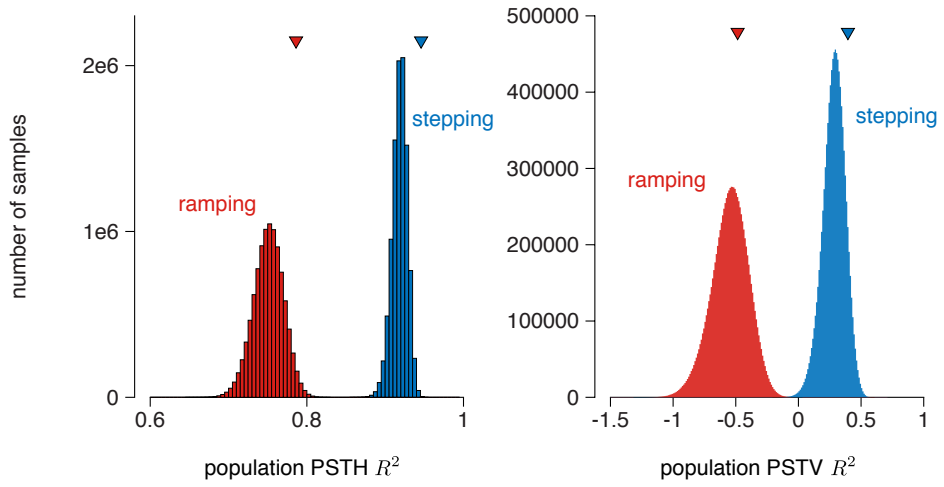

Fig. S2: Distribution of sampled R^2 values of the model predictions of the coherence-sorted PSTH (left) and PSTV (right). The distributions compare the ramping model PSTH/PSTV predictions (red) to the predictions from the stepping model (blue). Triangles indicate the R^2 values calculated for the PSTHs or PSTVs plotted in Fig. 4A, which were computed using the posterior mean parameters.

a trial j, we calculated the distribution over step times, marginalizing over step direction

$$P(z_j = z|y_{j,1:T_j}, \bar{\Psi}) = \sum_{d\in\{1,2\}} P(z_j = z, d_j = d|y_{j,1:T_j}, \bar{\Psi})P(d_j = d|\bar{\Psi}) \tag{91}$$

$$\propto P(y_{j,1:T_j}|z_j = z, d_j = d, \bar{\Psi})P(z_j = z|\bar{\Psi})P(d_j = d|\bar{\Psi}) \tag{92}$$

where the final terms $P(y_{j,1:T_j}|z_j = z, d_j = d, \bar{\Psi})$, $P(z_j = z|\bar{\Psi})$, and $P(d_j = d|\bar{\Psi})$ are all given by the model definition. For computational tractability, we truncated the possible step times at a maximum of 1500 time steps (15 times longer than the longest trial) and normalized the distribution based on the truncation. We then estimated the median step time as

$$\hat{z_j} = \underset{z\in\{0,\ldots,1500\}}{\arg\min} \frac{1}{2} \geq \left(\sum_{x=1}^{z} P(z_j = x|y_{j,1:T_j}, \bar{\Psi})\right) \tag{93}$$

We chose to use the median time instead of the mean because the distribution of step times tended to be highly skewed. A MAP estimator achieved similar results.

If the step time occurred after the trial end, we decoded no step on that trial. Otherwise, step direction was decoded as

$$\hat{d_j} = \underset{d\in\{1,2\}}{\arg\max} P(d_j = d|y_{j,1:T_j}, \bar{\Psi}) \tag{94}$$

$$= \underset{d\in\{1,2\}}{\arg\max} \sum_{z=0}^{1500} P(d_j = d|z_j = z, y_{j,1:T_j}, \bar{\Psi})P(z_j = z|\bar{\Psi}) \tag{95}$$

$$= \arg\max_{d \in \{1,2\}} \sum_{z=0}^{1500} P(y_{j,1:T_j} | d_j = d, z_j = z, \bar{\Psi}) P(d_j = d | \bar{\Psi}) P(z_j = z | \bar{\Psi}) \quad (96)$$

where the distributions $P(y_{j,1:T_j} | d_j = d, z_j = z, \bar{\Psi})$, $P(d_j = d | \bar{\Psi})$, and $P(z_j = z | \bar{\Psi})$ are all given by the model definition (product of independent Poissons, a Bernoulli distribution, and a negative binomial distribution respectively).

2.5.1 Step-aligned figures (Figs. 2,3,4, S13- S15)

For all step-aligned figures, spike rates on each trial were estimated by smoothing the spike trains with a centered boxcar filter (25 ms wide). The rates were then aligned to the step time on each trial and averaged. The step-aligned average included only the segment beginning 200 ms after motion onset until 200 ms after motion offset. This allowed our analyses to focus on the decision formation stage, separated in time from other events like the visual onset of the targets or the saccade. For trials in which we decoded no step (see Section 2.5), we aligned the trial to the end of our analysis period (before the go signal).

2.5.2 Model-based choice probability (Fig. 4)

Choice probability (CP) is a widely-used metric to quantify the relationship between spike count fluctuations and behavior. For a fixed stimulus, higher spike counts on single trials are often correlated with in-RF choices. CP is the probability that a spike count observed during an in-RF choice trial is greater than a spike count observed on an out-RF choice trial in response to the same stimulus. CP does not consider the percentage of in-RF or out-RF choices, only the distributions of spike counts observed for in-RF choice and out-RF choice trials. This way, we obtain a measure how the trial-to-trial fluctuations of a neuron correlate with choice, which discards any overall bias towards one choice.

CP is equivalent to decoding choice using a neuron-antineuron pair. The “anti-neuron” is a hypothetical neuron whose response distributions are equal to the recorded neuron, but with tuning to the targets reversed. A decoder would chose a target by selecting which of these two neurons gives the highest spike count on a single trial. An out-RF choice occurs when the antineuron gives the larger count, and the in-RF choice occurs when the antineuron has the lower count.

We calculated CP conditioned on stimulus coherence and direction and took the final CP as the average across conditions. We calculated CPs using dot coherences from $0 - 12.8\%$. We included only coherence levels with at least 3 in-RF choices and 3 out-RF choices into the final CPs. We chose to average CPs across conditions instead of $z-$scoring each response within the stimulus categories and calculating a single CP on the pooled responses, because $z-$scoring can be biased by conditions with unbalanced choice selection (*32*).

In addition to the classical, spike count-based CP, we calculated a model-based CP. We used the model fit to calculate the probability that the latent state had stepped up given a spike train. The model-based CP is the probability that an in-RF choice trial is more likely to have stepped into the up state than an out-RF choice trial (i.e., we calculated model-based CP by replacing the spike count observations in classical CP with the step probabilities).

The specific number we compare for trial j at time t is computed

$$P_{up}(j) = \frac{P(d_j = 1 | y_{j,1:t}, z_j < t, \bar{\Psi})}{P(d_j = 1 | y_{j,1:t}, z_j < t, \bar{\Psi}) + P(d_j = 2 | y_{j,1:t}, z_j < t, \bar{\Psi})} \tag{97}$$

where

$$P(d_j = 1 | y_{j,1:t}, z_j < t, \bar{\Psi}) = \sum_{z=0}^{t-1} P(z_j = z, d_j = 1 | y_{j,1:t}, z_j < t, \bar{\Psi}) \tag{98}$$

$$\propto \sum_{z=0}^{t-1} P(y_{j,1:t} | z_j = z, d_j = 1, \bar{\Psi}) P(z_j = z | \bar{\Psi}) P(d_j = 1 | \bar{\Psi}) \tag{99}$$

Here, the decoder assumes that a step has been made (i.e., it returns a "best guess" about the choice).

3 Supplementary Information

3.1 MCMC Results

To assess the ability of our MCMC methods to successfully fit and identify models, we simulated trials from the ramping and stepping models and applied the MCMC methods to the simulated data. We examined the ability of the sampler to recover true parameters and the DIC metric to identify the correct model.

3.1.1 Simulations: evaluating sampler effectiveness

First, to evaluate how well the MCMC algorithms were mixing (i.e., how quickly the chain could start to produce effectively independent samples from the true posterior), we calculated the autocorrelations in the parameter samples from the chain. This is a basic visual tool that can indicate if the chain has reached the true posterior, and how independent the samples are (with lower autocorrelation meaning more independence). We first tested this metric using all $50\,000$ samples from the MCMC output (Fig. S3, top). This indicated that the ramping parameters ω^2 and γ, and the stepping model parameter r had a high autocorrelation. However, once we thinned the chain by taking only every 5th sample (described in section 2), the autocorrelation was largely eliminated (Fig. S3, bottom). The autocorrelation from the MCMC output of the fit to the cell shown in Fig. 2 is shown in Fig. S4.

Additionally, we confirmed our ability to estimate the true parameters from the simulations (i.e., we asked, is the estimator consistent?). We ran the MCMC on various amounts of simulated data and compared the posterior mean estimates of the parameters to the true parameters. Figure S5 shows that the error is reduced with more data.

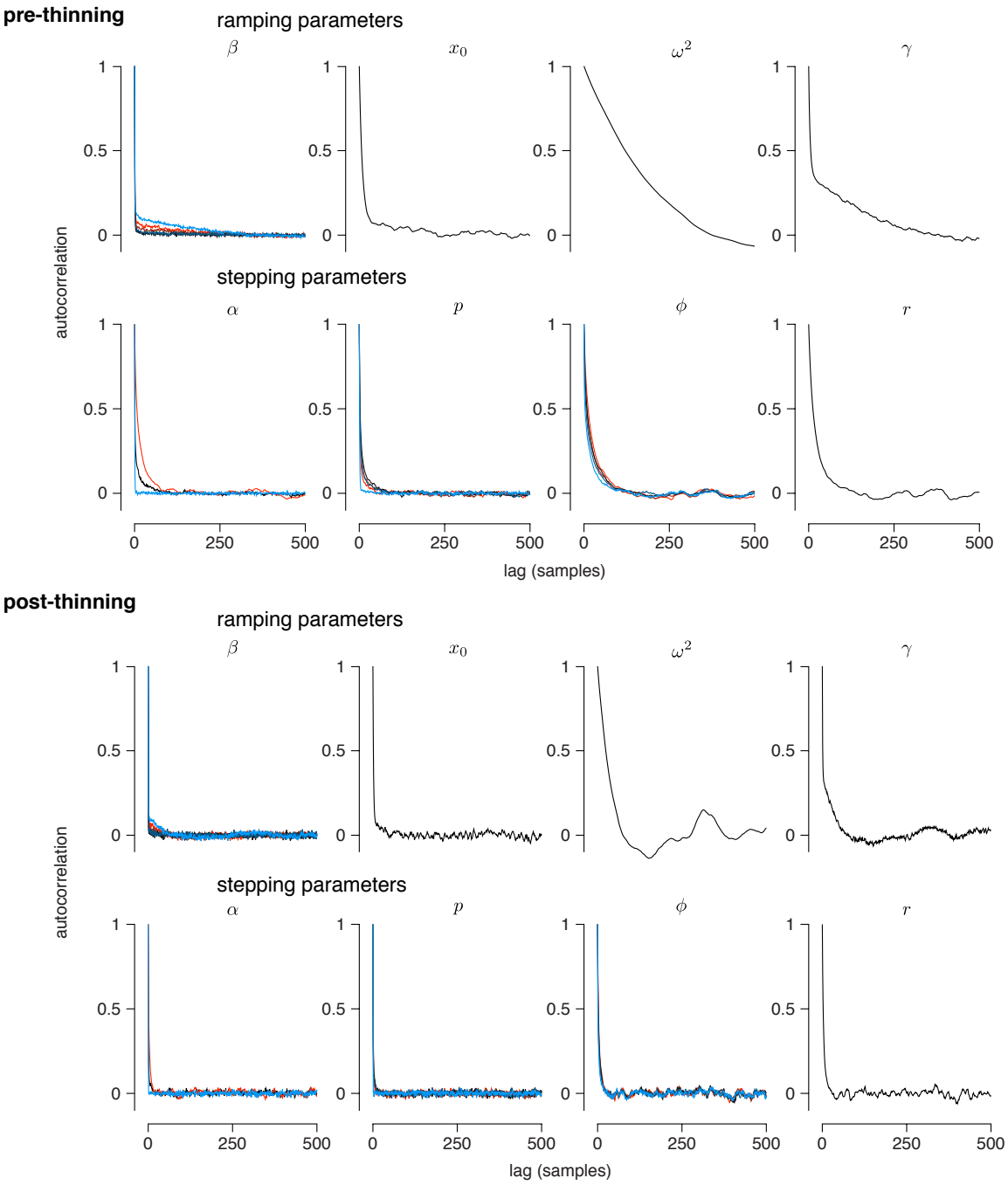

Fig. S3: Autocorrelation plots of all the parameter samples from our MCMC for all the parameters for both models for simulated data. The abscissa is in units of iterations of the MCMC algorithm. The top two rows show the autocorrelation of the samples from the original MCMC output. The bottom two rows show the autocorrelation after we thinned the chain, taking only every 5th sample. The chain was run on datasets containing 500 trials, with 100 trials of each of the 5 possible motion coherence levels.

