nature communications

Article

https://doi.org/10.1038/s41467-022-34805-5

Visual motion perception as online hierarchical inference

Received: 21 October 2021

Accepted: 7 November 2022

Published online: 01 December 2022

Check for updates

Johannes Bill [1,2] ✉, Samuel J. Gershman [2,3,4,5] & Jan Drugowitsch [1,3,5]

Identifying the structure of motion relations in the environment is critical for navigation, tracking, prediction, and pursuit. Yet, little is known about the mental and neural computations that allow the visual system to infer this structure online from a volatile stream of visual information. We propose online hierarchical Bayesian inference as a principled solution for how the brain might solve this complex perceptual task. We derive an online Expectation-Maximization algorithm that explains human percepts qualitatively and quantitatively for a diverse set of stimuli, covering classical psychophysics experiments, ambiguous motion scenes, and illusory motion displays. We thereby identify normative explanations for the origin of human motion structure perception and make testable predictions for future psychophysics experiments. The proposed online hierarchical inference model furthermore affords a neural network implementation which shares properties with motion-sensitive cortical areas and motivates targeted experiments to reveal the neural representations of latent structure.

Efficient behavior requires identification of structure in a continuous stream of volatile and often ambiguous visual information. To identify this structure, the brain exploits statistical relations in velocities of observable features, such as the coherent motion of features composing an object (Fig. 1a). Motion structure thus carries essential information about the spatial and temporal evolution of the environment, and aids behaviors such as navigation, tracking, prediction, and pursuit[1–8]. It remains, however, unclear how the visual system identifies a scene's underlying motion structure and exploits it to turn noisy, unstructured, sensory impressions into meaningful motion percepts.

In recent years, Bayesian inference has provided a successful normative perspective on many aspects of visual motion perception[9–17]. Human perception of motion stimuli spatially constrained by an aperture is well-explained by Bayesian statistical inference[9–11,14], and neural circuits that integrate local retinal input into neural representations of motion have been identified[18–23]. For the perception of structured motion spanning multiple objects and larger areas of the visual field, however, a comprehensive understanding is only beginning to emerge[15,24–27]. While common fate, that is, the use of motion coherence for grouping visual features into percepts of rigid objects, received some experimental support[24,28], the perception of natural scenes requires more flexible structure representations (e.g., nested motion relations and non-rigid deformations) than common fate alone. Recent theoretical work[15] has introduced a representation of tree structures for the mental organization of observed velocities into nested hierarchies. Theory-driven experiments subsequently demonstrated that the human visual system indeed makes use of hierarchical structure when solving visual tasks[16], and that salient aspects of human motion structure perception can be explained by normative models of Bayesian inference over tree structures[17]. Because these studies were restricted to modeling motion integration only with regard to the perceptual outcome—they analyzed presented visual scenes offline using ideal Bayesian observer models—it remained unclear how the visual system solves the chicken-and-egg problem of parsing (in real time) instantaneous motion in a scene while simultaneously inferring the scene's underlying structure.

We address this question by formulating visual motion perception as online hierarchical inference in a generative model of structured motion. The resulting continuous-time model is able to explain human perception of motion stimuli covering classical psychophysics

[1]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [2]Department of Psychology, Harvard University, Cambridge, MA, USA. [3]Center for Brain Science, Harvard University, Cambridge, MA, USA. [4]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA. [5]These authors jointly supervised this work: Samuel J. Gershman, Jan Drugowitsch. ✉e-mail: johannes_bill@hms.harvard.edu

https://doi.org/10.1038/s41467-022-34805-5

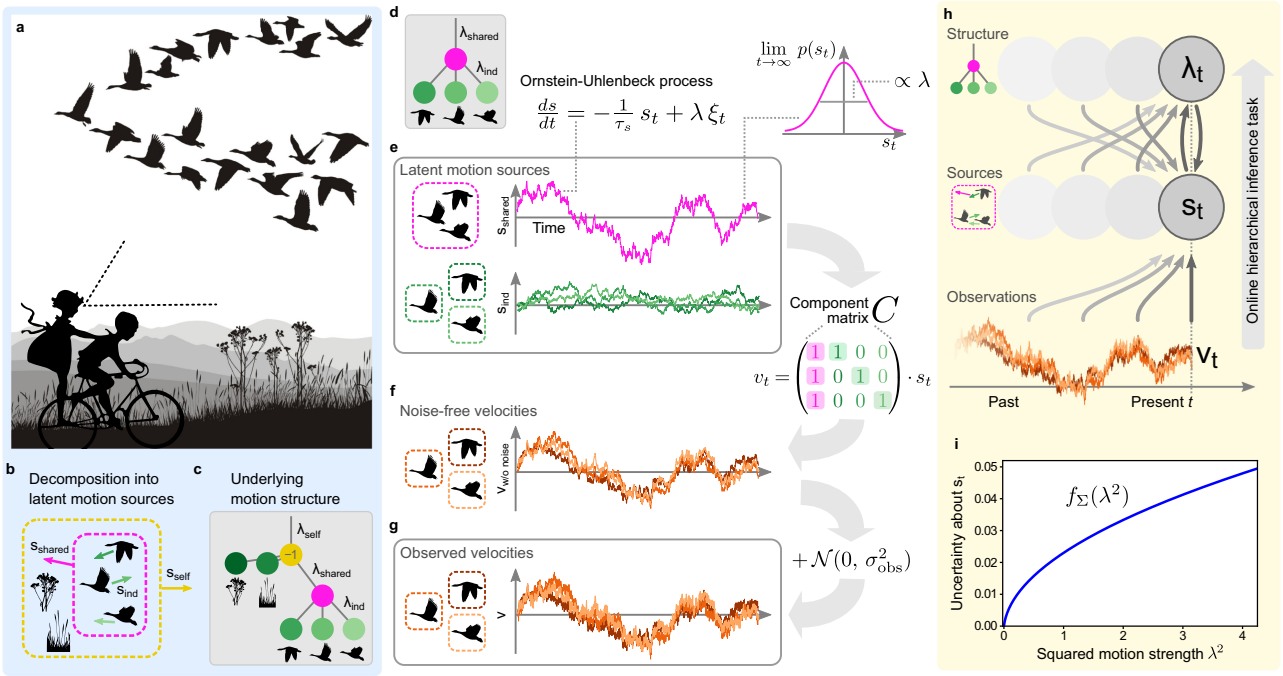


Fig. 1 | Visual motion perception as an online hierarchical inference task. **a** Scene with nested motion relations. Observed velocities reaching the observer's retina are perceived as a combination of self-motion, flock motion and every bird's individual motion relative to the flock. **b** Formal decomposition of the scene's motion into latent motion sources. **c** Tree-structured graph representation of the underlying motion structure with nodes corresponding to latent motion sources. Self-motion contributes in the opposite direction to retinal velocity (−1). Vertical distances between nodes, termed motion strengths, $\boldsymbol{\lambda}$, describe the long-term average speed of the source. Vanishing motion strength indicates that the corresponding motion source is not present the scene. **d**–**g** Generative model of structured motion. **d** Graph for a simpler motion scene with three flocking birds and a stationary observer. **e** Latent motion sources follow independent Ornstein–Uhlenbeck processes. **f** The component matrix, $\boldsymbol{C}$, composes noise-free velocities from the motion sources, such that each velocity is the sum of all its ancestral sources. **g** Observed velocities are noisy versions of the noise-free velocities. **h** Inverting the generative model according to Bayes' rule poses an online hierarchical inference task characterized by inter-dependent updates of motion sources and structure. **i** Using an adiabatic approximation, the motion sources' posterior variances reduce to a function of the motion strengths. Panels **a**–**h** are derived from artwork by Vladimír Čerešňák ("Migrating geese in the spring and autumn" licensed from Depositphotos Inc.) and Gordon Dylan Johnson ("Vintage Brother And Sister Bicycle Silhouette" from Openclipart.org, public domain).

experiments, ambiguous motion scenes, and illusory motion displays. The model, which relies on online Expectation-Maximization[29–31], separates inference of instantaneous motion from identifying a scene's underlying structure by exploiting the fact that these evolve on different time-scales. The resulting set of interconnected differential equations decomposes a scene's velocities with the goal of minimizing prediction errors for subsequent observations. Beyond capturing human percepts in many psychophysics experiments qualitatively, the model explains human motion structure classification quantitatively with higher fidelity than a previous ideal observer-based model[17]. Furthermore, the model provides a normative explanation for the putative origin of human illusory motion perception, and yields testable predictions for future psychophysics experiments.

Finally, we address how motion structure discovery could be supported by neural circuits in the brain. Studying the neural representations underlying motion structure perception is challenging, as the perceived structure often has no direct physical counterpart in the environment (e.g., the concept of a flock velocity in Fig. 1a). We derive a recurrent neural network model that not only implements the proposed online hierarchical inference model, but shares many properties with motion-sensitive middle temporal area (MT)[21] and dorsal medial superior temporal area (MSTd)[19,32]. The network model in turn allows us to propose a class of stimuli for neuroscientific experiments that make concrete predictions for neural recordings.

Results

In what follows, we first present the online model for simultaneous hierarchical inference of instantaneous motion and of the scene's underlying structure. Next, we demonstrate the model's ability to explain human motion perception across a set of psychophysics experiments and discuss testable predictions for future studies. Finally, we propose a biologically realistic neural implementation of online hierarchical inference and identify targeted experiments to reveal neural representations of latent structure.

Online hierarchical inference in a generative model of structured motion

A structural understanding of the scene in Fig. 1a requires the observer to decompose observed velocities of objects or their features into what we call latent motion sources, s, that, together, compose the scene (Fig. 1b). These latent sources might or might not have a direct counterpart in the physical world. In Fig. 1b, for instance, each bird's velocity on the observer's retina can be decomposed into the observer's self-motion, s_{self}, the flock's motion, s_{shared}, plus a smaller, animal-specific component, s_{ind}. Here, flock motion is an abstract mental concept that is introduced to organize perception, but doesn't have an immediate physical correlate. A correct decomposition leads to motion sources that aid interpretation of the visual scene, and thus supports behaviors such as navigation, tracking, prediction and pursuit. Such decomposition requires knowledge of the scene's structure, like the presence of a flock and which birds it encompasses (Fig. 1c). Wrong structural assumptions might lead to faulty inference of motion sources, like wrongly attributing the flock's motion in the sky to self-motion. Thus, the challenge for an observer is to simultaneously infer motion sources and structure online from a stream of noisy and ambiguous visual information.

We formalized the intuition of structured motion in the generative model shown in Fig. 1d–g. The stochastic model, first

introduced in ref. 16 and formally defined in Supplementary Note 1, accommodates fundamental principles of physics (isotropy and inertia) and psychophysics (continuity of trajectories[33] and slow-velocity priors[9]), without making assumptions on specific object trajectories. For example, the motion of three flocking birds viewed by a stationary observer (motion tree in Fig. 1d) can be decomposed into four independent motion sources—one shared (magenta) and three individual (green, one per bird)—that evolve according to Ornstein–Uhlenbeck processes[34], generating smooth motion with changes typically occurring at time scale τ_s (Fig. 1e). The resulting speed (absolute velocity) distribution of each motion source is governed by an associated motion strength, λ, such that the expected speed is proportional to λ. The observable velocities, $\boldsymbol{v}_t$, are in turn noise-perturbed (noise magnitude σ_{obs}; Fig. 1g) sums of the individual motion sources (collected in vector $\boldsymbol{s}_t$), with the contribution of each individual motion source specified by a different column of the component matrix $\boldsymbol{C}$ (see Fig. 1f). This formalizes the intuition that observable velocities are the sum of their ancestral motion sources in the tree.

In this model, the structure of a scene is fully characterized by the vector of motion strengths, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m, \ldots, \lambda_M)$, which describe the presence ($\lambda_m > 0$) or absence ($\lambda_m = 0$) of motion components, as well as their typical speed. In other words, given a reservoir of components, $\boldsymbol{C}$, which might have been learned to occur in visual scenes in general, knowing $\boldsymbol{\lambda}$ is equivalent to knowing the motion structure of the scene. Inferring this structure in turn becomes equivalent to inferring the corresponding motion strengths.