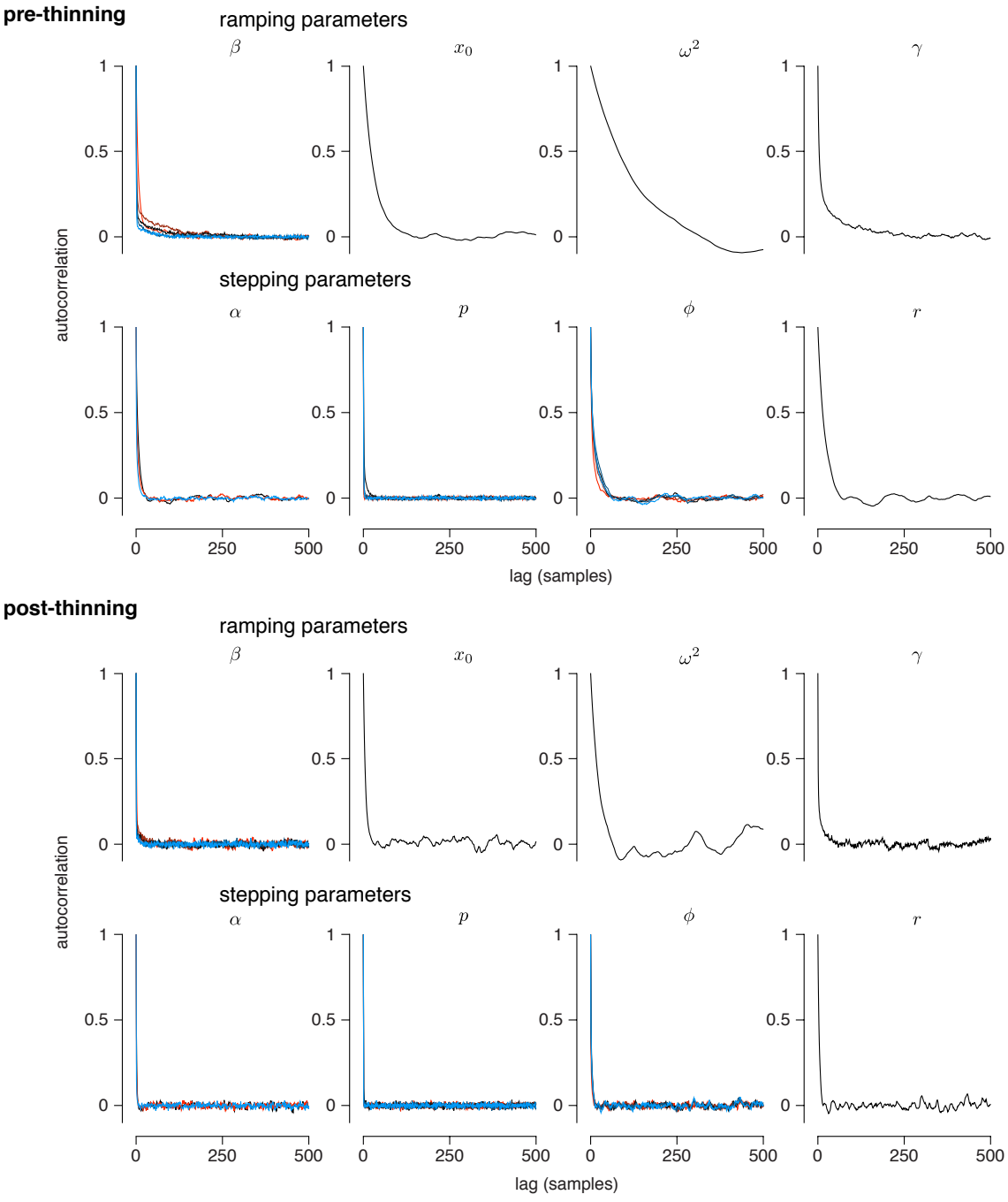

Fig. S4: Autocorrelation plots of all the parameter samples from our MCMC for all the parameters for both models for the cell shown in Fig. 2. The result is comparable to the simulation in Fig. S3

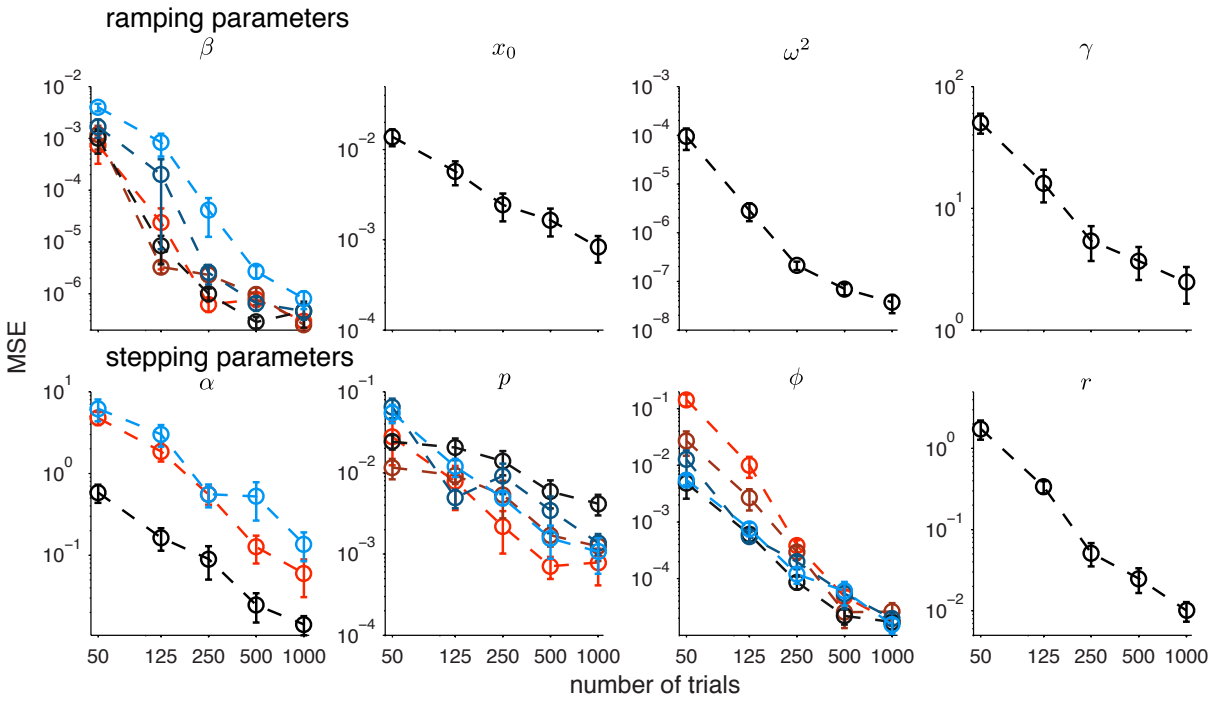

Fig. S5: Mean-squared errors of the MCMC posterior mean estimate (log scaled) of the model parameters as a function of the amount of simulated data. Simulated datasets containing an equal number of trials of each of 5 possible motion coherence levels. For the 50 and 125 samples populations, 20 independent simulations and MCMC runs were used. 10 runs were used for the remaining sample sizes. Error bars show one standard error of the estimate of the MSE. The true parameters were chosen to be similar to the parameters estimated for a real LIP neuron: $\beta = \{-4.7e\times10^{-3}, -2.4\times10^{-3}, -1.3\times10^{-3}, 6\times10^{-4}, 3.4\times 10^{-3}\}, x_0 = 0.72, \omega^2 = 1.7\times10^{-3}, \gamma = 39.7, \alpha = \{4.1, 0.57, 41.0\}, \phi = \{0.10, 0.30, 0.71, 0.82, 0.98\}, p = \{0.990.98, 0.98, 0.975, 0.97\}, r = 1.05$

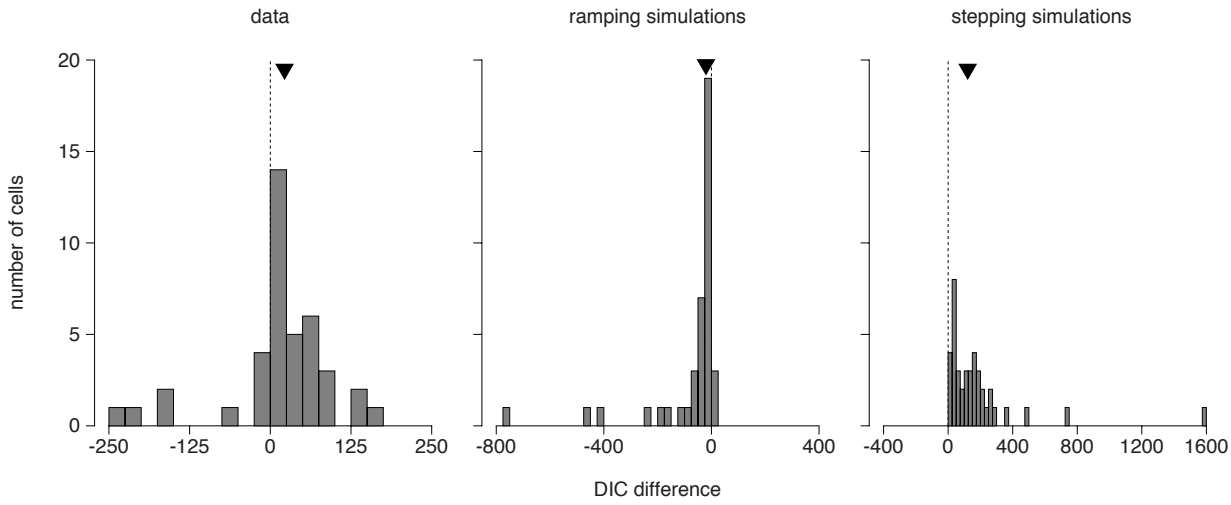

Fig. S6: We confirmed our model comparison techniques by performing the model comparison procedure on simulated data generated from the model fits to all 40 cells. Each simulation contained the same number of trials as in the actual data (on an individual cell basis). 25 out of 40 of the real cells (left) show strong evidence of stepping. For all 40 simulations of the stepping model (right), the model comparison showed strong evidence (DIC difference greater than 10) towards the correct model. For the ramping simulations (center), 31 simulations showed strong evidence supporting the ramping model. Our model comparison showed that 3 ramping simulations could be explained better by stepping, but none of these simulations offered strong support for stepping (maximum DIC difference of 3, well below our threshold of 10 for strong support). Median DIC differences are given by the triangles: data=22.1, ramping simulations=-20.3, and stepping simulations=121.2.

3.1.2 Simulations: model comparison

Figure S6 demonstrates that the model comparison works on simulated data. We simulated responses using parameters found in the model fits to actual data. The model comparison is able to consistently identify the simulated models, although many cells show a small DIC difference in these parameter regimes.

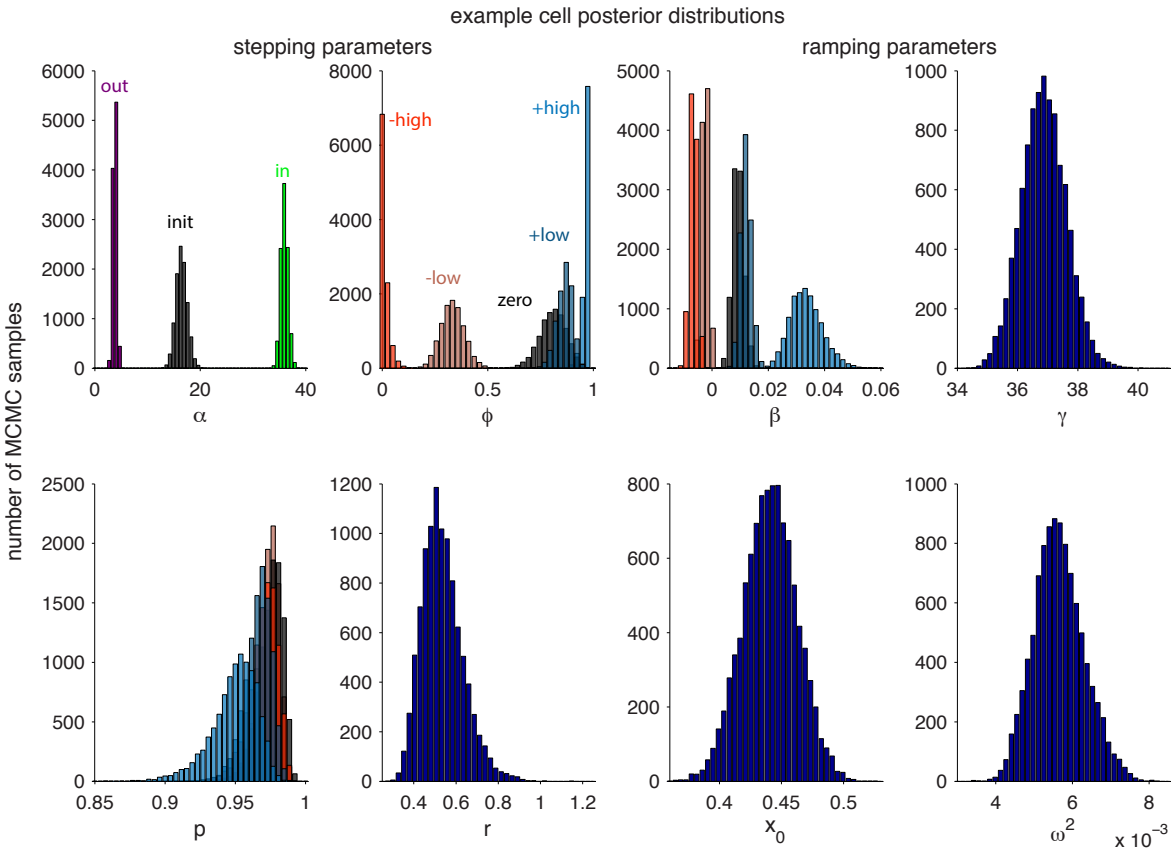

Fig. S7: Example estimates of the posterior distributions for all the parameters for both models (stepping model left 2 columns, ramping model right 2 columns) for a single example LIP cell. For the ϕ, p, and β parameters, distributions for all 5 coherences are shown.