An agent faces two challenges when performing inference in this generative model (Fig. 1h). First, inference needs to be performed on the fly (i.e., online) while sensory information arrives as an ongoing stream of noisy velocity observations. Second, how observed motion is separated into latent motion sources, $\boldsymbol{s}$, and motion structure, $\boldsymbol{\lambda}$, is inherently ambiguous, such that inference needs to resolve the hierarchical inter-dependence between these two factors. We address both challenges by recognizing that motion structure, $\boldsymbol{\lambda}$, typically changes more slowly than the often volatile values of motion sources, $\boldsymbol{s}$, facilitating the use of an online Expectation-Maximization (EM) algorithm to infer both. This separation of time scales yields a system of mutually dependent equations for updating $\boldsymbol{\lambda}$ and $\boldsymbol{s}$ and furthermore affords a memory-efficient, continuous-time online formulation that is amenable to a neural implementation (see Methods for an outline of the derivation, and Supplementary Note 2 for the full derivation). While the algorithm is approximate, it nonetheless performs adequate online hierarchical inference and closely resembles more accurate solutions, even for deeply nested motion structures (see Supplementary Fig. 1).

Our online model computes, at any time, a posterior belief over the latent motion sources, $\boldsymbol{s}_t$, which is Gaussian with mean vector $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$, as well as an estimate, $\boldsymbol{\lambda}_t$, of the underlying structure. The dynamics of $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, and $\boldsymbol{\lambda}_t^2$ (the inference is more elegantly formulated on the squared values) read:

$$\partial_t \boldsymbol{\lambda}_t^2 = -\frac{1}{\tau_\lambda}\boldsymbol{\lambda}_t^2 + \boldsymbol{\alpha} \odot \left(\boldsymbol{\mu}_t^2 + \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right) + \boldsymbol{\beta}, \tag{1}$$

$$\partial_t \boldsymbol{\mu}_t = -\frac{1}{\tau_s}\boldsymbol{\mu}_t + \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2) \odot \boldsymbol{C}^\top \boldsymbol{\epsilon}_t \text{ with } \boldsymbol{\epsilon}_t = \frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} - \frac{\boldsymbol{C}\boldsymbol{\mu}_t}{\sigma_{\text{obs}}^2}, \tag{2}$$

$$\text{and } \boldsymbol{\Sigma}_t = \text{diag}\left[\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right]. \tag{3}$$

The coupled Eqs. (1)–(3) support the following intuition. Equation (1) calculates a running average of the motion strengths $\boldsymbol{\lambda}_t^2$ by use of a low-pass filter with time scale τ_λ. Here, $\odot$ denotes element-wise multiplication and the function $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$ (Fig. 1i) estimates the variance of the s-posterior distribution according to an adiabatic approximation (cf. Eq. (3), see Methods). The constants $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contribute a sparsity-promoting prior, $p(\boldsymbol{\lambda}^2)$, for typical values of the motion strengths (see Methods for their full expressions). By Eq. (2), the motion source means $\boldsymbol{\mu}_t$ are estimated by a slightly different low-pass filter that relies on a prediction error, $\boldsymbol{\epsilon}_t$, between the model's expected velocities, $\boldsymbol{C}\boldsymbol{\mu}_t$, and those actually encountered in the input, $\boldsymbol{v}_t$ (both normalized by observation noise variance to facilitate the later network implementation). This prediction error on observable velocities is transformed back to the space of latent motion sources via the transposed component matrix $\boldsymbol{C}^\top$ and then, importantly, gated by element-wise multiplication ($\odot$) with the variance estimates $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$. This gating implements a credit assignment as to which motion source was the likely cause of observed mismatches in $\boldsymbol{\epsilon}_t$, and thus uses the scene's currently inferred motion structure to modulate the observed velocities' decomposition into motion sources. For flocking birds, for example, a simultaneous alignment in multiple birds' velocities would only be attributed to the shared flock velocity if such a flock had been detected in the past (λ_{shared} large, and λ_{ind} small). Otherwise it would be assigned to the birds' individual motions, s_{ind}.

Together, Eqs. (1) and (2) implement a coupled process of structure discovery and motion decomposition, which distinguishes them through different time-scales. Notably, the proposed model is not a heuristic, but is derived directly from a normative theory of online hierarchical inference. Next, we explored if the model can explain prominent phenomena of human visual motion perception.

Online inference replicates human perception of classical motion displays

To explore if the proposed online model can qualitatively replicate human perception of established motion displays, we simulated two classical experiments from Gunnar Johansson[25] and Karl Duncker[35]. These experiments belong to a class of visual stimuli which we refer to as object-indexed experiments (Fig. 2a) because the observed velocities, $\boldsymbol{v}_t$, belong to objects irrespective of their spatial locations. (A second class, which we refer to as location-indexed experiments, will be discussed below.)

In Johansson's experiment, three dots oscillate about the screen with two of the dots moving horizontally and the third dot moving diagonally between them (see Fig. 2b and Supplementary Movie 1). Humans perceive this stimulus as a shared horizontal oscillation of all three dots, plus a nested vertical oscillation of the central dot. Similar to previous offline algorithms[15], our online model identifies the presence of two motion components (Fig. 2c): a strong shared motion strength, λ_{shared} (magenta) and weaker individual motion, λ_{ind}, for the central dot (green). The individual strengths of the outer two dots (light and dark green), in contrast, decay to zero. Most motion sources within the structure are inferred to be small (dotted lines in Fig. 2d). Only two sources feature pronounced oscillations: the x-direction of the shared motion source, $\mu_{\text{shared},x}$, (magenta, solid line) and the y-direction of the central dot's individual source, $\mu_{\text{ind},y}$, (green, solid line), mirroring human perception. As observed velocities are noisy, they introduce noise in the inferred values of $\boldsymbol{\mu}_t$, which fluctuate around the smooth sine-functions of the original, noise-free stimulus. As expected from well-calibrated Bayesian inference, the magnitude of these fluctuations is correctly mirrored in the model's uncertainty, as illustrated by the posteriors' standard deviation $\sqrt{f_\Sigma(\lambda_t^2)}$ (shaded areas in Fig. 2d).

In the second experiment, known as the Duncker wheel, two dots follow the motion of a rolling wheel, one marking the hub, the other marking a point on the rim (Fig. 2e). The two dots describe an intricate trajectory pattern (see Fig. 2f and Supplementary Movie 2), that, despite its impoverished nature, creates the impression of a rolling object for human observers, a percept that has been replicated by offline algorithms[15]. Likewise, our online model identifies a shared (magenta in Fig. 2g) plus one individual (dark green) component, and decomposes the observed velocities into shared rightward motion

https://doi.org/10.1038/s41467-022-34805-5

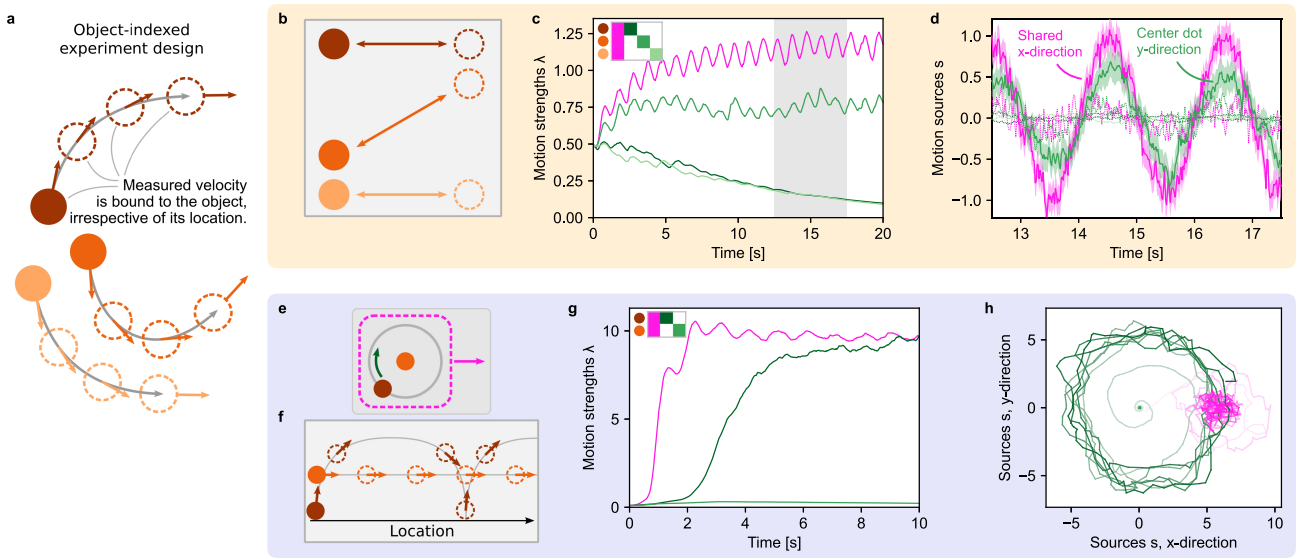


Fig. 2 | Online hierarchical inference replicates human perception of classical motion displays. a In object-indexed experiment designs, every observable velocity is bound to an object irrespective of its location. Many psychophysics studies fall into this class of experiment design. **b** Johansson's 3-dot motion display. Humans perceive the stimulus as shared horizontal motion with the central dot oscillating vertically between the outer dots. **c** The online model's estimate of the motion strengths, $\boldsymbol{\lambda}_t$ (a single motion strength is shared across both spatial dimensions). The component matrix, $\boldsymbol{C}$, is shown in the top-left as a legend for the line colors. Circles next to the matrix show the assignment of the rows in $\boldsymbol{C}$ to the dots in panel **b**. **d** The model's posterior distribution over the motion sources, $\boldsymbol{s}_t$, during the gray-shaded period in panel **c**. Shown are the mean values, $\boldsymbol{\mu}_t$, as lines along with the model's estimated standard deviation (shaded, only for two components for visual clarity). **e** The Duncker wheel resembles a rolling wheel of which only the hub and one dot on the rim are visible. **f** Despite its minimalist trajectory pattern, humans perceive a rolling wheel. **g** Inferred motion strengths, $\boldsymbol{\lambda}_t$. The model identifies shared motion plus an individual component for the revolving dot. **h** Inferred motion sources, $\boldsymbol{\mu}_t$, for the duration in panel **g**. Color gradients along the lines indicate time (from low to high contrast). For visual clarity, $\boldsymbol{\mu}_t$ has been smoothed with a 50 ms box filter for plotting. Source data are provided as a Source Data file.

plus rotational motion for the dot on the rim (see Fig. 2h). Notably, the shared motion component is discovered before the revolving dot's individual motion, leading to a transient oscillation in the inferred shared motion source, μ_{shared} (see light magenta trace in Fig. 2h) — an onset effect that could be tested experimentally.

In summary, the online hierarchical inference model successfully identified the structure underlying the motion displays, provided Bayesian certainty estimates for the inferred motion, and replicated human perception in these classical psychophysics experiments.

Online inference outperforms ideal observers in explaining human structure perception

Having qualitatively replicated motion structure inference in common motion displays, we next asked if our online model could quantitatively explain human motion structure perception. To address this question, we reevaluated behavioral data from Yang et al.[17], where participants had to categorize the latent structure of short motion displays (see Fig. 3a). Motion scenes followed one of four structures (Fig. 3b) and were generated stochastically from the same generative model underlying our hierarchical inference model. Owing to their stochastic generation, scenes often were ambiguous with regard to their latent structure, prompting distinct error patterns in human responses (see confusion matrix in Fig. 3c). For instance, independently moving dots were more frequently misclassified as clustered motion (I-C element) than vice versa (C-I element), global motion was highly recognizable, and nested hierarchical motion was more frequently misperceived as clustered than as global.

To test if human responses arise from normative, Bayesian motion structure inference, Yang et al. modeled these responses in two steps (blue branch in Fig. 3d): first, an offline Bayesian ideal observer, which was provided with the trajectories of all objects within a trial, calculated the likelihood for each of the four structures. Then, these four probabilities were fed into a choice model with a small set of participant-specific fitting parameters (see Methods). This model captured many aspects of human responses, including task performance, typical error patterns, single-trial responses, and participant-specific differences. Yet, the model arrived at these probabilities by comparing the likelihoods of the full sequences for all four candidate structures, and so had no notion of how a percept of structure could emerge over the course of the trial.

Thus, we next asked if our online model, which gradually infers the structure during the stimulus presentation, was better able to account for the observed response pattern. As our model by design inferred real-valued motion strengths $\boldsymbol{\lambda}$ rather than only discriminating between the four structures used in the experiment, we added an additional stage that turned the inferred motion strengths into a likelihood for each of the four structures at trial end (red branch in Fig. 3d, see Methods). To do so, we computed five hand-designed features from the seven-dimensional vector $\boldsymbol{\lambda}_t$ (besides one global and three individual strengths, there are three possible two-dot clusters), and trained a multinomial logistic regression classifier on the features to obtain likelihood values for each of the structures. The classifier was trained on the true structures of the trials, and thus contained no information about human responses. Finally, we fitted the same choice model as Yang et al. to the participants' responses.

The confusion matrix predicted by our model shows an excellent agreement with human choices, both when averaged across participants (Fig. 3e), and on a per-participant basis (see Supplementary Figs. 3 and 4). Indeed, our model beats the original computational model in terms of response log-likelihoods for all of the 12 participants (see Fig. 3f; $p < 0.001$, two-sided paired Wilcoxon signed-rank test). Furthermore, the online model overcomes the systematic underestimation of global motion (G-G matrix element) that previous, ideal observer-based approaches suffered from[16,17]. Importantly, in our model, any information connecting the stimulus to the eventual choice is conveyed through the motion strengths, $\boldsymbol{\lambda}_t$, as a bottleneck. The fact that the online hierarchical inference-based approach describes human responses better than the ideal observer-based model of Yang

https://doi.org/10.1038/s41467-022-34805-5

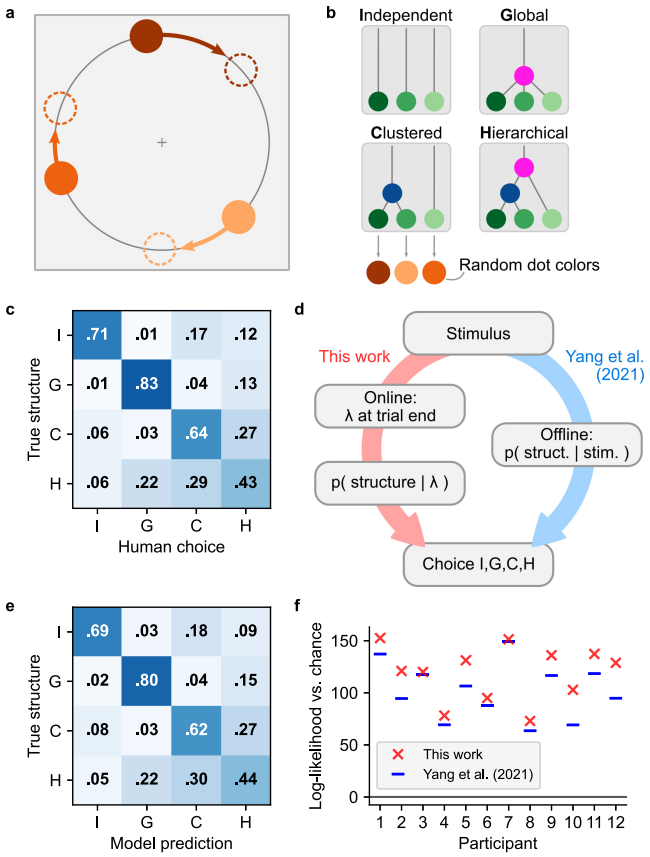


Fig. 3 | The model quantitatively explains human perception of nested and ambiguous motion scenes. a Stochastic motion stimulus from Yang et al.[17] consisting of three dots rotating on a circle. **b** Each trial followed one of four motion structures. If clustered motion was present (C or H structure), any pair of dots could form the cluster. **c** Confusion matrix of human responses, averaged over all 12 participants. **d** Models for predicting human responses. Yang et al. employed a Bayesian ideal observer as the basis for fitting a participant-specific choice model. Our model, in contrast, calculates the likelihood for each structure from the motion strengths, $\boldsymbol{\lambda}_t$, at trial end and then fits the same choice model as Yang et al. for translating probabilities into human responses. **e** Confusion matrix of our model. **f** Log-likelihood of human responses relative to chance level, for both models. The analyses in panels **e** and **f** are leave-one-out cross-validated to prevent overfitting. Source data are provided as a Source Data file.

et al. indicates that our model may share mechanistic features with the human perceptual apparatus.

Explaining motion illusions that rely on spatial receptive fields

In contrast to the object-indexed experiments discussed above, another class of psychophysics experiments employs velocity stimuli that remain at stationary locations (see Fig. 4a), typically in the form of apertures of moving dots or drifting Gabors. This class, which we refer to as location-indexed experiments, is furthermore popular in neuroscience as it keeps the stimulus' local visual flow within an individual neuron's spatial receptive field throughout the trial[21]. We investigated our model's ability to explain illusory motion perception in two different types of location-indexed experiments: motion direction repulsion in random-dot kinematograms (RDKs)[36–41], see Fig. 4, and noise-dependent motion integration of spatially distributed stimuli[42,43], see Fig. 5.

We modeled perception in these experiments by including a self-motion component and added a vestibular input signal to the observables (see Fig. 4b, and cf. Fig. 1a–c). The vestibular input, which we fixed to have zero mean plus observation noise, complemented the visual input, which is ambiguous with regard to self-motion and globally shared object motion and can induce illusory self-motion ("vection")[44–46]. In turn, we model the subjectively perceived velocity of objects, relative to the stationary environment, as the sum of all inferred motion sources excluding self-motion (see Fig. 4c and Methods).

In the RDK experiment, a participant fixates the center of an aperture in which two groups of randomly positioned dots move linearly with opening angle γ (see Fig. 4d) and subsequently reports the perceived opening angle. Motion direction repulsion occurs if the perceived angle is systematically biased relative to the true opening angle.

As previously reported, the repulsion bias can change from an under-estimation of the opening angle for small angles to an over-estimation for large angles (data from ref. 36 reprinted as black dots in Fig. 4e). We replicated this effect by simulating two constant dot velocities with opening angles that varied across trials. Our model decomposed the stimulus into self-motion, shared motion and individual (group) motion. Across opening angles, it featured a triphasic psychometric function with angles smaller than ~40° being underestimated, angles between ~40° and ~110° being over-estimated, and even larger angels being unbiased (purple curve in Fig. 4e). The match with human biases arose without systematic tuning of simulation parameters (the simulations presented in this manuscript were mostly performed with a set of default parameters, see Methods). Inspecting the model's inferred motion components revealed that, for small γ, the negative bias arose from integrating all dots into a single, coherent motion component while disregarding individual dot motions (left inset in Fig. 4e). Intermediate γ, in contrast, caused the shared component to be correctly broken up into two individual components—plus a small illusory self-motion component (right inset in Fig. 4e). This self-motion, which is ignored in the perceived velocities, widened the perceived opening angle between the two groups of dots. For even larger γ, the illusory self-motion vanished yielding unbiased percepts.

For fixed opening angles, motion direction repulsion is furthermore modulated by relative contrast and speed difference between the two motion components. Specifically, for an opening angle of $\gamma = 45°$, Chen et al.[37] have shown that increasing the contrast of one dot group inflates the perceived opening angle—here measured relative to horizontal to separate cause and effect—of the other, constant-contrast group (Fig. 4f, left). We replicated this effect in simulations that operationalized visual contrast as an (inverse) multiplicative factor on the observation noise variance, σ^2_{obs}. For an opening angle of $\gamma = 45°$, our model featured a positive and monotonically increasing repulsion bias as the second group's contrast increases (purple line in Fig. 4f, right), similar to what has been previously reported. For smaller opening angles, in contrast, our model predicts an inversion of the repulsion bias, which first decreases at low contrast and then increases again for higher contrast (blue line in Fig. 4f, right)—a prediction that remains to be tested. Increasing the speed of one motion component for large opening angles also introduces a positive bias in the perceived opening angle of the other component in human participants[36,38]. We replicated this effect by increasing the second group's speed, which, for a $\gamma = 90°$ opening angle, yielded a relatively stable bias of ~5° across different motion speeds (dashed line in Fig. 4g), in line with the aforementioned experimental data from Braddick et al.[36] and, for a $\gamma = 60°$ opening angle (purple line in Fig. 4g), qualitatively replicated the initial rise and then gradual decline in the bias, as reported for this opening angle by Benton and Curran[38]. Furthermore, our model predicts that the speed-dependent bias changes to a biphasic curve for smaller opening angles (blue line), providing another testable prediction.

Extending the basic MDR experiment from Fig. 4d, Takemura et al.[39] investigated how motion in a surrounding annulus affects the perceived directions of inner RDKs, see sketch in the top left of Fig. 4h. Two inner RDKs move to the left and right, respectively, while two

https://doi.org/10.1038/s41467-022-34805-5

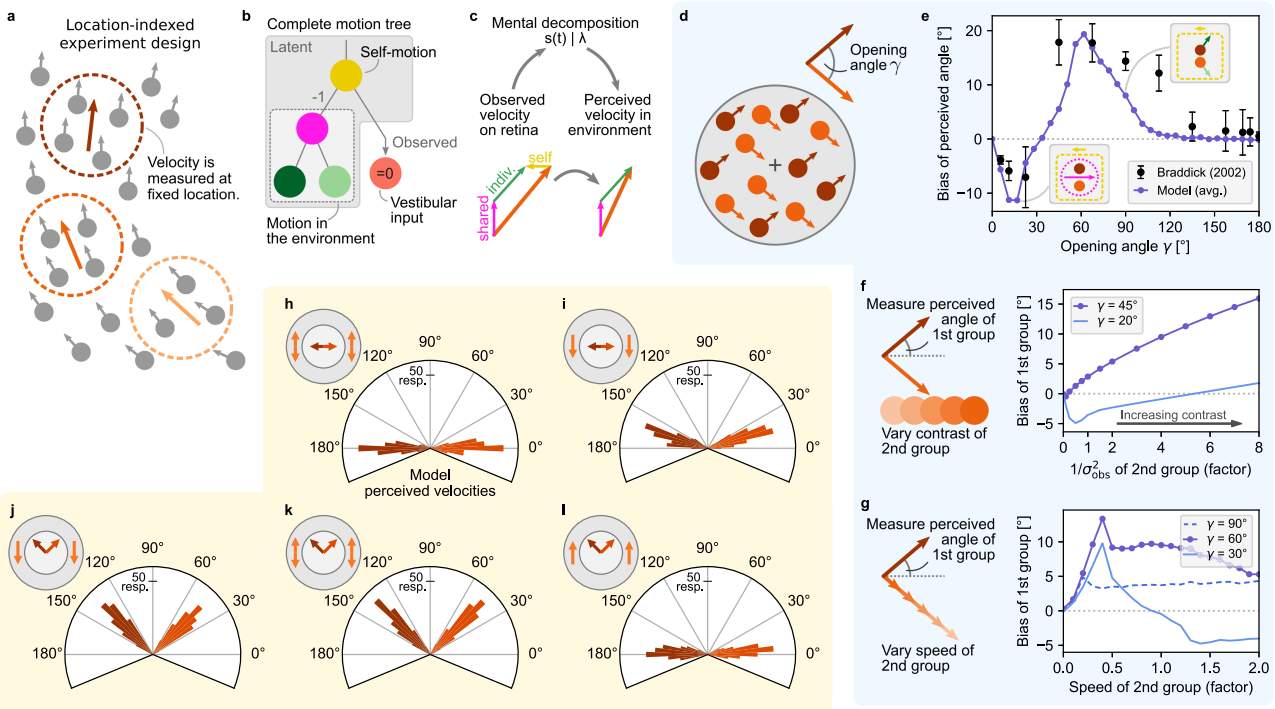


Fig. 4 | Hierarchical inference explains motion illusions in location-indexed experiments. a In location-indexed experiments, motion flow is presented at stationary spatial locations. **b** Considered latent motion components. Self-motion, which affects all retinal velocities in the opposite direction (−1) integrates both visual input and a vestibular signal (here: zero + noise). **c** Perceived object velocities, relative to the environment, are the sum of all inferred motion components excluding self-motion. **d** In motion direction repulsion experiments, two groups of dots move at constant velocity with opening angle γ. **e** The direction in which human perception of the opening angle is biased depends on the true opening angle. Black dots: human data, reproduced from ref. 36, error bars denote S.E. of the mean across subjects; $n = 3$ subjects, 80 trials per angle and subject. Purple line: model percept. Insets: the model's inferred motion decomposition. **f** Varying the contrast of one dot group modulates the biased percept of the angle of the other group. Purple: model percept for $\gamma = 45°$, qualitatively matching data from ref. 37. Blue: predicted inversion of the bias for smaller opening angles. **g** Same as panel **f**, but for varying the speed of the second group. Purple: model percept for $\gamma = 60°$, qualitatively matching data from ref. 38. Dashed blue: model percept for $\gamma = 90°$, qualitatively matching data from ref. 36. Solid blue: predicted biphasic function for smaller opening angles. **h**–**l** Extended experiment from ref. 39 which surrounds the two central RDKs with additional RDKs in an annulus. The hierarchical inference model replicates human perception in various conditions. **h** A surround with dots moving vertically both up- and downwards ("bi-directional surround" in ref. 39, indicated by orange arrows in the top-left sketch's annulus) causes no repulsion in the perceived directions of horizontally moving RDKs in the center (darker orange arrows in the top-left sketch's center). Our model replicates this perception as shown in the histogram of 200 trial repetitions. **i** Coherently moving annulus RDKs cause the perceived inner velocities to be biased away from the surround direction. **j** For diagonally moving inner RDKs, the same coherent downward surround has no noticeable effect. **k** Neither does a bi-directional surround bias the percept of diagonally moving inner RDKs. **l** An upward surround, in contrast, biases the percept of the inner RDKs to close-to-horizontal motion. Source data are provided as a Source Data file.

additional RDKs in the annulus move up and down, respectively. For this stimulus human observers show no direction repulsion[39]. We simulated this extended MDR experiment with our hierarchical inference model by extending the motion tree of Fig. 4b to include two group components for the outer and inner RDKs, respectively, on the third level, and four individual components (one per RDK) as leaves, on the forth level (cf. Supplementary Fig. 5). Across 200 simulated trials (see Methods), the distribution of inner RDK directions perceived by the model at trial end (see histogram in Fig. 4h) match the reported unbiased perception of humans.