3.1.3 Data: parameter estimates

The posterior distributions for the model parameters estimated from the MCMC for one example LIP cell are provided in figure S7. These estimated posteriors are simply histograms of the samples from the Markov chain. For visualization purposes, these distributions are the marginal posterior distributions for each parameter given the set of spike trains. However, the samples come from the joint posterior. Figure S8 shows samples from the posterior distribution over the latent firing rates for the stepping model for 15 example trials from one cell. Figure S9 shows samples from the posterior distribution over the firing rates for the ramping model for the same trials.

As discussed in the methods, we used the posterior mean as a specific estimate for the model parameters for a cell. Figures S10-S11 plot the population summary of the parameters estimates for both models, and the exact values are provided in Tables S1-S2 along with the model comparison results for each cell.

In addition to the parameter estimates, we also estimated the step times for each trial. Figure S12A shows the step times of 3 example cells. Figure S12C shows the average and variance of step times for each coherence, which was calculated by first taking the average (or standard deviation) of step times for each cell individually, then averaging across the population.

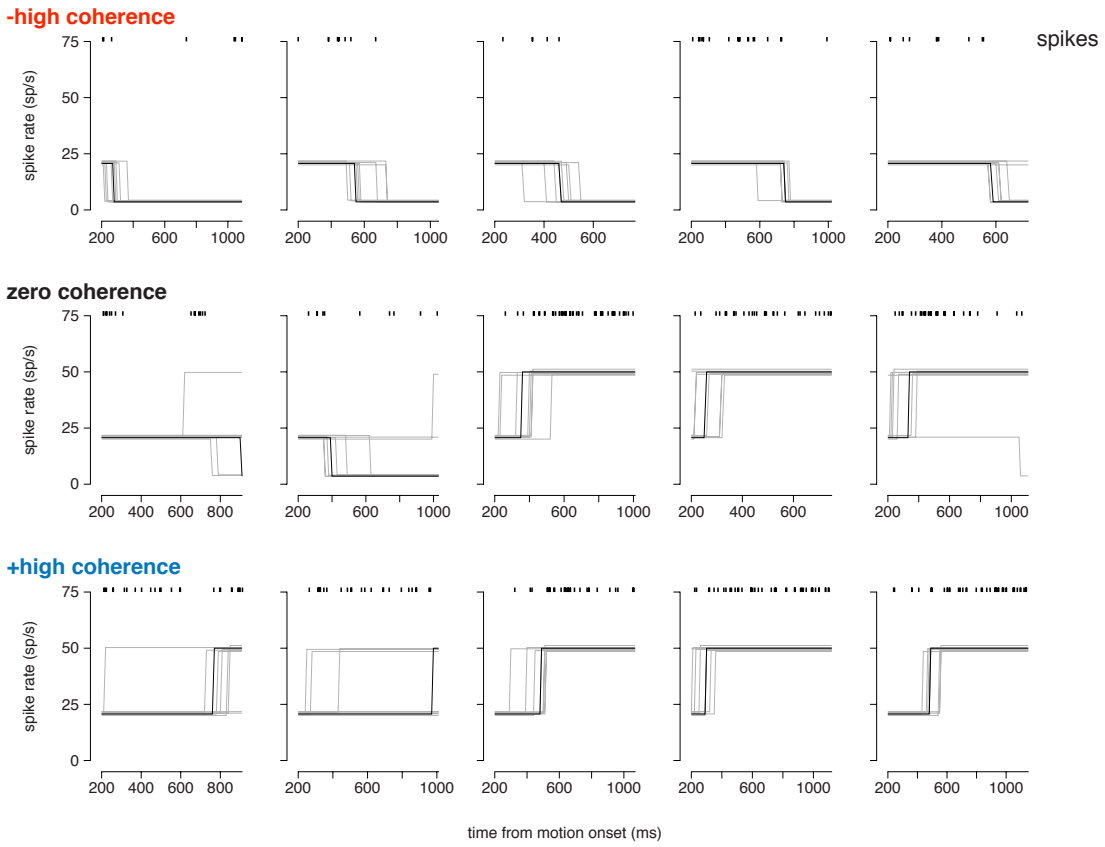

Fig. S8: Each plot shows samples from the posterior distribution over the latent firing rates in the stepping model (gray traces) for five example trials at each of three motion coherence levels. The black trace shows the trial's decoded step that was used for the step-aligned plots (Section 2.5). Rasters above each plot indicate the spike times on each trial. (Top row) High negative coherence trials. (Middle row) 0% coherence trials. (Bottom row) high positive coherence trials. The trials shown are from cell 10.

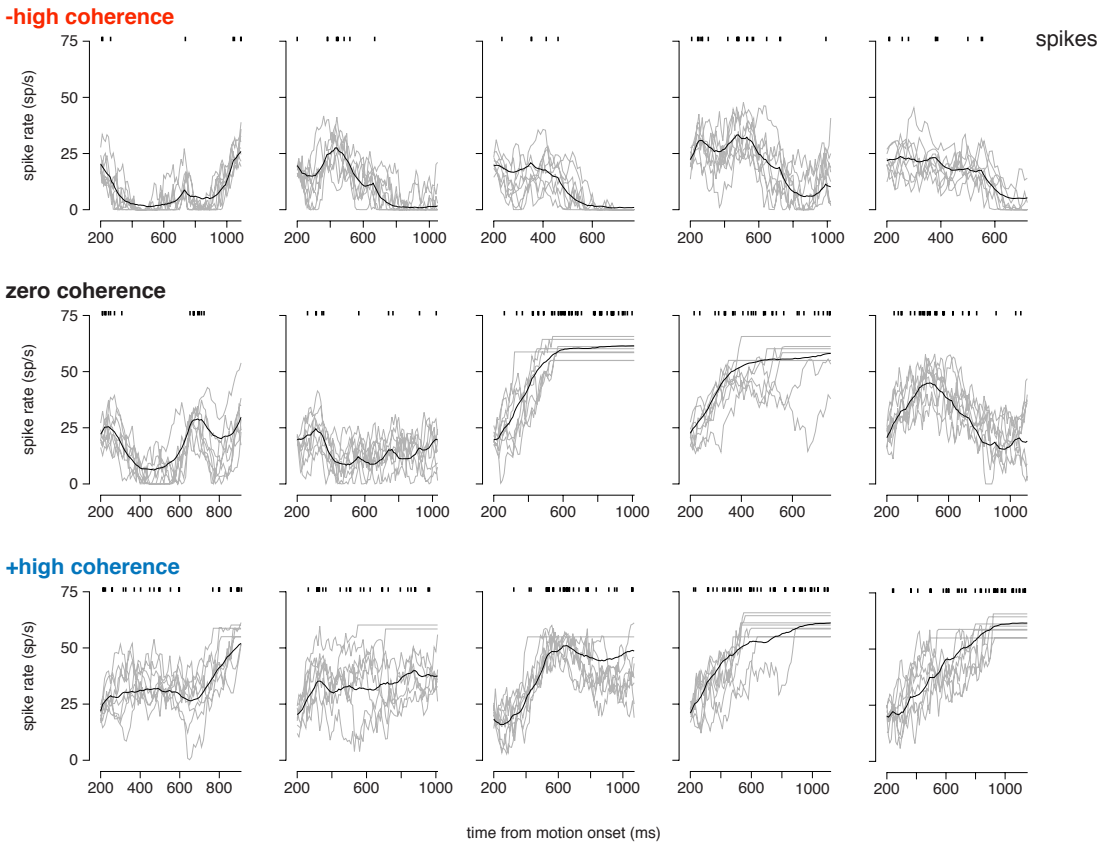

Fig. S9: Each plot shows samples from the posterior distribution over the latent firing rates in the ramping model (gray traces) for the same trials as shown in Fig. S8. The black trace shows the trial's posterior mean firing rate. Rasters above each plot indicate the spike times on each trial. (Top row) High negative coherence trials. (Middle row) 0% coherence trials. (Bottom row) high positive coherence trials.

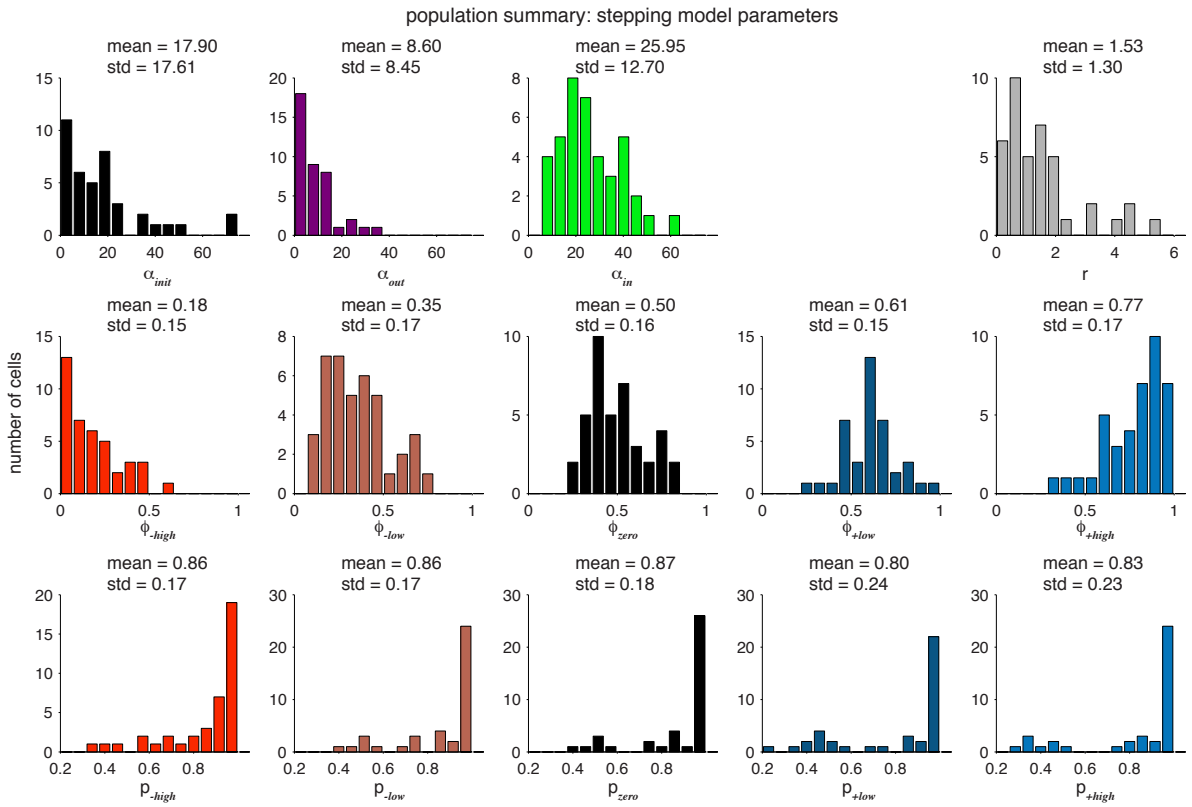

Fig. S10: The stepping model parameters fit to all 40 cells.

Cell	α_{init}	α_{out}	α_{in}	p_{-high}	p_{-low}	p_{zero}	p_{+low}	p_{+high}	ϕ_{-high}	ϕ_{-low}	ϕ_{zero}	ϕ_{+low}	ϕ_{+high}	r
1*	16.8	4.1	36.3	0.972	0.975	0.977	0.968	0.951	0.02	0.35	0.82	0.88	0.98	0.55
2*	7.7	0.7	19.9	0.815	0.957	0.977	0.982	0.987	0.01	0.20	0.36	0.48	0.94	0.79
3*	8.4	3.5	20.8	0.960	0.969	0.954	0.971	0.974	0.15	0.23	0.25	0.60	0.89	1.63
4	70.9	11.1	33.2	0.944	0.968	0.972	0.981	0.991	0.46	0.51	0.53	0.59	0.80	0.94
5*	3.4	0.8	9.5	0.953	0.979	0.944	0.980	0.957	0.03	0.39	0.50	0.52	0.78	1.42
6*	18.4	6.6	23.7	0.801	0.902	0.871	0.912	0.276	0.04	0.09	0.42	0.51	0.97	3.02
7	1.7	0.4	24.1	0.852	0.947	0.958	0.971	0.964	0.02	0.48	0.74	0.93	0.97	1.71
8*	1.3	10.9	29.2	0.994	0.990	0.990	0.987	0.970	0.07	0.26	0.57	0.59	0.62	0.38
9*	74.2	11.0	39.7	0.902	0.984	0.986	0.991	0.993	0.16	0.43	0.58	0.69	0.93	0.64
10	21.0	3.6	49.5	0.975	0.984	0.982	0.983	0.992	0.08	0.25	0.51	0.62	0.86	0.68
11*	11.8	3.3	11.0	0.829	0.737	0.406	0.374	0.360	0.37	0.72	0.84	0.84	0.92	2.13
12*	2.5	3.0	8.6	0.887	0.724	0.829	0.451	0.800	0.44	0.26	0.40	0.26	0.76	4.57
13*	5.1	12.0	21.9	0.986	0.985	0.975	0.972	0.924	0.58	0.29	0.37	0.44	0.72	0.86
14*	49.4	2.6	11.0	0.401	0.524	0.981	0.444	0.489	0.03	0.18	0.38	0.44	0.88	0.04
15*	38.9	5.4	16.9	0.972	0.989	0.990	0.985	0.977	0.45	0.67	0.54	0.67	0.80	0.38
16*	14.5	8.2	20.8	0.700	0.421	0.536	0.526	0.346	0.23	0.45	0.46	0.45	0.65	1.64
17*	11.6	31.5	60.9	0.978	0.957	0.951	0.964	0.957	0.34	0.31	0.38	0.69	0.72	0.54
18*	43.6	37.2	42.7	0.467	0.445	0.433	0.461	0.462	0.17	0.12	0.38	0.49	0.90	1.80
19*	11.2	3.6	13.3	0.545	0.695	0.762	0.559	0.436	0.15	0.20	0.34	0.48	0.81	1.87
20*	21.2	10.3	26.1	0.900	0.545	0.507	0.387	0.850	0.32	0.70	0.74	0.79	0.89	2.11
21*	35.6	24.3	27.4	0.361	0.757	0.779	0.878	0.908	0.17	0.39	0.69	0.64	0.64	5.42
22*	3.5	8.5	20.4	0.970	0.964	0.965	0.882	0.957	0.41	0.44	0.44	0.37	0.65	0.91
23	19.6	8.6	41.7	0.945	0.962	0.966	0.975	0.967	0.04	0.23	0.44	0.72	0.94	2.01
24*	21.3	1.7	10.4	0.965	0.967	0.953	0.962	0.982	0.15	0.48	0.75	0.64	0.58	0.73
25	8.6	23.6	44.9	0.996	0.993	0.991	0.988	0.985	0.23	0.14	0.37	0.43	0.63	0.27
26*	4.8	15.3	33.8	0.994	0.991	0.983	0.976	0.980	0.10	0.20	0.39	0.34	0.44	0.57
27*	3.1	13.8	21.7	0.994	0.963	0.972	0.727	0.967	0.38	0.68	0.75	0.69	0.83	0.12
28*	32.9	17.6	38.3	0.766	0.876	0.934	0.542	0.323	0.04	0.26	0.26	0.60	0.65	1.53
29*	1.6	7.1	14.2	0.986	0.985	0.974	0.969	0.953	0.09	0.40	0.40	0.61	0.36	1.44
30*	19.1	5.3	25.3	0.941	0.848	0.947	0.677	0.946	0.05	0.60	0.54	0.75	0.87	1.19
31*	19.9	10.9	29.4	0.871	0.859	0.882	0.449	0.850	0.06	0.41	0.40	0.59	0.80	0.84
32	4.1	0.6	41.0	0.988	0.983	0.982	0.975	0.972	0.10	0.30	0.71	0.82	0.98	1.05
33*	24.1	2.5	9.7	0.940	0.987	0.992	0.989	0.971	0.24	0.43	0.30	0.62	0.64	0.17
34	12.6	1.5	23.7	0.935	0.971	0.980	0.973	0.987	0.03	0.37	0.53	0.70	0.55	1.56
35*	2.5	7.2	18.8	0.910	0.925	0.862	0.923	0.739	0.27	0.57	0.31	0.58	0.31	4.21
36*	7.5	10.4	26.9	0.591	0.829	0.730	0.870	0.785	0.10	0.25	0.47	0.64	0.86	4.32
37	26.7	1.4	44.1	0.955	0.985	0.984	0.986	0.985	0.05	0.21	0.54	0.53	0.81	0.70
38*	5.8	1.7	10.8	0.646	0.538	0.569	0.325	0.882	0.14	0.33	0.59	0.71	0.96	3.39
39	5.8	1.1	16.8	0.979	0.987	0.987	0.990	0.983	0.10	0.16	0.33	0.62	0.82	0.74
40*	23.5	11.3	19.3	0.689	0.531	0.497	0.252	0.399	0.24	0.18	0.53	0.70	0.88	2.37

Table S1: Posterior mean stepping model parameters for all cells. Stars next to the cell number indicate those cells we identified as tentative steppers (positive DIC difference of any magnitude).

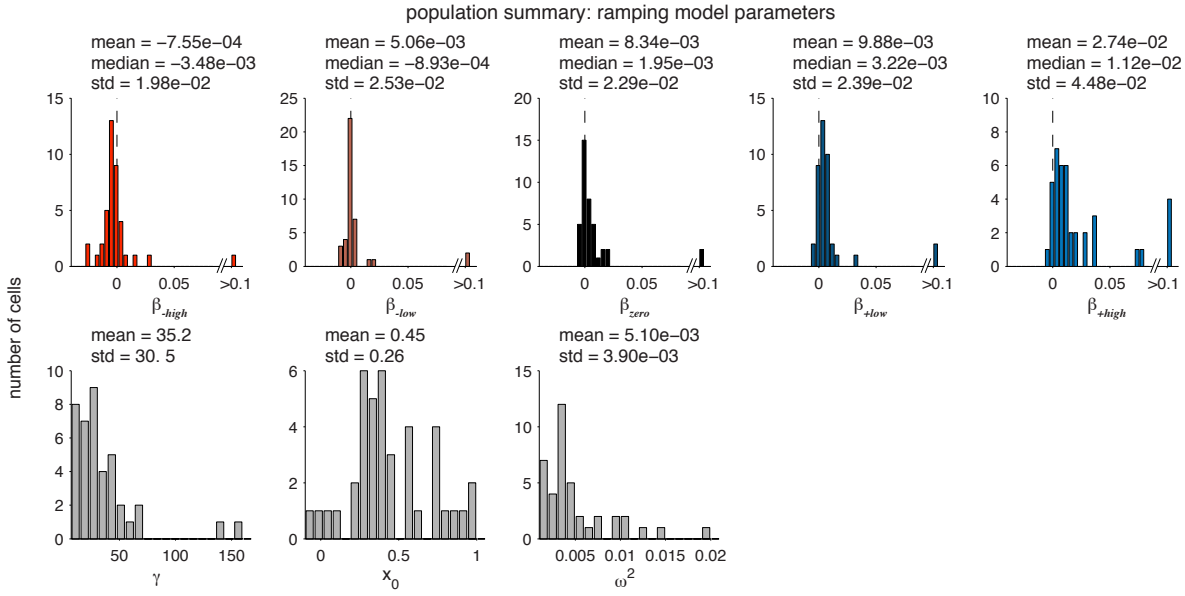

Fig. S11: The ramping model parameters fit to all 40 cells.

Cell	β_{-high}	β_{-low}	β_{zero}	β_{+low}	β_{+high}	x_0	ω^2	γ	DIC difference
1	-5.59e-03	-1.43e-03	1.02e-02	1.31e-02	3.45e-02	0.44	5.72e-03	37.0	141.8
2	-2.55e-02	-1.05e-02	-5.34e-03	8.48e-05	1.37e-02	0.36	1.09e-02	19.2	41.8
3	-1.47e-03	-3.21e-04	3.01e-05	3.28e-03	7.42e-03	0.40	3.05e-03	19.8	4.7
4*	-4.42e-03	-3.71e-03	-3.77e-03	-2.35e-03	3.94e-04	0.45	1.93e-03	138.6	-151.6
5	-8.72e-03	-6.58e-04	5.76e-04	9.20e-04	6.00e-03	0.38	3.80e-03	9.4	10.9
6	-8.47e-03	-7.17e-03	-3.56e-03	-2.16e-03	1.19e-01	0.76	1.56e-03	23.4	44.0
7*	-1.52e-02	-5.89e-04	4.69e-03	5.87e-03	1.21e-02	-0.05	1.01e-02	35.9	-164.4
8	-2.11e-03	1.73e-03	4.04e-03	6.12e-03	1.27e-02	0.19	6.75e-03	33.0	66.8
9	-3.45e-03	-1.97e-03	-7.70e-04	-9.58e-05	4.75e-04	0.36	1.71e-03	158.1	128.9
10*	-5.08e-03	-2.55e-03	1.36e-03	2.66e-03	4.82e-03	0.32	4.92e-03	62.0	-54.4
11	-1.30e-02	-6.06e-03	7.19e-03	1.49e-03	7.89e-02	0.93	5.18e-03	10.8	22.5
12	2.37e-03	1.67e-03	2.94e-03	5.53e-03	1.45e-02	0.29	2.87e-03	7.9	53.5
13	5.99e-03	4.03e-03	6.12e-03	7.78e-03	1.32e-02	0.30	3.16e-03	21.7	10.5
14	-2.79e-03	5.44e-04	6.52e-03	9.12e-03	2.21e-01	0.27	7.17e-03	11.7	70.6
15	-3.26e-03	-1.15e-03	-1.68e-03	-1.52e-03	-1.47e-03	0.38	2.67e-03	63.9	60.7
16	-2.02e-03	8.35e-04	1.17e-03	1.93e-03	4.38e-03	0.56	2.46e-03	23.4	28.4
17	3.01e-03	3.82e-03	4.97e-03	8.25e-03	1.03e-02	0.35	3.44e-03	70.4	96.0
18	1.08e-01	1.13e-01	1.09e-01	1.03e-01	1.17e-01	0.97	9.27e-03	39.5	27.5
19	-3.50e-03	-2.55e-03	5.31e-04	2.63e-03	3.51e-02	0.48	4.43e-03	14.5	22.1
20	-4.56e-03	-2.45e-03	1.51e-02	3.14e-03	2.73e-02	0.86	3.14e-03	26.0	0.5
21	2.77e-02	1.10e-01	9.63e-02	1.15e-01	1.10e-01	0.98	4.29e-03	29.5	170.5
22	-1.32e-05	2.67e-03	2.54e-03	4.09e-03	5.81e-03	0.29	4.31e-03	25.9	1.1
23*	-1.91e-03	-3.18e-04	1.36e-03	3.07e-03	5.32e-03	0.39	1.49e-03	46.4	-19.9
24	-1.07e-02	-6.20e-03	-3.44e-03	-5.04e-03	-3.94e-03	0.59	3.60e-03	25.5	38.4
25*	-2.36e-04	-6.52e-04	9.40e-04	3.17e-03	7.00e-03	0.40	3.45e-03	52.7	-211.3
26	2.43e-04	7.32e-04	3.59e-03	4.24e-03	4.86e-03	0.23	3.44e-03	39.8	19.6
27	3.77e-03	1.41e-02	1.72e-02	3.19e-02	2.67e-02	0.71	1.43e-02	20.5	9.9
28	-4.65e-03	-2.38e-03	-1.25e-03	5.69e-04	3.42e-03	0.72	1.69e-03	39.7	63.0
29	1.28e-03	2.89e-03	6.13e-03	9.51e-03	1.21e-02	0.09	4.00e-03	13.6	8.9
30	-8.49e-03	-2.40e-03	-1.66e-03	1.24e-03	6.76e-03	0.71	3.98e-03	26.5	57.3
31	-4.27e-03	2.47e-05	-1.11e-04	2.48e-03	4.74e-03	0.58	1.57e-03	32.2	76.4
32*	-4.55e-03	-2.68e-03	4.64e-03	8.97e-03	1.63e-02	0.03	9.99e-03	50.4	-227.4
33	-6.09e-03	-1.67e-03	-2.64e-03	-1.71e-05	1.04e-05	0.36	3.45e-03	26.6	80.7
34*	-1.23e-02	-2.32e-03	2.89e-04	2.51e-03	7.81e-04	0.55	3.56e-03	23.0	-9.9
35	1.57e-02	1.91e-02	2.12e-02	1.78e-02	3.76e-02	0.01	1.93e-02	11.0	24.6
36	1.89e-03	4.17e-03	7.94e-03	7.48e-03	1.92e-02	0.27	1.93e-03	28.8	4.7
37*	-2.44e-02	-9.12e-03	-3.99e-03	-3.45e-03	5.35e-03	0.65	7.49e-03	42.9	-8.9
38	-7.20e-03	-4.51e-03	4.67e-03	8.67e-03	2.11e-02	0.42	1.21e-02	11.6	22.0
39*	-4.72e-03	-2.48e-03	-1.43e-04	3.28e-03	7.49e-03	0.29	3.29e-03	18.1	-0.3
40	-1.76e-03	-1.13e-03	2.03e-02	1.11e-02	7.41e-02	0.81	2.40e-03	18.2	15.9

Table S2: Posterior mean ramping model parameters for all cells, and the DIC differences from the model comparison (positive indicating support for the stepping model). Stars next to the cell number indicate those cells we identified as tentative rampers (negative DIC difference of any magnitude).