Our model was further able to replicate human perceptual biases for various other combinations of dot motion in the inner and surrounding RDKs explored by Takemura et al. (see Fig. 4i–l, and Supplementary Fig. 5 for example trials). The percepts to all combinations are qualitatively replicated by our model. When both surrounding RDKs move downward, as shown in Fig. 4i, the perceived motion of the inner RDKs is slightly biased upward. The reason for the bias in the model's percept is a small illusory self-motion component in upward direction which necessitates a slight diagonal upward tilt of the inner RDKs' individual motions for explaining their horizontal retinal velocities. When modifying the stimulus such that the inner RDKs move diagonally with a 90 degree opening angle (see Fig. 4j–l), human and model percepts remain unbiased in the case of downward (Fig. 4j) and bi-directional surrounding motion (Fig. 4k). In both cases, the directional contrast of the presented velocities obviates the illusory identification of self-motion, thereby implicating unbiased percepts of the model. If, however, the surrounding RDKs move upwards, strong direction repulsion on the inner dots was reported[39] leading to their perceived motion to become almost horizontal (Fig. 4l). In the model, this effect originates from illusory downward self-motion arising from the general alignment of the presented velocities. Overall, our hierarchical inference model replicated biased and unbiased perception across a variety of stimulus conditions.

Turning to noise-dependent motion integration of spatially distributed stimuli, we investigated a motion illusion by Lorenceau[42] which has received little attention in the literature (see Fig. 5). Two groups of dots oscillate in vertical and horizontal orientation, respectively (see Fig. 5a and Supplementary Movie 3). Both groups follow sine-waves with identical amplitude and frequency, but maintain a relative phase shift of $\pi/2$ that is consistent with an imaginary global clockwise (CW) rotation (indicated by a gray arrow in Fig. 5a). This stimulus can be considered to be location-indexed, as the small oscillation amplitude of less than 1 degree of visual angle caused the stimulus to conveniently fit into the receptive fields of individual neurons of the human homolog of area MT[47]. Interestingly, the stimulus' percept changes once disturbances orthogonal to the axes of

https://doi.org/10.1038/s41467-022-34805-5

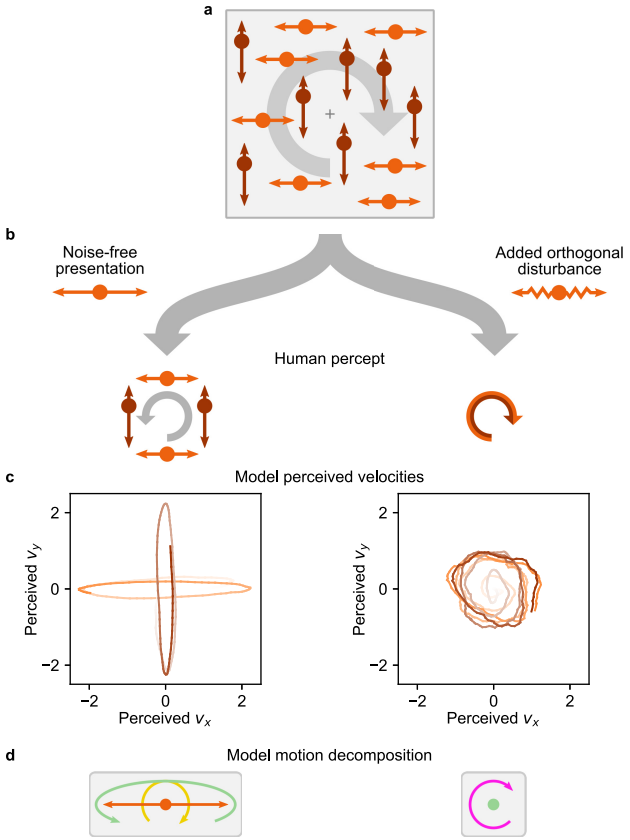


Fig. 5 | Noise-dependent perceptual changes for motion integration of spatially distributed stimuli. a In the motion illusion from Lorenceau[42], a vertically and a horizontally oscillating group of dots maintain a 90°-phase shift consistent with global clockwise rotation (indicated as gray arrow). **b** The noise-free stimulus (left branch) evokes transparent motion with an additional counter-clockwise rotating percept in human observers. Adding motion noise by disturbing dot trajectories orthogonally to their group's oscillation axis (right branch; modeled by increased observation noise σ^2_{obs}) flips the percept to a single coherent rotation of all dots in clockwise direction. **c** The model's perceived velocities in both stimulus conditions (time = color gradient from low to high contrast; $t \le 2$ s in noise-free condition; $t \le 5$ s in noisy condition). For visual clarity, perceived velocities have been smoothed with a 200 ms box filter for plotting. **d** Illustration of the model's inferred motion decomposition. For noise-free stimuli, clockwise rotating self-motion is compensated by counter-clockwise rotating group motion (sketched here for the horizontal group). With motion noise, only a single, clockwise rotating shared motion component is inferred for all dots. Source data are provided as a Source Data file.

oscillation are added (called "motion noise" in ref. 42, see Fig. 5b). Without motion noise, participants perceive transparent motion, that is, the dots within either group are combined to a rigidly moving object according to common fate, and both groups are perceived as moving separately. Their movement, however, is not perceived as strictly vertically and horizontally, but rather the stimulus induces an impression of slight counter-clockwise (CCW) rotation, that is, "opposite to veridical"[42]. With motion noise, in contrast, the percept switches in two ways: all dots appear to move coherently along a circle, and the perceived direction of movement becomes CW. These percepts are illustrated in Fig. 5b.

Applied to this stimulus, our model replicates the perceived rotation direction reversal with increased motion noise, which we simulated through an increase in the observation noise σ^2_{obs}. Specifically, the model's perceived velocities for both groups of dots featured a slight global CCW rotation on top of two generally separated groups for the noise-free stimulus, and a single global CW rotation once observation noise is increased (Fig. 5c). Inspecting the model's motion decomposition provides a possible answer to how this flip in perceived rotation emerges, which is illustrated in Fig. 5d by the example of the horizontal group. On noise-free presentation, dot motion was decomposed into clockwise rotating self-motion (golden arrow) plus a horizontally elongated, yet slightly CCW rotating group motion (green arrow), leading to the transparent CCW motion percept. Once observation noise increased, the inferred motion structure discarded the separated groups in favor of a single global motion component (magenta), leading to the percept of coherent CW rotation for all dots (see Supplementary Fig. 6 for trajectories of the motion strengths and sources under both conditions).

Object recognition and perceptual switching of nested structure-from-motion displays

Motion relations do not only aid dynamic tasks, such as tracking and prediction, but also provide essential cues for object recognition. Structure-from-motion (SfM), the perception of 3D objects from 2D visual displays, is well-studied in psychology[48–54] and neuroscience[55–58]. We asked whether our model can support SfM perception and replicate the salient phenomenon of perceptual switching when presented with ambiguous stimuli (see Fig. 6a). Furthermore, using the model, we identified SfM displays of nested objects which could inspire future psychophysics experiments studying how structure interacts with perceptual ambiguity.

Typical SfM displays, like the point cloud-cylinder in Fig. 6a and Supplementary Movie 4, involve rotational motion in three dimensions, contrasting with the translational motion in two dimensions considered so far. Our generative model supports such 3D rotation in location-indexed experiments: as illustrated in Fig. 6b, introducing a rotational motion source, s^{rot}, which describes the cylinder's angular velocity around the y-axis, yields a linear dependence of the observed retinal velocities on s^{rot} at every input location (dashed orange circles) owing to the locations' fixed coordinates. Thus, rotational motion is supported naturally by the component matrix, $\boldsymbol{C}$, (cf. Fig. 1e) and integrates without any changes into our hierarchical inference model.

Ambiguous SfM displays, such as the considered frontal view of a rotating cylinder, furthermore feature equivocal correspondences of spatially overlapping inputs to the cylinder's surface at the front and back. Mentally assigning the overlapping left- and rightward retinal velocities to their depth locations is key to forming a coherent percept of the 3D object. To support such percepts in our model, we added a basic assignment process: spatially overlapping velocities are assigned to their depth location on the cylinder (front or back) such that the assignment locally minimizes the model's prediction error, $\boldsymbol{\epsilon}_t$, in Eq. (2). Furthermore, this assignment is independently re-evaluated at each input location with a uniform probability in time (see Methods). We tested the model's ability to perceive SfM by using a motion tree with self-motion, rotational motion and individual motion (see Fig. 6c, and Supplementary Fig. 7 for a control simulation with more motion components). As shown in Fig. 6d, the model swiftly identifies rotational motion across all input locations at a constant angular speed, matching the human percept of a rotating cylinder. Subsequently, the percept switches randomly and abruptly between CW and CCW rotation, with inter-switch-intervals following a Gamma distribution (see Fig. 6e). The resulting stochastically switching percepts with typical durations of a few seconds match the reported bistable perception of humans[53,57].

To explore how more complex structures could interact with SfM perception, we asked how our model interprets the rotation of nested point-cloud cylinders (see Fig. 6f and Supplementary Movie 4). Their rotation is easily identified by humans[49], and features a more complex structure than basic SfM displays that only require a single rotational motion source. To present this stimulus to the model, the extended graph in Fig. 6g features rotational sources not only for the inner and outer cylinders (light and dark