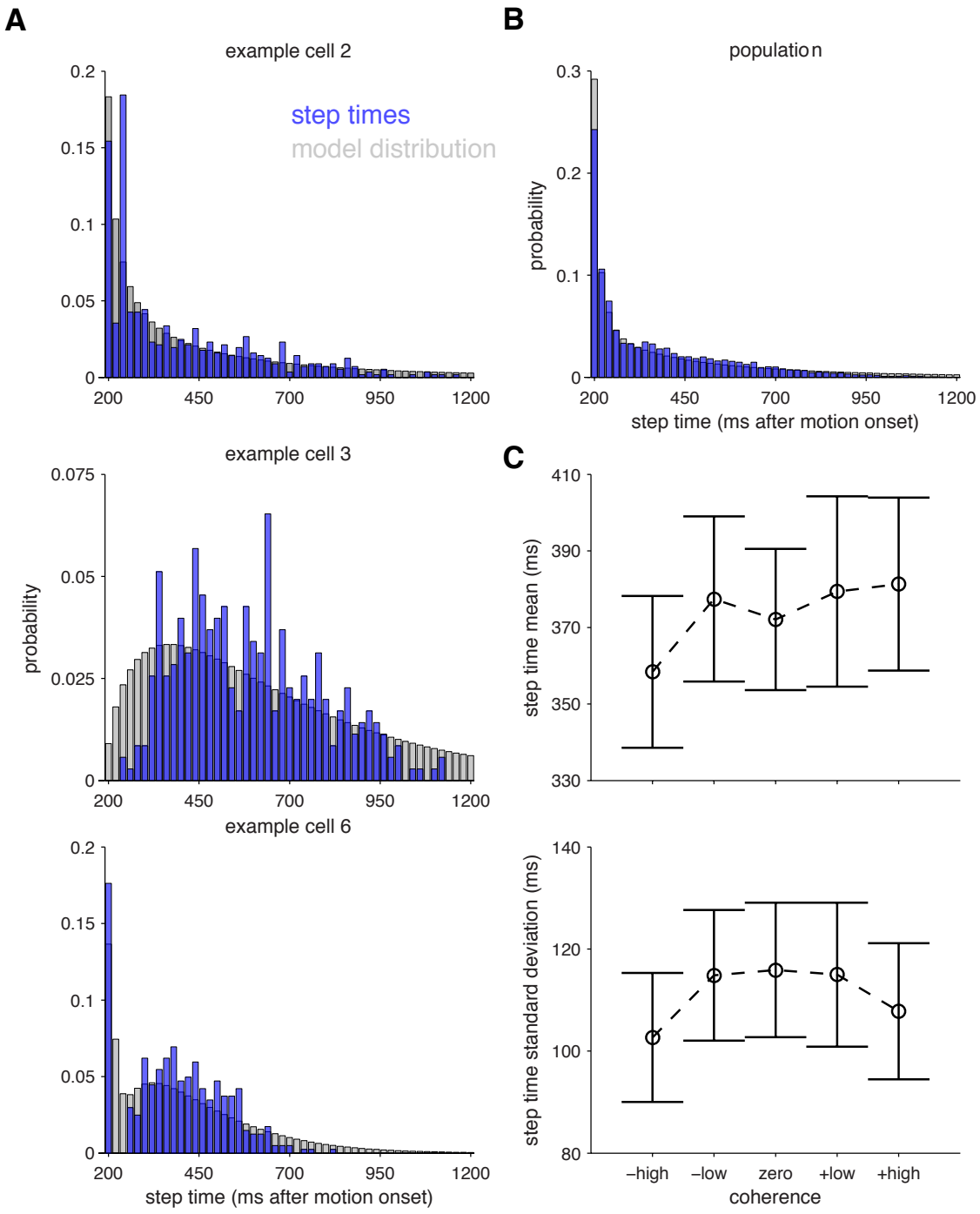

Fig. S12: Model-estimated step times for our LIP cells. (**A**) The step times relative to motion onset estimated for 3 example cells (blue) across all dot coherences. The model fit step time distributions are shown in grey. (**B**) The step time distribution across the entire population. (**C**) The mean step times averaged across cells for each coherence level (top) and the standard deviation of step times averaged across cells (bottom). Only trials for which we could estimate a step time were used in this figure.

3.2 Stepping model results

3.2.1 Related to main text Figure 2: single-cell examples

Here we include more single cell examples of the step-based analysis, similar to Figure 2 in the main text (Figures S13-S15). We performed the same analysis on spike trains from a simulated ramping cell S16.

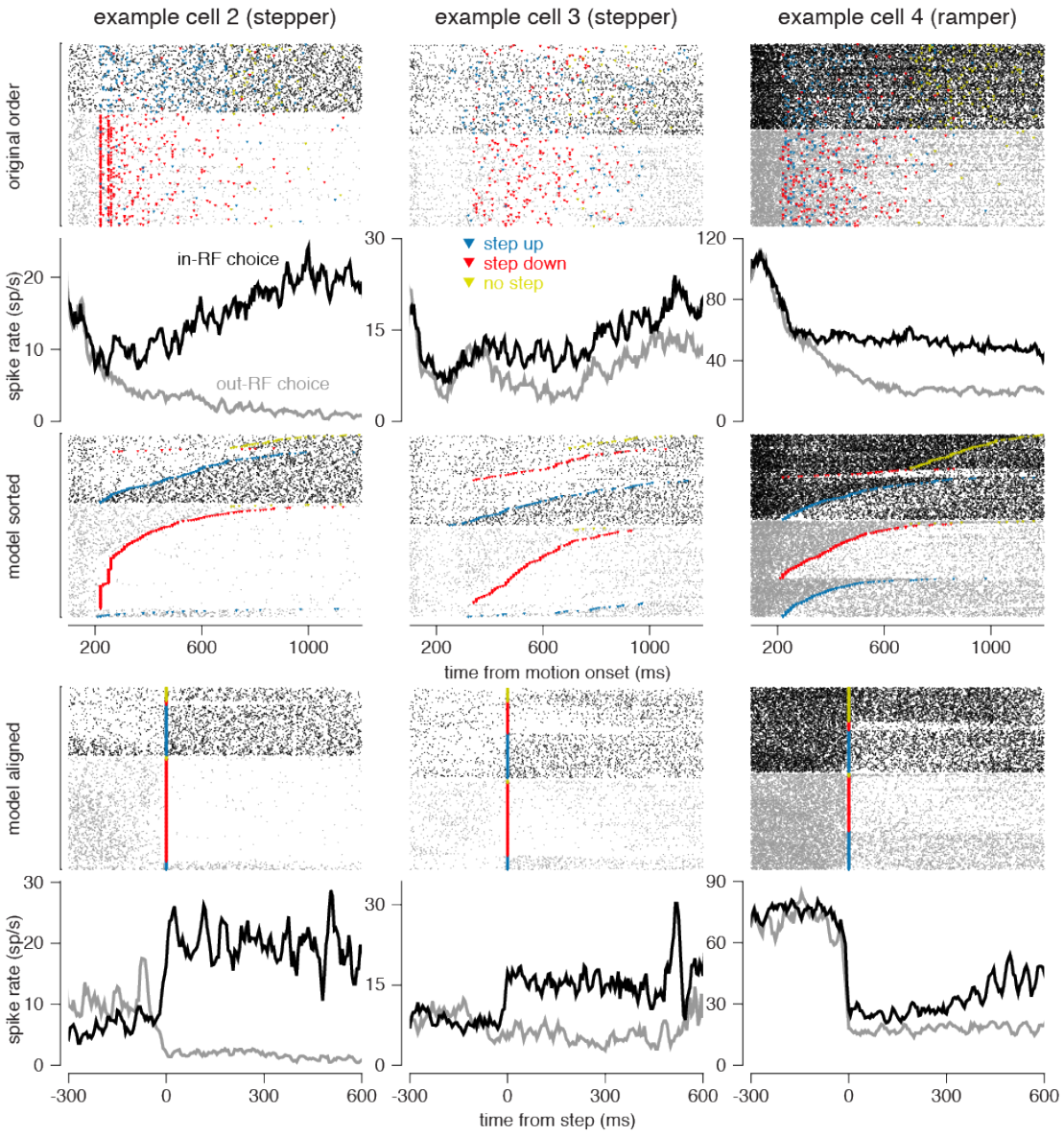

Fig. S13: Each column shows the responses of an example LIP cell with the same step-model analysis performed on the cell in Figure 2A. The top row shows all the trials aligned to stimulus onset, sorted by choice, and ordered in the order the trials were collected. 2nd row shows the average firing rate aligned to motion onset. 3rd row is the same data as in row 1, although the trials have been ordered by step time. In the 4th row, the trials have been aligned to the step time. The spike rate aligned to step time is given in the bottom row. Cells that were fit better by the ramping model are labeled as “rampers” and cells better fit by the stepping model are labeled “steppers”.

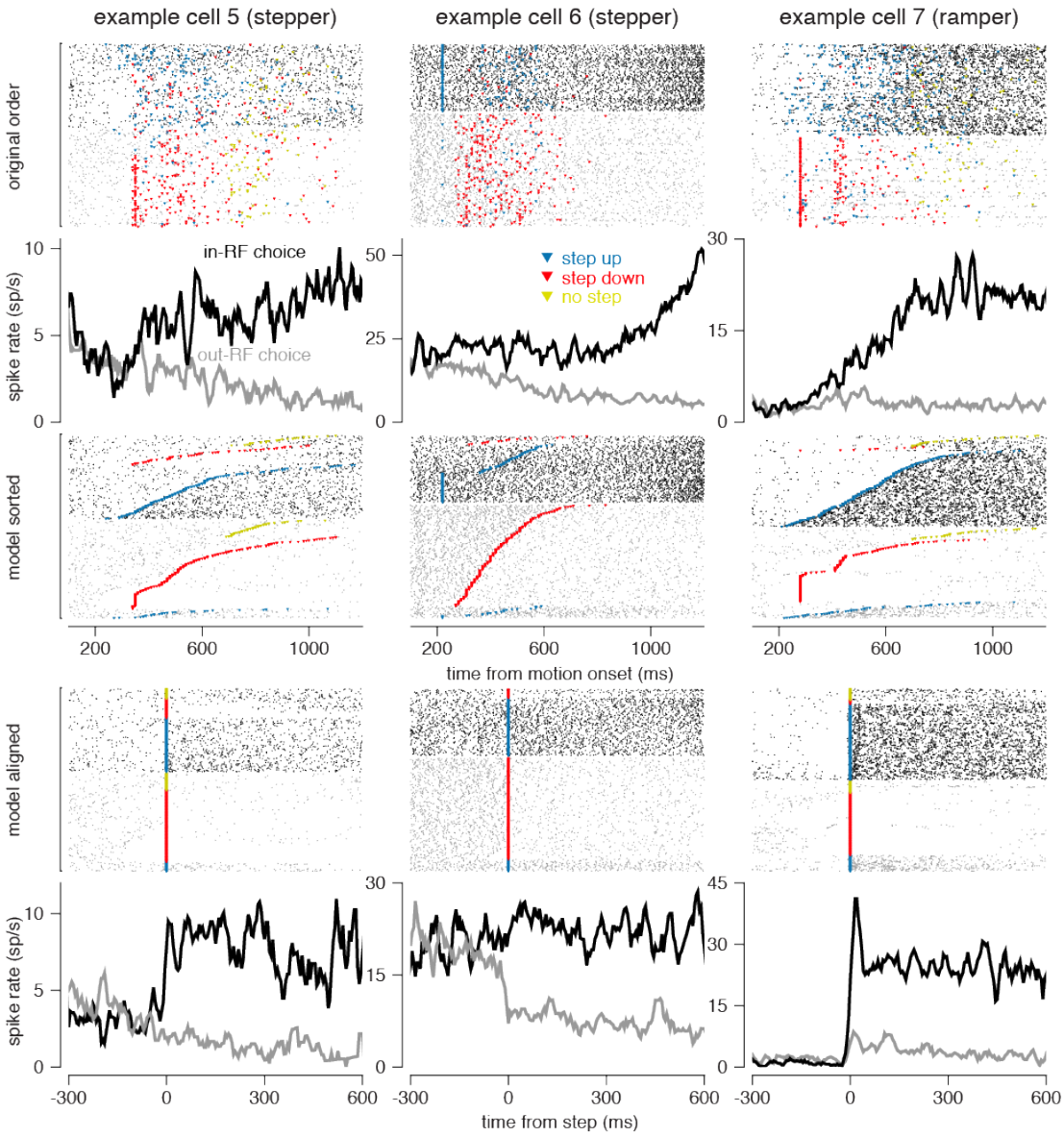

Fig. S14: Same as Figure S13 for 3 more cells.

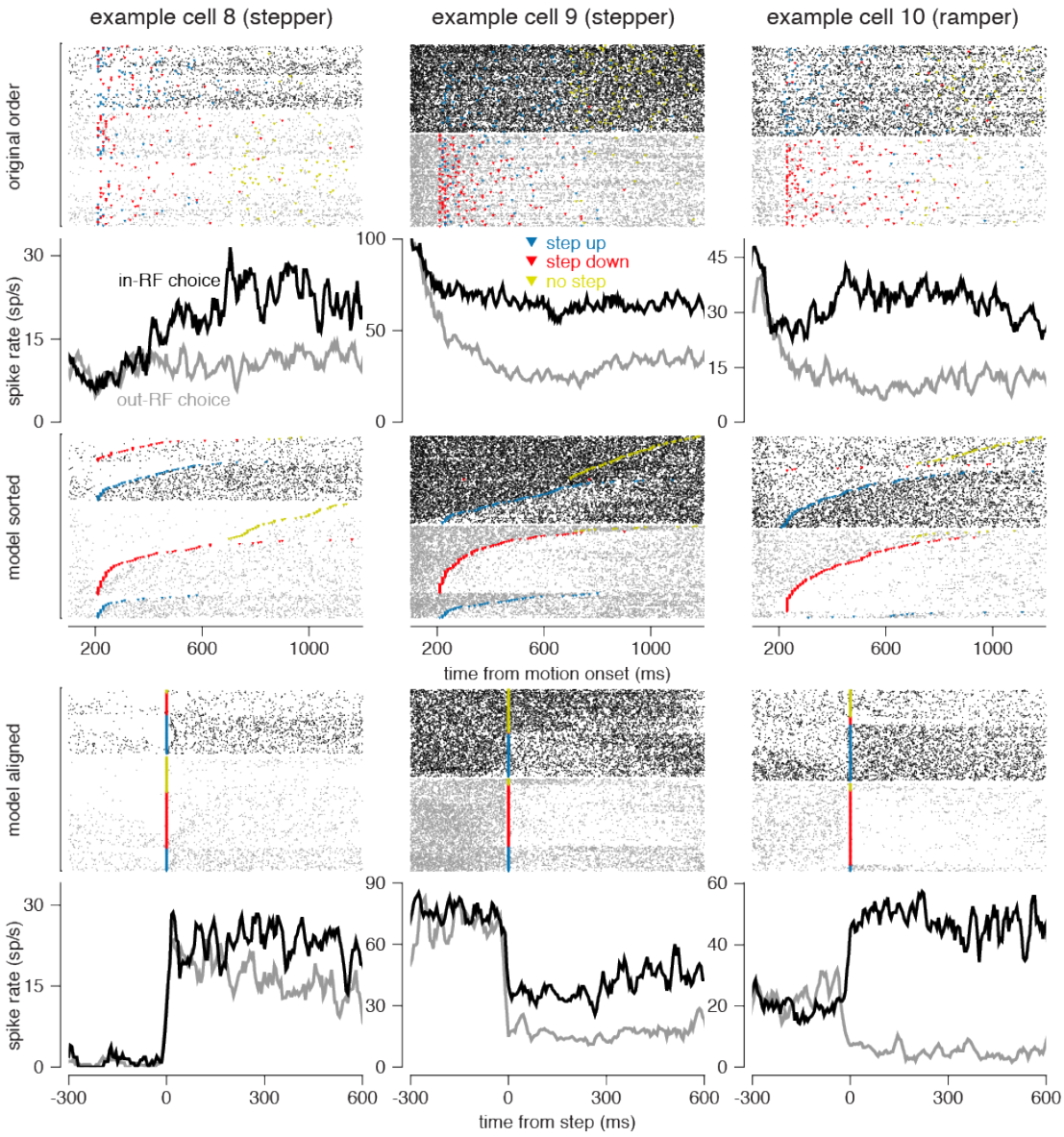

Fig. S15: Same as Figure S13 for 3 more cells.