https://doi.org/10.1038/s41467-022-34805-5

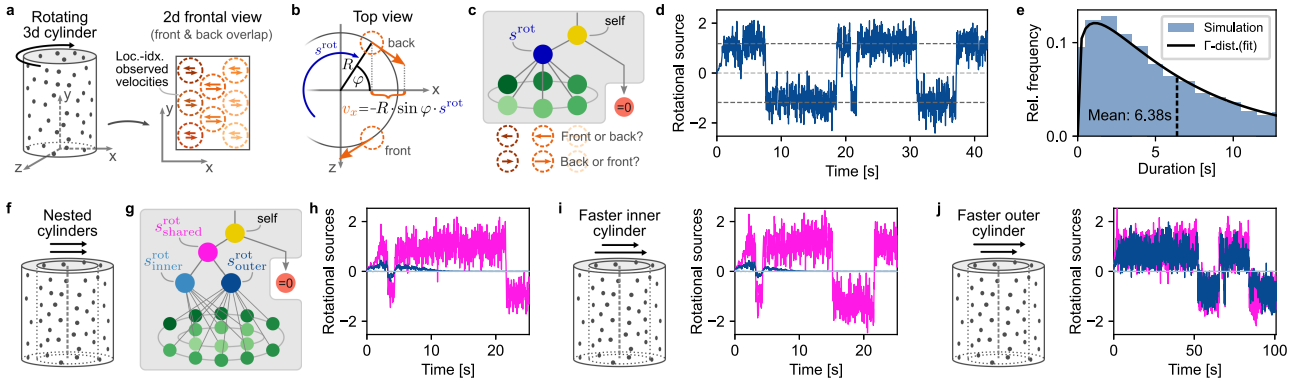


Fig. 6 | Object recognition and perceptual switching of nested structure-from-motion (SfM) displays. a Cylindrical SfM stimulus. A random point cloud on the surface of a rotating, transparent cylinder (left) supports two possible percepts when viewed from the front without depth information (right). Humans perceive the structured motion of this 2D projection as a rotating 3D cylinder, albeit with bistable direction of the perceived rotation. **b** Top view illustration of how the generative model supports rotational motion. The rotational motion source, s_t^{rot}, describes angular velocity about the vertical axis ($s^{\text{rot}} > 0$ for CCW rotation, by definition). In location-indexed experiments, observed velocities, $\boldsymbol{v}_t$, at a (fixed) location with angle φ and radius R are a linear function of the rotational motion source. In the frontal view of SfM experiments, only the x-component, $v_x = -R\sin(\varphi)\, s_t^{\text{rot}}$, and the vertical y-component, $v_y = 0$, are visible. **c** Motion tree and correspondence problem. The graph contains self-motion, rotational motion of the entire cylinder, and individual motion for every location. For any x-y coordinate, there exist two overlapping observed velocities which are ambiguous regarding their depth position (front or back). We performed the assignment of observations to their perceived depth (front or back) such that the prediction error, $\boldsymbol{\epsilon}_t$, in Eq. (2) is minimized. **d** 3D percept and perceptual bistability. Like humans, the model identifies rotation as the single motion component. The value of s_t^{rot} switches randomly between CW and CCW rotation with constant angular speed. **e** Distribution of perceptual switches. The distribution of duration-of-percepts closely follows a Gamma distribution, as commonly reported in human psychophysics. **f** Extension of the SfM display adding a smaller point cloud-cylinder, nested within the original cylinder. **g** Motion tree for the extended experiment. Three rotational components are provided: shared rotation of both cylinders, rotation of the outer cylinder, and rotation of the inner cylinder. The correspondence problem now demands assigning 4 observations where both cylinders overlap. **h** Perceived structure for identical angular speed of both cylinders. The model infers a single shared rotational component. **i** Fast inner cylinder. When increasing the angular speed of the inner cylinder by 50% (sketch on the left), the inferred structure is unaffected (right): the cylinders are perceived as having the same angular velocity. **j** Fast outer cylinder. In contrast, when increasing the angular speed of the outer cylinder by 50% (left), the cylinders' speeds are perceived as separated (right). For visual clarity, the trees in panels c and g show only 5 and 3 receptive field locations for the outer and inner cylinder, respectively, while for the simulations, we used 7 and 5 locations. Source data are provided as a Source Data file.

blue, respectively), but also the possibility of shared motion (magenta) affecting both cylinders. Where both cylinders overlapped, the assignment now minimized the prediction error over 4 overlapping retinal velocities (24 possible combinations per location), but remained otherwise unchanged. When both cylinders rotated with the same angular velocity of 90°/s, the model inferred a single shared rotational component (see Fig. 6h) leading to the impression of rigid rotation in which perceptual switches occur simultaneously for both cylinders. Identifying a structure with a single component rather than separate rotations for both cylinders is the result of the model's preference of simple structures. Increasing the angular velocity of the inner cylinder by 50% to 135 °/s (see Fig. 6i) did not change the model's percept of a rigidly shared rotation, but led to a slightly higher perceived speed of rotation. Inspecting the inference process revealed that the assignment process often assigned fast-moving dots of the inner cylinder to the outer cylinder and, vice versa, slower moving outer dots to the inner cylinder. This assignment yielded a sufficiently coherent interpretation of all retinal velocities as originating from a single rotation (within the bounds of perceptual acuity, σ_{obs}) for the model to prefer the simpler structure. Finally, a display in which the outer cylinder rotates faster than the inner cylinder (135 °/s and 90 °/s, respectively; see Fig. 6j) changed the model's inferred structure to perceiving different rotational speeds for both cylinders. Yet, even though each cylinder had its distinct perceived rotation, their rotational directions remained aligned and perceptual switches still occurred simultaneously, a perceptual linkage known from related experiments[54].

The nested SfM displays in Fig. 6f–j provide testable predictions for future psychophysics studies (see Supplementary Movie 4 for a video of all conditions). The model's percepts across all conditions matched the percept of the authors.

Experimental predictions from a biological network model of hierarchical inference

Finally, we asked whether and how a biologically plausible neural network could implement our online hierarchical inference model. To this end, we devised a recurrent neural network model of rate-based neurons. Naturally, such modeling attempt relies on many assumptions. Nonetheless, we were able to identify several experimentally testable predictions that could help guide future neuroscientific experiments.

Following Beck et al.[59], we assumed that task-relevant variables can be decoded linearly from neural activity ("linear population code") to support brain-internal readouts for further processing, actions and decision making. Furthermore, we employed a standard model for the dynamics of firing rates, $r_i(t)$, and assumed that neurons can perform linear and quadratic integration[59–62]:

$$\tau_i\, \partial_t r_i = -r_i + f_i(\boldsymbol{w}_i^{\top}\boldsymbol{r} + \boldsymbol{r}^{\top}\boldsymbol{Q}^{(i)}\boldsymbol{r}), \tag{4}$$

with time constant τ_i, activation function $f_i(\cdot)$, weight vector $\boldsymbol{w}_i$ and matrix $\boldsymbol{Q}^{(i)}$ for linear and quadratic integration, respectively. The rate vector, $\boldsymbol{r}(t)$, here comprises all presynaptic firing rates, including both input and recurrent populations. With these assumptions, we derived a network model with the architecture shown in Fig. 7a, which implements the online model, given by Eq. (1)–(3), via its recurrent interactions and supports linear readout of all task-relevant variables. That is, for every task-relevant variable, x, there exists a vector, $\boldsymbol{a}_x$, such that $x = \boldsymbol{a}_x^{\top}\boldsymbol{r}$ (see Supplementary Note 4 for the derivation).

The network consists of three populations. The input population (bottom in Fig. 7a) encodes the observed velocities, $\boldsymbol{v}_t/\boldsymbol{\sigma}^2_{\text{obs}}$, and observation precision, $1/\boldsymbol{\sigma}^2_{\text{obs}}$, in a distributed code. While any code that supports linear readout of these variables could serve as valid neural input, we chose a specific model that, as shown below,

https://doi.org/10.1038/s41467-022-34805-5

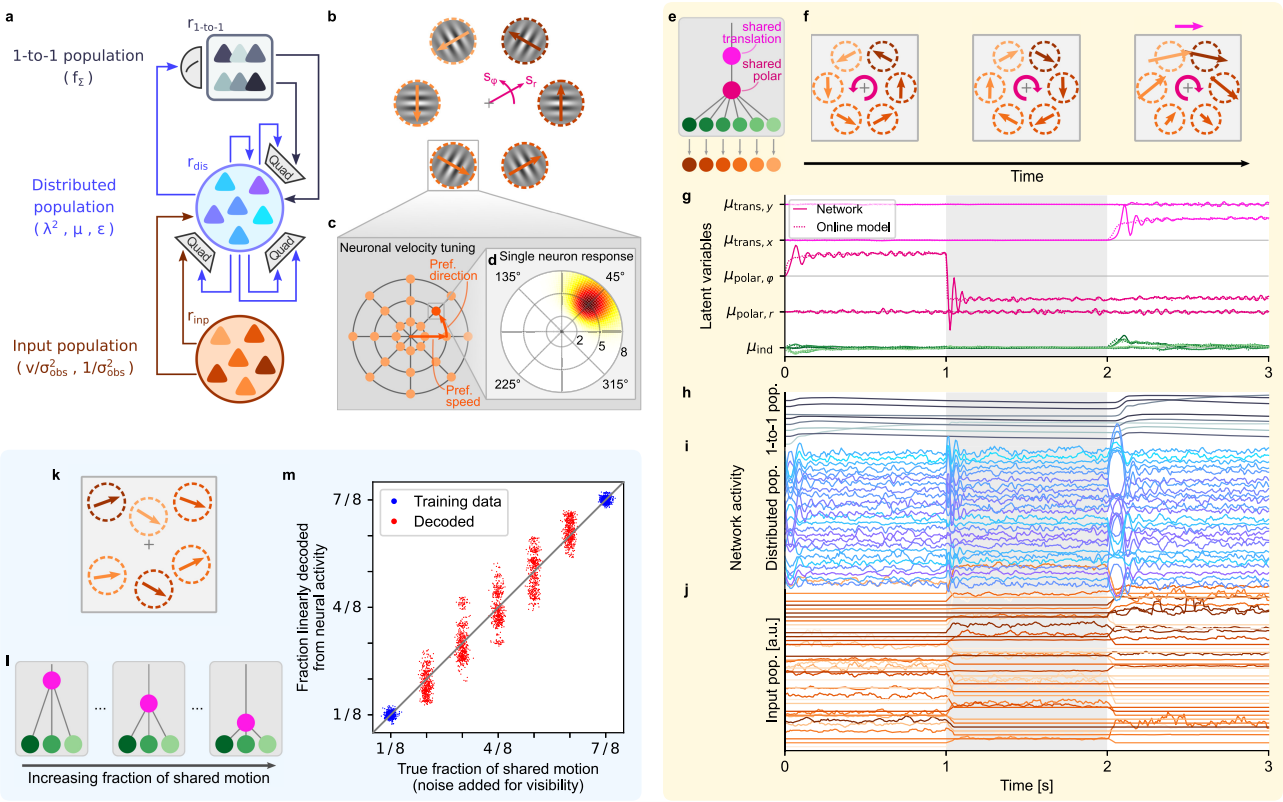


Fig. 7 | Hierarchical inference can be performed by a biologically realistic network model. a Network model implementing the online hierarchical inference model. Linear and quadratic interactions are indicated by direct arrows and Quad boxes, respectively. In parentheses, the variables represented by each population. **b** Rotational stimulus in a location-indexed experiment. Besides translational (Cartesian) motion, the model also supports rotational, s_φ, and radial motion, s_r. **c** Tuning centers in a model of area MT. A local population of neurons, which share the spatial receptive field highlighted in panel **b**, cover all directions and speeds with their velocity tuning centers. **d** Response function for the neuron highlighted in panel **c**. The neuron responds strongly to local velocities into the upper-right direction with a speed of ~5°/sec. Max. rate = 29.5 spikes/s. **e** Motion structure used for the network simulation in panels **f**–**j**, including simultaneous translational, rotational and radial motion sources. **f** Illustration of the stimulus. After 1*s* of counter-clockwise rotation around the fixation cross, the rotation switches to clockwise. At $t = 2$ s, rightward translation is superimposed on the rotation. **g** Motion sources inferred by the network (solid lines: distributed population readout; dotted lines: solution by the online model given by Eqs. (1)–(3)). Shown is μ_t for translational, rotational, radial and individual motion. Only 4 individual components (2 x- and 2 y-directions) are shown for visual clarity. **h** Firing rates of the 1-to-1 population. Rates are in arbitrary units (a.u.) because the theory supports scaling of firing rates with arbitrary factors. **i** Same as panel **h**, but for a random subset of 25 neurons of the distributed population. **j** Same as panel **h**, but for a random subset of 40 neurons of the input population, and smoothed with a 50 ms box filter for plotting. **k** Stimulus of a proposed neuroscience experiment. Velocities in distributed apertures follow the generative model from Fig. 1 using shared motion and individual motion. **l** Different trials feature different relative strengths of shared and individual motion, ranging from close-to-independent motion (left) to highly correlated motion (right). **m** Linear readout of the fraction of shared motion from neural activity. Seven different fractions of shared motion were presented (x-axis; noise in x-direction added for plotting, only). A linear regression model was trained on the outermost conditions (blue dots). Intermediate conditions were decoded from the network using the trained readout (red dots). Only a subset of $7 \times 500 = 3500$ points is shown for visual clarity. Source data are provided as a Source Data file.

captures many properties of motion-sensitive area MT. The distributed population (center in Fig. 7a) simultaneously represents the squared motion strengths, $\boldsymbol{\lambda}_t^2$, mean of the sources, $\boldsymbol{\mu}_t$, and prediction errors, $\boldsymbol{\epsilon}_t$, in a distributed code with linear readout. For those, almost arbitrary readouts suffice, such that we chose randomly generated readout vectors, $\boldsymbol{a}$. Notably, we propose the prediction errors, $\boldsymbol{\epsilon}_t$, to be linearly decodable, which allowed Eq. (2) to be implemented with the neuron model in Eq. (4) (see Supplementary Note 4, Sections 3 and 4). All neurons in the distributed population have simple activation functions, $f_i(\cdot)$, that are linear around some baseline activity. The linear decodability of $\boldsymbol{\lambda}_t^2$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\epsilon}_t$ are testable predictions. Finally, the 1-to-1 population (top in Fig. 7a) represents the uncertainty, $\boldsymbol{\Sigma} = \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}^2)$, in a one-to-one mapping, $r_m \propto f_\Sigma(\lambda_m^2)$, with r_m being the firing rate of either a single cell or, more likely, a small population. The theoretical motivation behind this representation is twofold: on the one hand, the non-linear form of $f_\Sigma(\cdot)$ prevents a distributed, linearly decodable representation (see Supplementary Note 4, Section 5); on the other hand, the particular shape of $f_\Sigma(\lambda_m^2)$, shown in Fig. 1i, mirrors the typical activation function of Type-I neurons[63], such that the proposed representation emerges naturally for the activation function, $f_\Sigma(\boldsymbol{a}_{\lambda_m^2}^\top \boldsymbol{r})$, in the 1-to-1 population (using the fact that λ_m^2 can be read out neurally with weights $\boldsymbol{w} = \boldsymbol{a}_{\lambda_m^2}$). Overall, the network structure predicts $\boldsymbol{\lambda}_t^2$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\epsilon}_t$ to be linearly decodable, and the components of $\boldsymbol{f}_\Sigma$ to be independently encoded in single neurons or small neural populations.