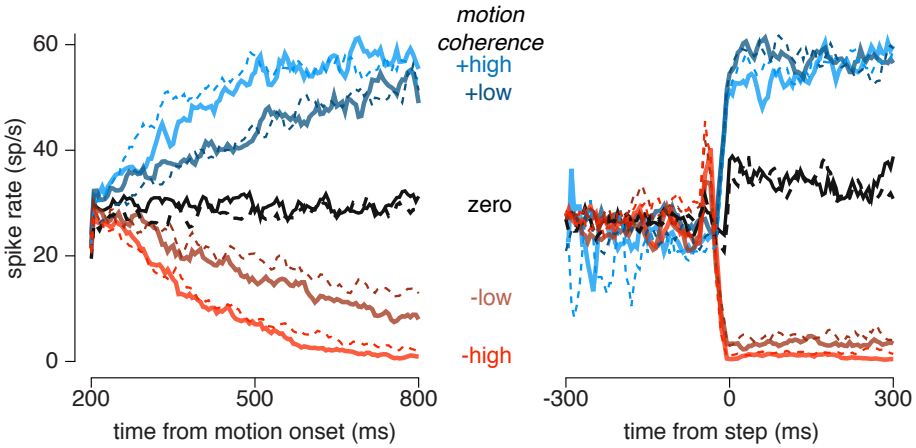

Fig. S16: Coherence-sorted PSTH for a simulated ramping cell (solid traces) aligned to motion onset (left) and estimated step times (right). The output of the stepping model fit to the simulated ramping cell is plotted in the dashed traces. The step-aligned PSTH revealed an apparent step in the ramping simulation. Although we found evidence of a nonzero slope in the step-aligned PSTHs from the ramping simulation before and after the step, these ramping slopes were small enough that visual inspection should be deemed insufficient (e.g., light blue curve, right panel). However, the quantitative model comparison (Section 2.3) correctly identified the simulated responses as arising from the ramping model. Thus, the average spike rate over trials aligned to a specific event can provide some evidence that a potential model is a viable description of the data, but conclusive tests between ramping and stepping models require additional quantitative assays. The Bayesian model comparison provides a stronger measure for hypothesis testing by quantifying which model best predicts the observed spike trains.

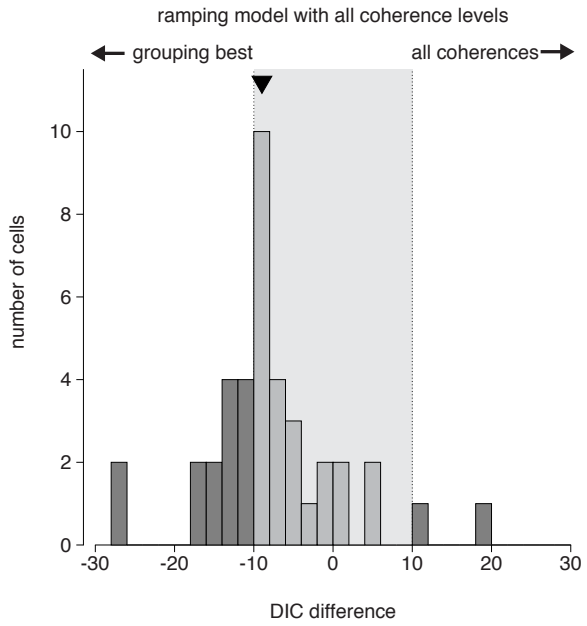

Fig. S17: Model comparison between the ramping model fit using all coherences compared to the ramping model fit with grouped coherence levels. The median DIC difference is denoted by the black triangle.

3.3 Model comparison results are unaffected by grouping coherence levels

We grouped stimulus coherence levels into 5 levels ($\pm$high,$\pm$low, and zero) in order to limit the number of model parameters. To ensure that this grouping did not bias our results in favor of the stepping model, we fit the ramping model to the data using all coherence levels (11 levels) and compared to the stepping model fits with grouped coherences. In general, model comparison showed that the ramping model with grouped coherence levels performed better than including all coherences (Fig. S17). Only 6 cells showed slightly better performance on all coherence levels. Of those cells, 2 were originally classified as "rampers". For 3 of the remaining 4 "stepping" cells, the grouped-coherence stepping model still provided a better fit than the all-coherence ramping model. For the final cell, the all-coherence ramping model provided a slightly better fit than the grouped-coherence stepping model (DIC difference -1.27). However, the stepping model fit with all coherences to this cell better described the cell than the all-coherence ramping model (DIC difference 8.84).

3.4 Model comparison results are consistent across start time of analysis

For our main analysis, we made the assumption (used in previous publications) that the motion integration signal in LIP began at approximately 200 ms (*11*). However, the population PSTH in figure 3A suggests that coherence sorting in the population may start earlier than 200 ms. We therefore repeated the model comparison using spike trains beginning at 160 ms, 180 ms, and 220 ms after motion onset, but still ending 200 ms after motion offset. The model comparison results were similar across all 3 analysis windows (Fig. S18). The median DIC difference was comparable across time points: 17.3, 22.4, 22.1, 23.4, for the 160, 180, 200, and 220 ms start times respectively. In the 160 ms analysis, 30 cells were better fit by the stepping model (24 showed strong support) and 10 cells were better fit

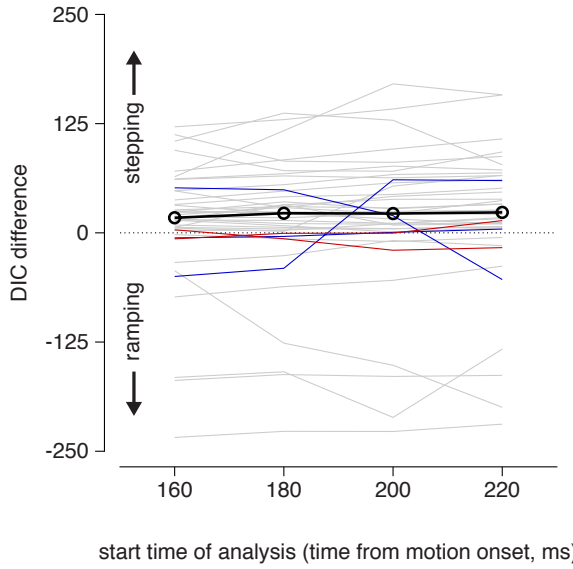

Fig. S18: Model comparison between the ramping model fit using different start times for the analysis. The median DIC (black) was significantly greater than 0 for all time points (sign test, $p < 0.004$ for each start time). 37 cells were consistently classified as a ramper or stepper across all times (grey traces). Two cells (red trace) were classified as a steppers by the 200 ms analyses, but showed slightly better support for stepping at the 160 ms, 180 ms, or 220 ms analyses. Three cells (blue traces) were classified as a stepper by the 200 ms analysis, but showed better support for ramping at one of the other start times.

by the ramping model (7 showed strong support). In the 180 ms analysis, 29 cells were better fit by the stepping model (24 showed strong support) and 11 cells were better fit by the ramping model (7 showed strong support). In the 220 ms analysis, 31 cells were better fit by the stepping model (27 showed strong support) and 9 cells were better fit by the ramping model (8 showed strong support).

Only 3 out of the original 31 "stepping" cells were better fit by the ramping model at a different analysis start point (Fig. S19). One of these cells showed weak support for stepping or ramping at all time points ($|\text{DIC difference}| < 10$). Two of the 9 original "ramping" cells were better fit by the stepping model at a different start point (Fig. S20). One of these cells showed only small support for the stepping model (DIC difference = 3.5) in the 160 ms analysis. The other cell showed only weak support for ramping in the original analysis. These model comparison changes seen in these cells do not alter the overall result. We believe these changes are not due to an earlier onset of integration, and are primarily the result of including responses to the onset of targets and dot motion that occur at the earlier portions of the trials (*14, 33*).

3.5 Comparison to existing methods

Here we compare our methods with two recent approaches for analyzing the latent dynamics of LIP spike trains.

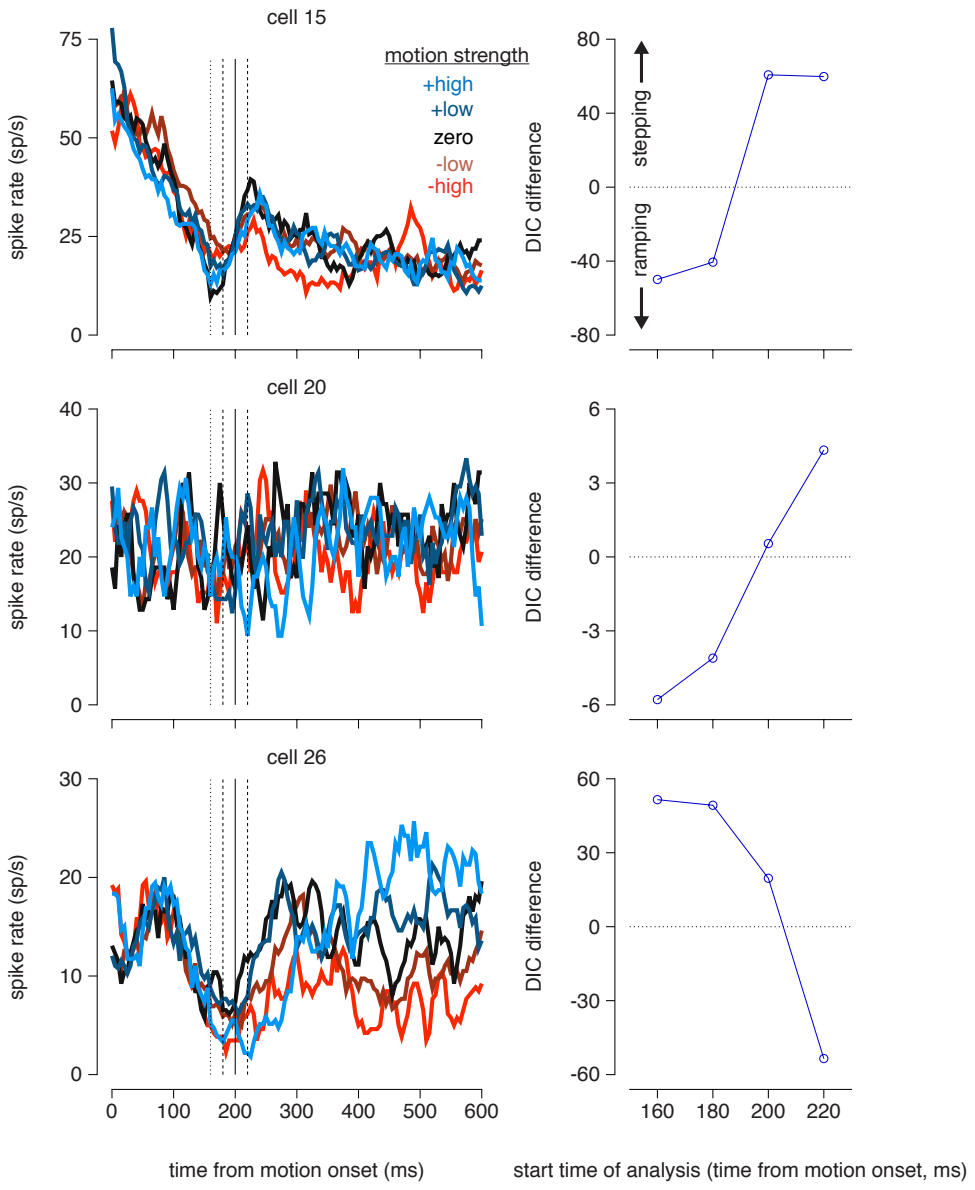

Fig. S19: The three rows show the coherence-sorted PSTH (left) of cells identified as steppers in the original analysis that were better fit bit the ramping model at different start times of analysis. (right) The model comparison metric for each cell is given for all start times. The PSTHs were estimated using only a 25 ms sliding window, which makes the firing rate estimate very noisy for single cells.