Even though the network model supports both the object-indexed and location-indexed experiments from Figs. 2–6, the retinotopic organization of the early visual system[21,64] brings a location-indexed perspective closer in line with our understanding of how the cortex encodes visual information. Furthermore, as we show in Supplementary Note 1, Section 5, our model can be extended to support motion sources in polar coordinates (see Fig. 7b), such that it supports salient real-world retinal input motifs, such as rotation and radial expansion/contraction about the fovea. (Note that rotation and expansion on the retina are conceptually distinct from the cylindrical rotation, s^{rot}, in structure-from-motion, discussed earlier.) Representations of angular motion, s_φ, and radial motion, s_r, can also coexist with translational motion (i.e., linear motion in Cartesian coordinates)

within the same population. Selective neural response to rotation, expansion/contraction and translation, as well as combinations thereof, such as spiraling, has been frequently reported in the dorsal medial superior temporal area (MSTd)[19,65].

Before demonstrating this capability in simulations, let us provide further information about the model's input population, and how it relates to known properties of area MT. To do so, consider the location-indexed stimulus in Fig. 7b. During fixation, each aperture stimulates a population in retinotopically organized, motion sensitive area MT[21]. Neurons in MT are tuned to respond preferentially to a certain direction and speed (Fig. 7c), such that the full population jointly covers all velocities in a polar grid[66,67]. The response of individual neurons to velocities within their spatial receptive field is commonly modeled by a log-normal function for speed[67] and a von Mises function for direction[68], leading to the bump-like response function shown in Fig. 7d. As a third factor, higher visual contrast (smaller σ^2_{obs}) leads to higher firing rates[69]. As we derive in Supplementary Note 4, Section 6, a neural population with these response functions supports linear readout of input velocities, $\boldsymbol{v}_t/\boldsymbol{\sigma}^2_{\text{obs}}$, and precision, $1/\boldsymbol{\sigma}^2_{\text{obs}}$, in Cartesian coordinates. This provided us with a biologically realistic and, at the same time, theoretically grounded input population model which we used in the following network simulations.

We tested the network's ability to perform online hierarchical inference in the simulation shown in Fig. 7e–j. To challenge the network, we employed a stimulus that combined shared rotation and shared translation (motion tree in Fig. 7e). Six input populations with receptive fields shown in Fig. 7f projected to a distributed population of 100 neurons and a 1-to-1 population of size 8 (one per motion strength). After one second of retinal velocities of counter-clockwise rotation (Fig. 7f, left), these velocities switched to clockwise rotation (center), followed by a superposition of clockwise rotation and rightward translation (right). As the network response for the three populations to this stimulus shows (Fig. 7h–j), input neurons fired sparsely and were only active if the stimulus matched their preferred direction or speed. Neurons in the distributed population, in contrast, showed fluctuating activity with little apparent structure, and exhibited population-wide transients upon changes of the input. Finally, the 1-to-1 population responded more graded and with a short delay, suggesting that every rate, r_m, describes a small cortical population rather than individual neurons. Knowledge of the (randomly drawn) vectors, a_x, of the simulated network, allowed us to read out the network's latent motion decomposition at each time point (solid lines in Fig. 7g). This revealed that the network correctly decomposed the input, including the overlaid rotational and translational motion, and closely matched the online model (dotted lines).

In experiments with humans and animals, we have no access to these readout vectors, $\boldsymbol{a}_x$. We therefore simulated a possible experiment that tests our model and doesn't require this knowledge (see Fig. 7k–m), while benefiting from precise stimulus control. Several apertures, located at the receptive fields of recorded neurons in motion sensitive areas (e.g., area MT or MSTd), present a motion stimulus according to the generative model from Fig. 1. Velocities across the apertures are positively correlated owing to a shared motion source, but also maintain some individual motion (see Fig. 7k and Supplementary Movie 5). A series of trials varies the fraction of shared motion in the stimulus, $q \coloneqq \lambda^2_{\text{shared}}/(\lambda^2_{\text{shared}} + \lambda^2_{\text{ind}})$, ranging from almost independent motion (Fig. 7l, left) to almost perfect correlation (right). According to the network model, $\boldsymbol{\lambda}^2$ can be read out linearly. For the simulation in Fig. 7m, we presented the network with trials of seven values of q. We then trained a linear regression model to predict q from the neural activity for the two most extreme structures (blue dots in Fig. 7m), and decoded q for the intermediate structures using this regression model (red dots in Fig. 7m). Owing to the stochastic stimulus generation, the network's motion structure estimates, $\boldsymbol{\lambda}_t$, fluctuate around the true strength—yet, on average, the trained linear readout correctly identified the fraction of global motion in the stimulus. This is a strong prediction of the network model, which could be tested in a targeted neuroscientific experiment.

Discussion

We have proposed a comprehensive theory of online hierarchical inference for structured visual motion perception. The derived continuous-time model decomposes an incoming stream of retinal velocities into latent motion components which in turn are organized in a nested, tree-like structure. A scene's inferred structure provides the visual system with a temporally robust scaffold to organize its percepts and to resolve momentary ambiguities in the input stream. Applying the theory to human visual motion perception, we replicated diverse phenomena from psychophysics in both object-indexed and location-indexed experiment designs. Furthermore, inspection of the model's internal variables provided normative explanations for putative origins of human percepts and spawned concrete predictions for psychophysics experiments. Finally, the online inference model afforded a recurrent neural network model with visual inputs reminiscent of cortical area MT and latent structure representations reminiscent of area MSTd.

Our online model shares features with predictive coding[70,71], a theory positing that "higher" brain areas provide expectations to earlier areas in a hierarchical model of sensory input and that neural processing aims to minimize prediction errors between top-down expectations and bottom-up observations. Like predictive coding, the dynamics in Eq. (2) update the values of motion sources to minimize prediction errors, $\boldsymbol{\epsilon}_t$, within the bounds imposed by the identified structure. Yet, structure identification according to Eq. (1) follows a different principle by computing a running average of motion source magnitudes. This contrasts with common theories of predictive coding in the brain[72,73], which assume that the same computational principle is repeated across cortical hierarchies, and demonstrates how hierarchical visual processing could combine multiple interacting algorithmic motifs. Moreover, the network model in Fig. 7a challenges the prevalent view[72,74] that error signals are necessarily represented by distinct neural populations (or alternatively distinct dendritic compartments[75]). While our network model supports the possibility of distinct error populations, we show that prediction errors could also be computed and conveyed by the same neurons representing other quantities, such as the motion sources, $\boldsymbol{\mu}_t$, and even the structure, $\boldsymbol{\lambda}^2_t$, using a distributed neural code.

In the main text, we have for the sake of clarity limited the presentation of the theory to a basic version that nonetheless covers all essential concepts. In Supplementary Note 3, we present several extensions that are naturally covered by our model:

(i) observation noise, σ_{obs}, can be time- and object-dependent, which is relevant for modeling temporary occlusion of a subset of stimuli;
(ii) observation noise can be non-isotropic (different values in x- and y-direction), which is relevant for angle-dependent edge velocities in apertures[76];
(iii) for optimal inference, different motion components can feature different time constants, since velocity is expected to change more slowly for heavy objects due to higher inertia;
(iv) different motion components may tend to co-occur or exclude one another in real-world scenes, which can be modeled by an interaction prior of pairwise component compatibility; and
(v) when motion components are not present for a long time, they will decay to zero, preventing their rediscovery, which can be mitigated by a prior on motion strengths.

The current theory is limited to velocities as input, thereby ignoring the well-documented influence of spatial arrangement on visual motion perception, such as center-surround modulation[77,78],

adjacency[26] or motion assimilation[79], as well as Gestalt properties[80]. Furthermore, the model does not solve the correspondence problem in object-indexed experiments, but simply assumes that velocities are correctly assigned to the input vector as objects move about the visual field. For location-indexed experiments, we have explored how structure inference in concert with a basic assignment process, which minimizes the observer's local prediction errors, could solve the correspondence problem during structure-from-motion perception. Our work focuses on the simultaneous inference of motion sources, $\boldsymbol{s}_t$, and motion strengths, $\boldsymbol{\lambda}_t$. Other quantities, such as time constants and, probably more importantly, the motion components, $\boldsymbol{C}$, have been assumed to be given. It is worth noting, however, that gradient-based learning of $\boldsymbol{C}$ is, in principle, supported by the theory on long time scales (see Supplementary Note 3, Section 5). Finally, limited experimental evidence of the neural correlates of motion structure perception required the neural network model to rely on many modeling assumptions. The model's predictions should act as a starting point for further scientific inquiry of these neural correlates.

Even though the sensory processes underlying object-indexed motion perception necessarily differ from those of location-indexed perception, our model describes human perception for both types of experiments. Thus, both types might share the same underlying neural mechanisms for structure inference. This raises the intriguing question whether there exist stable, object-bound neural representations of velocity. Furthermore, our work points towards a tight link between neural representations of latent structure and representations of uncertainty in that the estimated motion strengths, $\boldsymbol{\lambda}_t$, determine the credit assignment of prediction errors through the gating function, $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$—a function that also computes the variance of motion components, e.g., the brain's uncertainty about flock velocity. Behaviorally, sensory noise directly impacts the perceived structure of a scene as demonstrated experimentally by the perceptual reversal in the Lorenceau-motion illusion[42] (cf. Fig. 5). More generally, our theory predicts that the visual system will organize its percepts into simpler structures when sensory reliability decreases. Moreover, the reliability of visual cues plays a role in multisensory integration[81], with area MSTd[82,83], but not area MT[84], exhibiting tuning to vestibular signals. Thus, MSTd may be a candidate area for multisensory motion structure inference. Overall, we expect our theoretical results to guide targeted experiments in order to understand structured visual motion perception under a normative account of statistical information processing.

Methods

In what follows, we provide an overview of the generative model, the online hierarchical inference model, the computer simulations, and the data analysis. A more detailed presentation is found in the Supplementary Information.

Generative model of structured motion

We consider K observable velocities, $v_{k,d}(t)$, in D spatial dimensions. For notational clarity, we will consider in this Methods section only the case $D=1$ and use the vector notation, $\boldsymbol{v}_t=(v_1(t),..,v_K(t))^\mathsf{T}$. The extension to $D>1$ is covered in Supplementary Note 1, Section 4. Observable velocities, $\boldsymbol{v}_t$, are generated by M latent motion sources, $s_{m,d}(t)$, abbreviated (for $D=1$) by the vector $\boldsymbol{s}_t=(s_1(t),..,s_M(t))^\mathsf{T}$. Velocities are noisy instantiations of their combined ancestral motion sources, $\boldsymbol{v}_t\sim\mathcal{N}(\boldsymbol{C}\,\boldsymbol{s}_t,\sigma^2_{\text{obs}}/\delta t\,\boldsymbol{I})$, where $C_{km}=+1,-1$, and 0 in $K\times M$ component matrix, $\boldsymbol{C}$, denote positive, negative and absent influence, respectively. For the formal definition, observations, $\boldsymbol{v}_t$, remain stable within a short time interval $[t,t+\delta t)$, and the observation noise variance, $\sigma^2_{\text{obs}}/\delta t$, ensures a δt-independent information content of the input stream. In the online inference model, below, we will draw the continuous-time limit, which will become independent of δt. In computer simulations, δt is the inverse frame rate of the motion display (default value: $1/\delta t=60$ Hz). Each motion source (in each spatial dimension) follows an Ornstein–Uhlenbeck process, $\mathrm{d}s_m=-s_m/\tau_s\,\mathrm{d}t+\lambda_m\,\mathrm{d}W_m$, with time constant τ_s, motion strength λ_m (shared across dimensions), and Wiener process W_m. The OU process's equilibrium distribution, $\mathcal{N}\left(0,\frac{\tau_s}{2}\lambda_m^2\right)$, introduces a slow-velocity prior which, as we note, has recently been proposed to originate from the speed-contrast statistics of natural images[85]. The resulting marginal stationary velocity distribution of v_k is $v_k\sim\mathcal{N}\left(0,\sigma^2_{\text{obs}}/\delta t+\frac{\tau_s}{2}\sum_{m=1}^M C_{km}^2\lambda_m^2\right)$.

Radial and rotational motion sources. In location-indexed experiments, the input's location (e.g., a neuron's receptive field) remains fixed. For $D=2$, the fixed input locations enable our model to support rotations and expansions around various axes. In this manuscript, we consider two cases: rotation around a vertical axis (SfM experiment in Fig. 6) and rotation/expansion around the fovea (network model in Fig. 7).

For rotations around a vertical axis, each input $\boldsymbol{v}_k$ has fixed polar coordinates (R_k,φ_k) as sketched in Fig. 6b. When describing rotation by means of a rotational motion source, s_t^{rot}, we obtain for the noise-free part of the observed velocity in Cartesian coordinates: $v_{k,x}=-R_k\,\sin(\varphi_k)\,s_t^{\text{rot}}$, $v_{k,y}=0$, and $v_{k,z}=-R_k\,\cos(\varphi_k)\,s_t^{\text{rot}}$. Owing to the linear dependence of $\boldsymbol{v}_k$ on s^{rot}, we can include the coefficients as a column in component matrix, $\boldsymbol{C}$, and s^{rot} as a motion source in the vector $\boldsymbol{s}_t$. Note that in the SfM experiments only the x- and y-directions are observed.

Similarly, for rotation/expansion around the fovea, each input $\boldsymbol{v}_k$ has fixed polar coordinates (R_k,ϑ_k) with radial distance R_k and angle ϑ_k, relative to the pivot point (we use different symbols than for vertical rotation for notational clarity). Denoting radial and rotational motion sources by s_r and s_φ, we obtain for the noise-free part of $\boldsymbol{v}_k$ in Cartesian coordinates: $v_{k,x}=s_r\,\cos\vartheta_k-s_\varphi R_k\sin\vartheta_k$, and $v_{k,y}=s_r\,\sin\vartheta_k+s_\varphi R_k,\cos\vartheta_k$. Since R_k and ϑ_k are fixed coefficients, the mapping $(s_r,s_\varphi)\mapsto(v_{k,x},v_{k,y})$ is linear and, thus, can be described by the component matrix $\boldsymbol{C}$. The full derivation and an illustration of the velocity relations in polar coordinates are provided in Supplementary Note 1, Section 5.

Online inference

The goal of motion structure inference is to simultaneously infer the value of motion sources, $\boldsymbol{s}_t$, and the underlying structure, $\boldsymbol{\lambda}$, from a stream of velocity observations. The number of spatial dimensions, D, component matrix, $\boldsymbol{C}$, time constant τ_s, and observation noise σ_{obs} are assumed to be known. The EM algorithm leverages that changes in $\boldsymbol{s}_t$ and $\boldsymbol{\lambda}$ (if changing at all) occur on different time scales, τ_s and τ_λ, respectively. For $\tau_\lambda\gg\tau_s$, the EM algorithm treats $\boldsymbol{\lambda}$ as a constant for inferring $\boldsymbol{s}_t$ (E-step), and optimizes an estimate, $\boldsymbol{\lambda}_t$, online based on the inferred motion sources (M-step).

E-Step. For fixed $\boldsymbol{\lambda}$, the posterior $p(\boldsymbol{s}_t\,|\,\boldsymbol{v}_{0:t};\,\boldsymbol{\lambda})$ is always a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)$, and can be calculated by a Kalman-Bucy filter[86,87]; see Supplementary Note 2, Sections 1, 2, and 3.1 for the derivation. This yields coupled differential equations for the time evolution of $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$. To reduce the computational complexity of the system, we perform an adiabatic approximation on the posterior covariance, $\boldsymbol{\Sigma}_t$, by assuming (a) that it has always converged to its stationary value, and (b) that off-diagonal values in $\boldsymbol{\Sigma}_t$ are zero, that is, we ignore correlations in uncertainty about latent motion sources in the posterior distribution. As shown in the full derivation in Supplementary Note 2, Section 3, the first assumption is warranted because the stationary value of $\boldsymbol{\Sigma}_t$ depends only on the current structure estimate, $\boldsymbol{\lambda}_t$; then, because $\boldsymbol{\Sigma}_t$ decays to stationarity at time scale $\tau_s/2$, it can always follow any changes in $\boldsymbol{\lambda}_t$ which happen at time scale $\tau_\lambda\gg\tau_s$. The

Table 1 | Default parameters of the computer simulations

Description	Variable	Object-indexed	Location-indexed	Network
Time const. motion sources	τ_s	0.300 s	0.100 s	0.100 s
Time const. motion strengths	τ_λ	1.000 s	0.333 s	0.333 s
Inv. observation frame rate	δt	1/60 s	1/60 s	1/120 s
Observation noise	σ_{obs}	0.05	$0.017 = 0.05 \div 3$	0.017
Initial motion strength	$\lambda_m(t=0)$	0.5	0.5	0.5
No. of pseudo observation	ν_m	0	0/−1	0
Val. of pseudo observations	κ_m	0	0	0
Vestibular input	v_{vst}	–	0	–
Obs. noise for vestibular input	σ_{vst}	–	0.05	–
Time const. for pred. err.	τ_ϵ	–	–	0.050 s

Most parameters are maintained throughout all computer experiments. Deviations from these parameters are listed in the respective experiment description. The value $\nu_m = -1$ in location-indexed experiments relates to self-motion. For $D = 2$ spatial dimensions, $\nu_{\text{self}} = -2/D = -1$ yields a uniform prior distribution (see Supplementary Note 2, Section 1.3).

second assumption is a modeling assumption: that biological agents might disregard the subtle (and complicated) interactions between the uncertainties of different motion sources and rely on their individual uncertainties, instead. Using the two assumptions we derive a closed-form solution for the posterior variance,

$$\Sigma_{mm} = \frac{\sigma_{\text{obs}}^2}{\tau_s \, \|\boldsymbol{c}_m\|^2} \left(-1 + \sqrt{1 + \frac{\tau_s{}^2 \, \|\boldsymbol{c}_m\|^2}{\sigma_{\text{obs}}^2} \lambda_m^2} \right) =: f_\Sigma(\lambda_m^2) \,, \tag{5}$$

with $\|\boldsymbol{c}_m\|^2 = \sum_{k=1}^{K} C_{km}^2$ denoting the vector-norm of the m-th column of $\boldsymbol{C}$. This is Eq. (3) of the main text. The plot in Fig. 1i has parameters $\|\boldsymbol{c}_m\|^2 = 4$, $\tau_s = 300$ ms, and $\sigma_{\text{obs}} = 0.05$. By plugging the adiabatic approximation of the variance into the time evolution of $\boldsymbol{\mu}_t$, we arrive at Eq. (2) of the main text (see Supplementary Note 2, Section 3.4 for the derivation).

M-step. Using the posterior from the E-step, motion strengths, $\boldsymbol{\lambda}$, are optimized to maximize the likelihood of the observed velocities. This optimization further incorporates prior distributions, $p(\lambda_m^2)$, most conveniently formulated over the squared motion strengths, for which we employ a scaled inverse chi-squared distribution,

$$p(\lambda_m^2) = \mathcal{I}\chi\left(\lambda_m^2 \,;\, \nu_m, \kappa_m^2\right) = \frac{1}{\lambda_m^{(2+\nu_m)}} \exp\left[-\frac{\nu_m \kappa_m^2}{2\lambda_m^2} - A\left(\nu_m, \kappa_m^2\right) \right], \tag{6}$$

owing to its conjugacy to estimating the variance of s_m (this is what λ_m^2 controls). The prior features two hyper-parameters, ν_m and κ_m^2, which give rise to an intuitive interpretation as ν_m pseudo-observations of average value κ_m^2. The partition function, $A(\nu_m, \kappa_m^2)$, only serves for normalization. By default, we employ a Jeffreys prior ($\nu_m = \kappa_m^2 = 0$), which is a typical choice as a non-informative prior in Bayesian statistics and promotes a preference for finding simple structures by assigning higher beliefs to small values of λ_m (and highest to $\lambda_m = 0$). The only exception is the motion strength assigned to self motion, λ_{self}, for which we employ a uniform prior distribution, formally by setting $\nu_{\text{self}} = -2$ and $\kappa_{\text{self}}^2 = 0$. These choices reflect the a-priori belief that motion components supported by $\boldsymbol{C}$ will usually be absent or small in any given scene—with the exception of self-motion-induced velocity on the retina, which occurs with every saccade and every turn of the agent's head (see Supplementary Note 2, Section 1.2 for the formal calculation of the M-step).

In the online formulation of EM (see Supplementary Note 2, Sections 2.3 and 3.4 for the derivation of the online EM algorithm and of the online adiabatic inference algorithm which constitutes our model, respectively), these priors give rise to the low-pass filtering dynamics in Eq. (1) for updating λ_m^2, with constants

$$\alpha_m = \frac{2}{\tau_s{}^2 \, (2 + \nu_m + \tau_\lambda/\tau_s)} \,, \text{ and} \tag{7}$$

$$\beta_m = \frac{\nu_m \, \kappa_m^2}{\tau_\lambda \, (2 + \nu_m + \tau_\lambda/\tau_s)} \,. \tag{8}$$

This completes the derivation of the online model for $D = 1$ spatial dimensions. The extension to multiple dimensions is straightforward and provided in Supplementary Note 2, Sections 1.3, 2.3 and 3.4 alongside the respective derivations.

Preference for simple structures. The above Jeffreys prior on motion strengths, $p(\lambda_m^2)$, facilitates the discovery of sparse structures. This property is important when a large reservoir of possible motion components in $\boldsymbol{C}$ is considered: the model will recruit only a small number of components from the reservoir. In Supplementary Fig. 2, we demonstrate this ability for the example of the Johansson experiment from Fig. 2b–d by duplicating the shared motion component, i.e., the first two columns in $\boldsymbol{C}$ are all 1's. As Supplementary Fig. 2 shows, the model recruits only one of the two identical components and discards the other. This example of identical components in the reservoir represents the theoretically hardest scenario for maintaining a sparse structure.

Computer simulations

Computer simulations and data analysis were performed with custom Python code (Python 3.8, Numpy 1.21, Scipy 1.7, scikit-learn 0.24, Matplotlib 3.4, Pandas 1.3, xarray 0.19). The code has been published on GitHub[88] and supports most of the extensions presented in Supplementary Note 3.

For the numerical simulation, input was drawn with observation noise variance $\sigma_{\text{obs}}^2/\delta t$, at the time points of input frames (every δt). The drawn input remained stable until the next frame. Between frames, the differential equations for online hierarchical inference were integrated with SciPy's explicit Runge-Kutta method RK45 which adapts the step size. This integration method combines numerical accuracy with a parameterization that is almost invariant to the input frame rate. The default parameters that we used are listed in Table 1. The data shown in the figures is provided in a supplementary source data file.

Hierarchical motion experiments (Fig. 2)

For the Johansson experiment, all $K = 3$ dots followed sinusoidal velocities with frequency 0.5 Hz. Horizontal amplitudes were $2\sqrt{\tau_s}$ for all dots; vertical amplitudes were 0 for the outer dots and $\cos(45^\circ) \cdot 2\sqrt{\tau_s}$ for the inner dot. For the Duncker wheel, we set the wheel radius to $R = 1$ and the rotation frequency to 1 Hz. This leads to the hub velocity

$v_{\text{hub, y}} = 0$ and $v_{\text{hub, x}} = 2\pi\,\text{s}^{-1}$ because the hub must travel $2\pi R$ during one period for slip-free rolling. For the rim velocities, being the derivatives of location, we thus find $v_{\text{rim,x}} = v_{\text{hub,x}} + R\,\omega\,\cos(\omega\,t)$ and $v_{\text{rim,y}} = -R\,\omega\,\sin(\omega\,t)$, with $\omega = 2\pi\,\text{s}^{-1}$. For the simulation, we increased the observation noise to $\sigma_{\text{obs}} = 0.15$ and set $\lambda_m(t=0) = 0.1$ to highlight the gradual discovery of the motion components.