3.5.1 Churchland et al. (2011): moment-based ("VarCE") method

Churchland et al. 2011 (*11*) introduced a method for analyzing the dynamics of spike trains based on the variance of the conditional expectation (or "VarCE"). This method employs the law-of-total-variance to divide the time-varying spike count variance (referred to above as the PSTV) into two components: one due to the point process or spiking variability, and another due to the variability of the underlying latent process. The Churchland et al. method assumes that the first component is proportional to the spike

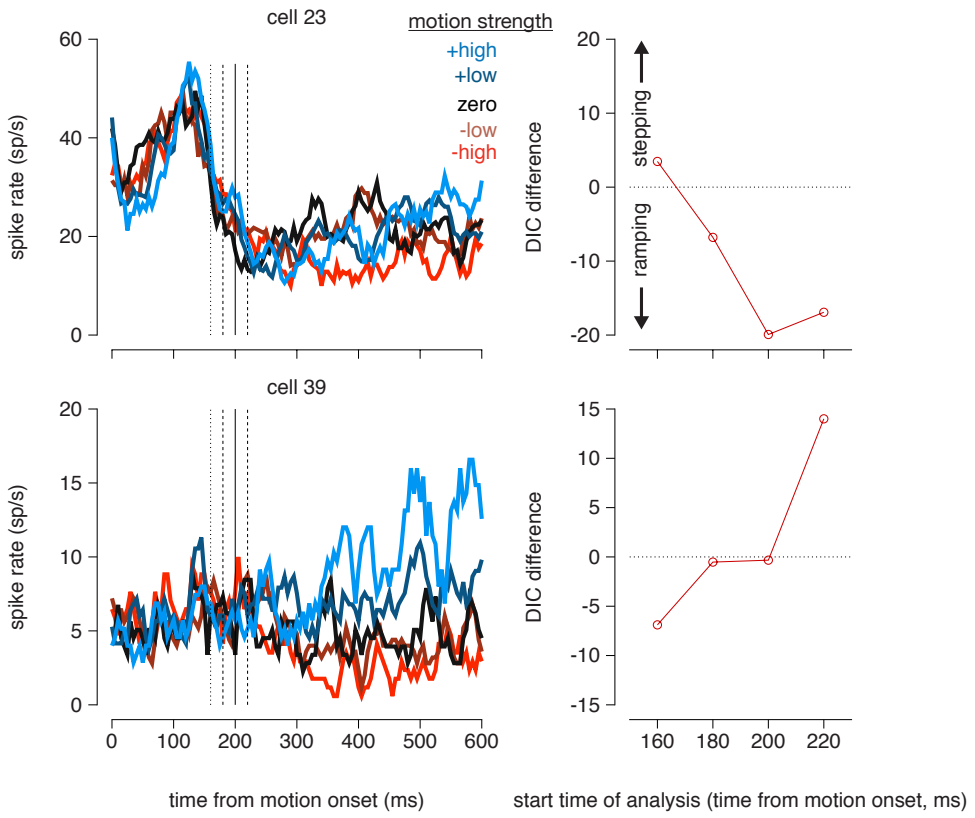

Fig. S20: The two rows show the coherence-sorted PSTH (left) of cells identified as rampers in the original analysis that were better fit bit the stepping model at different start times of analysis. (right) The model comparison metric for each cell is given for all start times.

rate, which holds true for any inhomogeneous renewal process (including the conditionally Poisson stepping and ramping models we have considered here). The second component is the VarCE, which is the quantity of interest in this analysis, since it corresponds to the variability of the latent process that drives spiking. The VarCE is calculated within a single time bin as

$$VarCE = \sigma_N^2 - \phi\sigma_{N|x}^2 \tag{100}$$

where σ_N^2 is the total spike count variance (PSTV), and $\phi\sigma_{N|x}^2$ is an estimate of the point-process variance, obtained by multiplying the mean spike count $\sigma_{N|x}^2$ (i.e., the PSTH for that time bin) by a scale factor ϕ. This scale factor is estimated by setting it to the cell's minimum observed Fano factor.

The basic intuition for this approach is that the shape of the VarCE over time should provide insight into the latent dynamics that underlie spiking. The VarCE of a continuous diffusion process should grow linearly, because noise accumulates linearly over time. A discrete stepping process, on the other hand, should have low VarCE at the beginning and end of a trial (assuming it always steps to the same final state), and high variance during the portion of the interval when steps are most likely.

Churchland et al. compared the shape of the VarCE curve estimated from neural data to that of a simulated stepping model, and concluded that LIP responses were inconsistent with stepping dynamics. However, the stepping simulations used to make this argument were restricted to "in"-RF choice, 0%

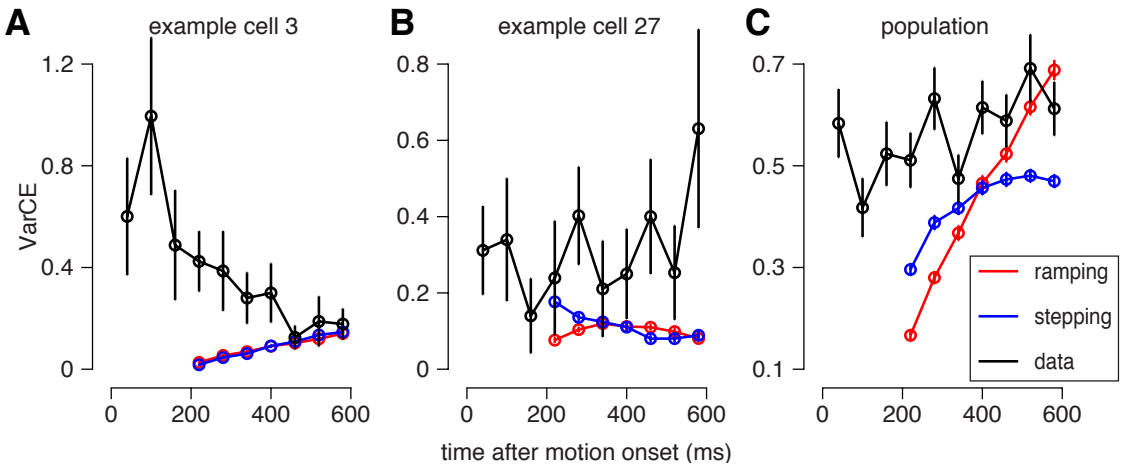

Fig. S21: VarCE from Churchland et al. 2011 (*11*) calculated from simulated spike trains and our data. The VarCE from the simulations was calculated in a 400 ms window beginning 200 ms after motion onset, and the VarCE from the data was calculated starting at motion onset. Spike count statistics were computed within a 60 ms time bin. (**A**) VarCE for the 0% coherence trials of an LIP cell (black). We simulated from ramping (red) and stepping fits to the cell (blue). Simulations included 50000 trials per cell– many more than can be collected in a real experiment. Cell parameters are given in Tables S1- S2. (**B**) Same as A for another cell. (**C**) VarCE for our LIP population (black) and from simulations of model fits to the population for 0% coherence trials.

coherence trials, and assumed that the mean response reflected a mixture of two response levels, one from early in the trial and one preceding the saccade. This produced a VarCE time course that was larger overall, and increased more steeply, than the VarCE from the data.

However, if a neuron's latent state is allowed to step up, step down, or to not step on some trials (which is analogous to a drift-diffusion path that does not hit the bound, or wanders downward on some "in" trials), then a stepping model (like ours) can produce a more flexible range of VarCE timecourses. We therefore decided to explore whether VarCE could definitively distinguish between the ramping and stepping dynamics implemented in our models.

Figure S21A shows the VarCE calculated for 0% coherence trials for an example cell (black traces), along with the VarCE of spike trains simulated from the two fitted models (red traces for ramping, blue traces for stepping). The ramping and stepping models produce nearly identical linear VarCE traces ($r^2 = 0.991$ and $r^2 = 0.992$ between a true linear ramp and the ramping and stepping model VarCEs respectively). Figure S21B shows an example cell for which the VarCE traces predicted by the two model fits show distinct nonlinear trends ($r^2 < 0.01$ and $r^2 = 0.83$ compared to a linear ramp for the ramping and stepping models respectively). However, estimates of VarCE are noisy for individual cells and the visual adequacy or superiority of either model is not particularly definitive. We therefore calculated the cell-averaged VarCE on the 0% coherence trials from the data and our model fits (Fig. S21C). We found that the VarCE of the stepping model fits (blue curve) matched the data more closely than ramping fits (red curve) (mean squared error ramping = 0.031, stepping = 0.026).

3.5.2 Bollimunta et al. (2012): a single-trial, spike train approach

Another recent study, Bollimunta et al. (*19*), reported a statistical analysis of LIP spike trains similar in spirit to our own. This paper examined specifically whether LIP spike trains were better fit by a ramping model or a discrete stepping model. It concluded that ramping model provided a better fit to LIP responses, in contradiction to the findings we have reported here. However, there are substantial differences in modeling and statistical methodology that may explain this discrepancy.

First, the Bollimunta et al. ramping model had a linearly increasing spike rate, without a diffusion component or a bound, which represents a significant departure from standard noisy accumulation-to-bound dynamics. Second, the Bollimunta et al. stepping model assumed a uniform distribution of step times and identical step directions for all trials being analyzed. However, we do not in general expect step times (like reaction times) to be uniformly distributed, nor to correlate perfectly with choice (see Fig. 2A).

There are also differences in statistical power. The model fitting techniques used in (*19*)had computational costs that made it infeasible to analyze more than 4 trials at a time. By considering only several trials at a time, the fitting procedure can produce different parameter fits for different trials. Bollimunta et al. used the Hannan-Quinn information criterion (HQIC) as a metric to compare models, which is similar to our use of DIC. The distribution of HQIC values computed across model fits of different trials was tested for being significantly greater than 0, instead of computing a single HQIC for all trials for a single cell. The magnitude of the median HQIC values reported by Bollimunta et al. are less than 0.01, which is several orders of magnitude smaller than the model comparison values we report here. Additionally, Bollimunta et al. used an analysis window consisting of 400ms preceding a saccade instead of the entire integration period. The MCMC methods we used to analyze a large number of trials, along with our definitions of both types of latent dynamics, increased the statistical power of our study. In summary, we feel these differences in modeling and methodology might explain the discrepancies in our findings.

3.6 Application to a response-time version of the task

We applied our model comparison to a publicly available dataset from Roitman & Shadlen (*17*) (downloaded from `https://www.shadlenlab.columbia.edu/`). The task was similar to our own, except the monkey was not given a "go" signal and instead viewed the dots until it chose to signal a decision with a saccade. The averaged responses of these choice-selective LIP cells are shown in Figure 7 of (*17*).

Similar to the analysis of our data, we applied our model fitting to spike trains starting from 200 ms after motion onset. Following (*19*), we considered spikes up until 50 ms before the saccade. We selected only trials that were at least 350 ms long, counting from motion onset until saccade. Therefore, every trial we analyzed had at least 100 ms of data for our analysis. After weeding out short trials, we selected cells with at least 8 remaining trials per signed coherence level. 16 cells from this dataset met this criterion, and their motion-aligned and choice-separated PSTHs are shown in Figure S22. The motion coherence-sorted responses along with the step model fits are shown in Figure S23. Our model comparison analysis indicated that 12 of these 16 cells were better fit by the stepping model than the ramping model (Fig. S25). Figure S24 shows the population average firing rate aligned to motion onset and inferred step times (similar to Fig. 3A in the main text for the fixed-duration task). The similar-

ity of these results to our model comparison using the larger experimenter-controlled dataset in the main paper suggests that the dynamics of LIP neurons are not fundamentally different in experimenter-controlled and reaction-time versions of the task, consistent with previous conclusions (*3*).

We note that the response time paradigm might allow for decision-related and saccade-related motor activity to overlap within a single spike train, even though we excluded a brief portion of the pre-saccadic activity from analysis. The models considered here were only intended to capture decision-related dynamics, so further analyses of LIP responses during the response time task will benefit from analyzing larger datasets with extended models that are explicitly designed to disentangle decision-related dynamics from pre-motor activity.

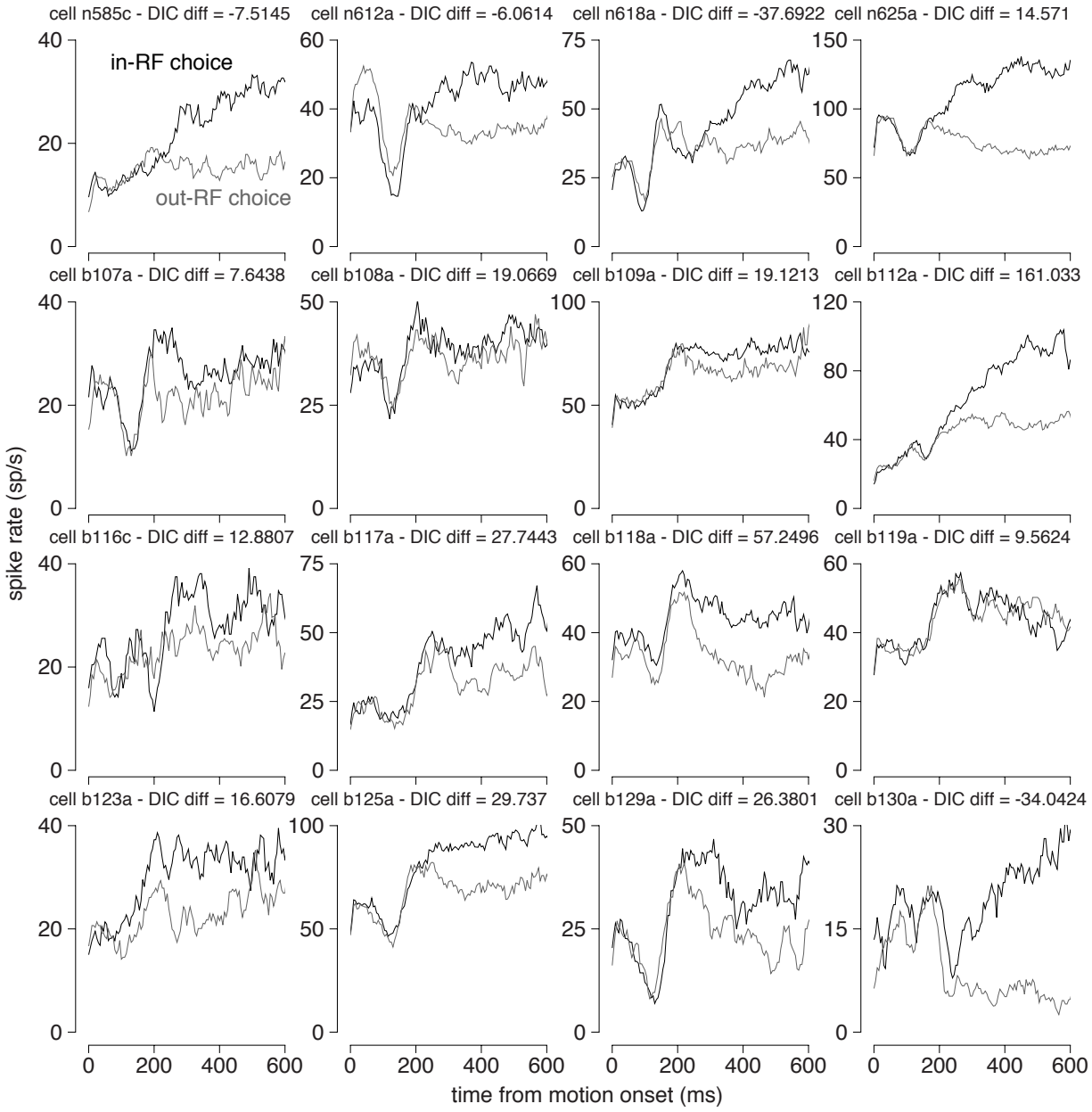

Fig. S22: PSTHs aligned to motion onset from the cells we analyzed from Roitman & Shadlen (*17*), sorted by the monkey's choice. DIC differences from our model comparison analyses are given for each cell (positive favoring the stepping model).

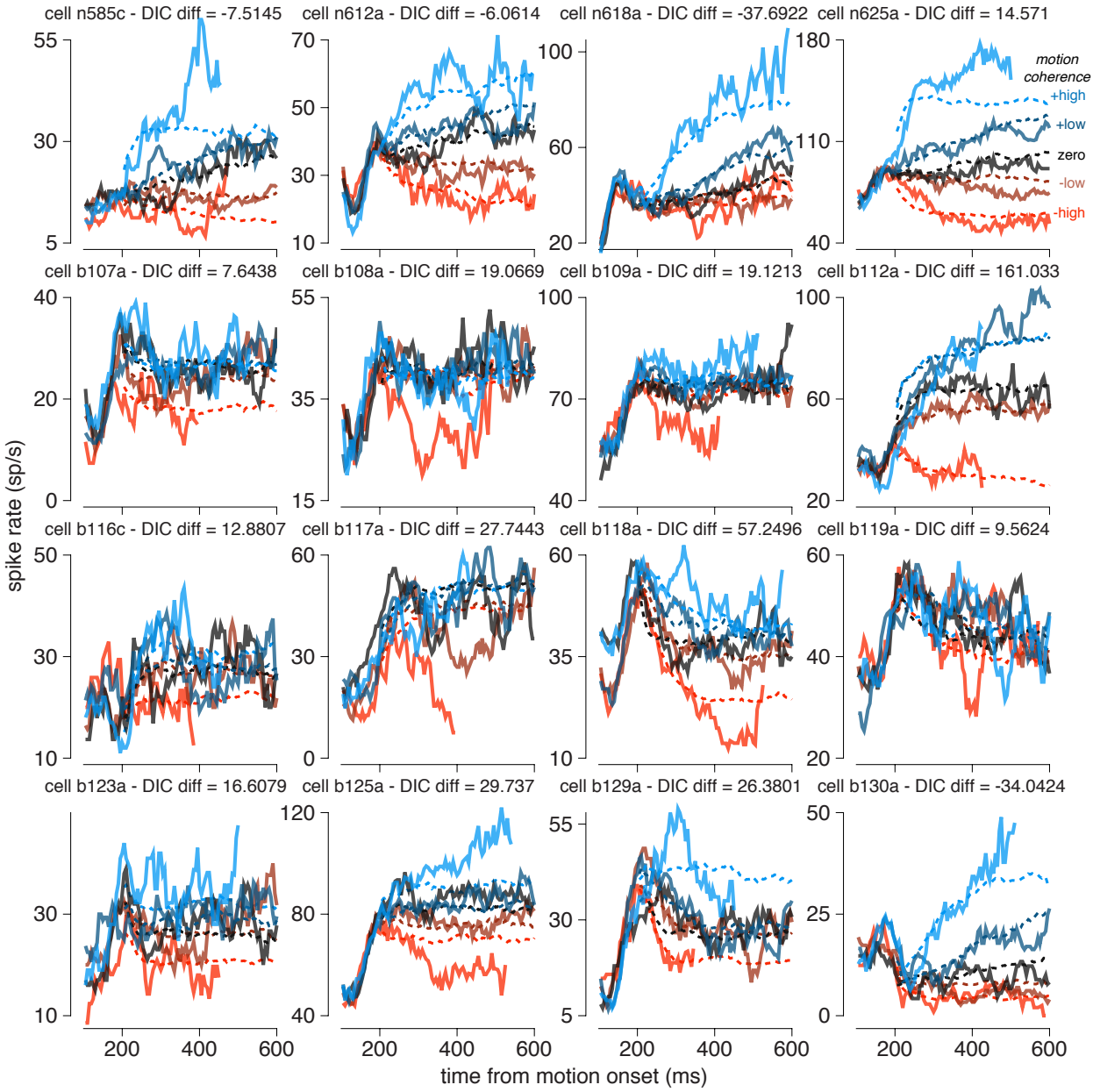

Fig. S23: PSTHs aligned to motion onset from the cells we analyzed from Roitman & Shadlen (*17*), sorted by motion coherence (solid lines). The stepping model fits are shown in the dashed lines. These PSTHs were smoothed using a 50 ms sliding average, rather than a 25 ms window, because the number of trials in each condition was limited. The average rates included activity up to 50 ms before the saccade. Firing rates are only shown at time points when at least 8 trials were available.

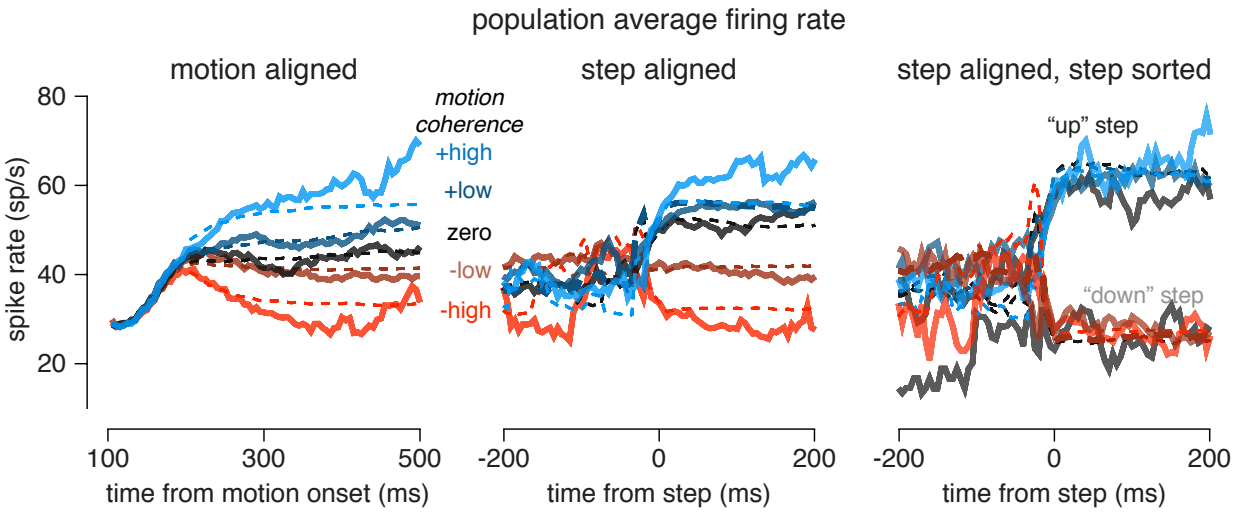

Fig. S24: Population average PSTH sorted by motion coherence computed from spike trains for the 16 cells we analyzed from Roitman & Shadlen (*17*): (left) aligned to motion onset and sorted by motion strength; (middle) aligned to step times inferred under the stepping model and sorted by motion strength; (right) aligned to step times and sorted by both motion strength and inferred step direction. Simulated results from the stepping model (dashed lines) provide a close match to the real data under all types of alignment and conditioning. These PSTHs were smoothed using a 50 ms sliding average, rather than a 25 ms window, because the number of trials in each condition was limited.

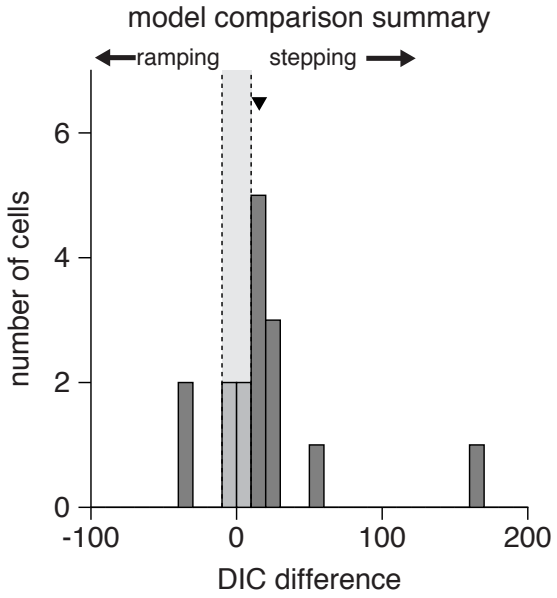

Fig. S25: Model comparison results for 16 cells from Roitman & Shadlen (*17*).

References

1. M. E. Mazurek, *Cerebral Cortex* **13**, 1257 (2003).
2. J. I. Gold, M. N. Shadlen, *Annual Review of Neuroscience* **30**, 535 (2007).
3. R. Kiani, T. D. Hanks, M. N. Shadlen, *Journal of Neuroscience* **28**, 3017 (2008).
4. R. Kiani, M. N. Shadlen, *Science* **324**, 759 (2009).
5. M. N. Shadlen, R. Kiani, *Neuron* **80**, 791 (2013).
6. M. N. Shadlen, W. T. Newsome, *Proceedings of the National Academy of Sciences* **93**, 628 (1996).
7. T. D. Hanks, *et al.*, *Nature* **520**, 220 (2015).
8. P. Miller, D. B. Katz, *The Journal of Neuroscience* **30**, 2559 (2010).
9. D. Durstewitz, G. Deco, *The European Journal of Neuroscience* **27**, 217 (2008).
10. M. S. Goldman, *Encyclopedia of Computational Neuroscience*, D. Jaeger, R. Jung, eds. (Springer, 2015), pp. 1177–1182.
11. A. K. Churchland, *et al.*, *Neuron* **69**, 818 (2011).
12. R. Ratcliff, J. N. Rouder, *Psychological Science* **9**, 347 (1998).
13. B. W. Brunton, M. M. Botvinick, C. D. Brody, *Science* **340**, 95 (2013).
14. M. L. R. Meister, J. A. Hennig, A. C. Huk, *The Journal of Neuroscience* **33**, 2254 (2013).
15. A. K. Churchland, R. Kiani, M. N. Shadlen, *Nature Neuroscience* **11**, 693 (2008).
16. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 583 (2002).
17. J. D. Roitman, M. N. Shadlen, *The Journal of Neuroscience* **22**, 9475 (2002).
18. I. H. Stevenson, K. P. Kording, *Nature Neuroscience* **14**, 139 (2011).
19. A. Bollimunta, D. Totten, J. Ditterich, *The Journal of Neuroscience* **32**, 12684 (2012).
20. R. Kiani, C. J. Cueva, J. B. Reppas, W. T. Newsome, *Current Biology* **24**, 1542 (2014).
21. M. T. Kaufman, M. M. Churchland, S. I. Ryu, K. V. Shenoy, *eLife* **4**, e04677 (2015).
22. K. P. Burnham, D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach* (Springer Science & Business Media, 2002).
23. N. Shephard, M. K. Pitt, *Biometrika* **84**, 653 (1997).
24. K. Yuan, M. Girolami, M. Niranjan, *Neural Computation* **24**, 1462 (2012).

25. N. Gordon, D. Salmond, A. Smith, *IEE Proceedings F Radar and Signal Processing* **140**, 107 (1993).

26. G. O. Roberts, O. Stramer, *Methodology And Computing In Applied Probability* **4**, 337 (2002).

27. M. Girolami, B. Calderhead, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 123 (2011).

28. E. S. Bromberg-Martin, M. Matsumoto, O. Hikosaka, *Neuron* **67**, 144 (2010).

29. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 485 (2014).

30. S. Chib, *Journal of the American Statistical Association* **90**, 1313 (1995).

31. S. Chib, I. Jeliazkov, *Journal of the American Statistical Association* **96**, 270 (2001).

32. I. Kang, J. H. Maunsell, *Journal of Neurophysiology* **108**, 3403 (2012).

33. I. M. Park, M. L. Meister, A. C. Huk, J. W. Pillow, *Nature Neuroscience* **17**, 1395 (2014).