Structure classification (Fig. 3)

The stimulus data and human responses were released by Yang et al.[17] on GitHub. The experiment is described in detail in ref. 17. There were 12 participants with each participant performing 200 trials. Each trial consisted of three dots moving on a circle for 4 s. Dots had different colors to prevent their confusion, but colors did not convey any information on the dots' roles within the structure. No data was excluded. Trials were generated stochastically from the same generative model that is considered in this work, with uniform probability for each of the four structures (Independent, Global, Clustered, Hierarchical) to underlie the trial. Motion strengths were chosen such that all dots had identical marginal velocity distributions, $p(v_k)$, across all structures—leaving motion relations as the only distinguishing information (see ref. 17, for detailed stimulus parameters and λ-values of all structures). Like Yang et al.[17], we treated the experiment as one-dimensional ($D = 1$), operating directly on the angular velocities. Noise-free angular velocities were calculated from the circular distance of subsequent stimulus frames, and we set $1/\delta t = 50$ Hz to match the experiment's frame rate.

For presenting the trials to our online inference model, we initialized each of the λ_m at its average value (average taken across the ground truth of all structures). At trial end, the model yielded $M = 7$-dimensional $\boldsymbol{\lambda}$-vectors associated with 1 shared component, 3 cluster components (one per possible pair), and 3 individual components (see Supplementary Fig. 3 for example trials). For logistic regression, we calculated 5 features, T_i, from $\boldsymbol{\lambda}$, namely:

$$
\begin{aligned}
&T_1 = \lambda_1 / \sum_m \lambda_m && \text{Does shared motion stand out?}\\
&T_2 = \max\{\lambda_2,\lambda_3,\lambda_4\} / \sum_{m=2,3,4} \lambda_m && \text{Does one cluster dominate the others?}\\
&T_3 = \max\{\lambda_5,\lambda_6,\lambda_7\} / \sum_{m=5,6,7} \lambda_m && \text{Does one individual component stand out?}\\
&T_4 = \lambda_c^2 / \sum_{m=c,\,\text{Ch}_1(c),\,\text{Ch}_2(c)} \lambda_m^2 \ \text{ with } c = \text{argmax}(\lambda_2,\lambda_3,\lambda_4) && \text{Does the strongest cluster dominate its children?}\\
&T_5 = \lambda_c^2 / \sum_{m=c,\,\neg\text{Ch}(c)} \lambda_m^2 \ \text{ with } c = \text{argmax}(\lambda_2,\lambda_3,\lambda_4) && \text{Does the strongest cluster dominate the 3rd dot?}
\end{aligned}
\tag{9}
$$

Here, $\text{Ch}_{1,2}(c)$ denote the individual motion components of the two dots within the cluster component c, and $\neg\text{Ch}(c)$ denotes the dot not being in cluster c. The features were hand-designed based on the reasoning that they may be useful for structure classification. Their most important property is that all information about a trial is conveyed through $\boldsymbol{\lambda}$ as a bottleneck. A multinomial logistic regression classifier was trained with L1-regularization on the feature vectors, $(T_1, \ldots, T_5)$, to classify the ground truth structures of the trials. Then, we fitted the same choice model as ref. 17 to the human choices, but replaced the ideal observer log-probability, $\log p(S \mid \boldsymbol{v}_{0:T})$, which was used in ref. 17, with the class probability from the trained classifier, $\log p(S \mid \boldsymbol{\lambda})$:

$$
P(\text{choice} = S) = \pi_{\text{L}} \frac{1}{4} + (1 - \pi_{\text{L}})\ \exp\left[\beta\left(\log p(S \mid \boldsymbol{\lambda}) + b_S\right)\right] / \text{Norm.}, \tag{10}
$$

with lapse probability, π_{L}, inverse temperature, β, and biases, b_S, for all structures, $S = G,\ C,\ H$, relative to the independent structure ($b_I = 0$ by convention). Note that, in contrast to ref. 17, we do not need to consider structure multiplicities here because the features are already symmetric with regard to the three possible cluster assignments. Like ref. 17, we did not apply observation noise to the presented velocities, but maintained a non-zero observation noise parameter, σ_{obs}, for the inference. Observation noise, σ_{obs}, and lapse probability, π_{L}, were shared parameters for all participants and were fitted jointly via a simple grid search. We obtained $\sigma_{\text{obs}} = 0.04$ and $\pi_{\text{L}} = 4\%$ (compared to 14% in ref. 17). The remaining 4 parameters, $\{\beta, b_G, b_C, b_H\}$, were fitted via maximum likelihood for each participant. All reported confusion matrices and log-likelihoods were obtained by fitting the 4 per-participant parameters using leave-one-out cross-validation. The log-chance level in Fig. 3f is $200 \cdot \log(1/4)$ since each participant performed 200 trials.

Location-indexed experiments (Figs. 4–6)

To support self-motion, we introduce a column of −1's in $\boldsymbol{C}$ as an additional component, which is connected to all visual velocity inputs and to a vestibular input v_{vst}. In our simulations, the vestibular input is always stationary, but noisy: $v_{\text{vst}} \sim \mathcal{N}(0, \sigma_{\text{vst}}^2)$. The associated self-motion strength, λ_{self}, uses a uniform prior (see discussion under Eq. (6)). Perceived velocities are the sum over all-except-self-motion: $\boldsymbol{v}_{\text{perceived}} = \sum_{m \neq \text{self}} \boldsymbol{C}_{*m} \boldsymbol{\mu}_m$.

Motion-direction-repulsion (MDR) experiments (Fig. 4)

In the MDR experiments with two RDKs, input was modeled as $K = 3$ velocities: two for the two groups of dots, plus the vestibular input. Repulsion angles were estimated from 20 repetitions of 30 s long trials, with $\boldsymbol{v}_{\text{perceived}}$ averaged over the last 10 s of each trial. Error bars from the simulations were too small to be shown in Fig. 4e–g.

In Fig. 4e, the velocities for opening angle, γ, were given by $(v_x, v_y) = v_0 \cdot (\cos(\gamma/2), \sin(\gamma/2))$ for the first group, with $v_0 = 2\sqrt{\tau_s}$, and $v_0 \cdot (\cos(\gamma/2), -\sin(\gamma/2))$ for the second group. As in Fig. 3 of ref. 36, the repulsion bias was measured with respect to the full opening angle.

In Fig. 4f, increasing contrast of the second group was modeled as dividing the observation noise variance by a factor, f, between 0.001 and 10, leading to variance σ_{obs}^2/f for this group's input. As in ref. 37, the repulsion bias was measured only with respect to the first group's perceived direction. The expressed similarity to experimental data refers to the "2-motion condition" in Fig. 7 of ref. 37.

In Fig. 4g, the velocity of the second group was multiplied by a factor between 0 and 2, and the repulsion bias was measured only with respect to the first group's perceived direction. For a 60° opening angle, we qualitatively replicate the experimental data from Fig. 2a, b in ref. 38. In order to maintain the simulation parameters from previous conditions, we did not attempt to quantitatively match the speed of targets and distractors in ref. 38. A direct quantitative comparison to the human data from Fig. 4b in ref. 36 is difficult because they had measured the point of subjective equality (PSE) to a 90° opening angle for this stimulus condition, finding a 10° bias for the full opening angle.

For the Takemura experiment[39] in Fig. 4h–l, we used $K = 5$ inputs: two inner RDKs, two outer RDKs, and the vestibular signal, which were organized in the motion tree shown in Supplementary Fig. 5a. If not mentioned otherwise, the simulation parameters matched those from the basic MDR experiment in Fig. 4e. The inner stimuli had $v_x = \pm v_0$, and $v_y = v_0$ if non-zero. The outer stimuli had $v_x = 0$, and $v_y = \pm v_0$. The standard deviation of the observation noise of the outer RDKs was divided by factor 6, reflecting that in ref. 39 the outer RDKs covered a three-times larger area and had twice the dot density of the inner RDKs. Each histogram is based on 200 trial repetitions, which use identical initial conditions but different realizations of observation noise, with perceived velocities measured at trial end. The conditions in panels Fig. 4h–l correspond to figure panels 4a, 4b, 6 left, 6 center, 6 right, in ref. 39. Besides transparent motion (i.e., two perceived velocities), Takemura et al. reported also coherent motion (i.e., only one perceived velocity) for the inner RDKs in a fraction of trials. In our computer simulations, we focused only on the biased perception of two velocities.

Lorenceau illusion (Fig. 5)

For the Lorenceau illusion, we modeled each dot's velocity as a separate input owing to the spatially distributed nature of the stimulus. As in ref. 42, the two groups of 10 dots each oscillated at a frequency of 0.83 Hz. For the oscillation amplitude, we chose $R = 1/2$ (arbitrary units), leading to velocities $v_x(t) = R\,\omega\,\cos(\omega\,t)$ for the horizontal group and $v_y(t) = -R\,\omega\,\sin(\omega\,t)$ for the vertical group, with $\omega = 2\pi \cdot 0.83\,\mathrm{s}^{-1}$. As shown in Supplementary Fig. 6, the model decomposes this stimulus into a deeply nested structure comprising self-, shared-, group-, and individual motion. For the noise-free stimulus condition, we used the default simulation parameters. For the condition with motion noise, the observation noise, σ_{obs}, of the visual inputs (not the vestibular input) was multiplied by 25.

Structure-from-motion (SfM) experiments (Fig. 6)

We treat SfM as a location-indexed experiment owing to experimental findings[50,52]. For computer simulations, we model each cylinder as a ring in the x-z-plane, conflating its height into one receptive field (the simulations still run in 2D with x- and y-dimensions being modeled). The outer cylinder had radius $R = 1.5$, and the inner, if present, $R = 1.0$. Normal rotation speed was 90°/s, and fast speed was 135°/s. Velocities were observed at seven equidistant receptive field locations along the x-axis, $x_{\mathrm{RF}} \in \{1.2, 0.8, 0.4, \ldots, -1.2\}$. These correspond to angles, φ_k, on the cylinders via $x_{\mathrm{RF}} = R\,\cos(\varphi_k)$ with the inner cylinder covering only five RF locations (cf. Fig. 6b). When presenting velocity observations, v_k, each RF location x_{RF} had multiple overlapping v_k (2 for one cylinder, 4 for nested cylinders where they overlapped). The observation noise for velocity inputs, σ_{obs}, was multiplied by 20 for the single cylinder-condition and by 30 for the nested cylinders-conditions, reflecting the high local ambiguity when measuring multiple overlapping speeds and directions[89]. For consistency with other simulations, we provided a vestibular input with the same parameters as in previous location-indexed experiments, although this signal plays no computational role in the SfM simulations.

The model's component matrix, $\boldsymbol{C}$, comprised translational self-motion, rotational motions for the outer cylinder, the inner cylinder (only in nested conditions), and shared for both cylinders (only in nested conditions), as well as translational individual components for each v_k (see Fig. 6c, g). Rotational motion is naturally covered by our model as presented in *Radial and rotational motion sources* earlier in Methods and sketched in Fig. 6b. To solve the correspondence problem of overlapping v_k, we devised the following assignment process. At every integration time step, δt, and for every RF location, x_{RF}, keep the previous assignment with probability 0.7, and continue to the next x_{RF}. Else, that is if the assignment is re-evaluated, calculate the Euclidean distance between the model's expected velocities, $\boldsymbol{C}\boldsymbol{\mu}_t$, and the observed velocities, $\boldsymbol{P}_j\,\boldsymbol{v}_t$, for all permutations, $\boldsymbol{P}_j$, of the overlapping inputs within this RF. Then choose the assignment, $\boldsymbol{P}_j$, that minimizes the Euclidean distance, i.e., the local prediction error, $\boldsymbol{\epsilon}_t$, within the RF. Once all x_{RF} were processed in this manner, perform the integration of $\partial_t\boldsymbol{\mu}_t$ according to Eq. (2) using the assigned permutations of $\boldsymbol{v}_t$. This completes the model for SfM perception. We note that, since the integration is performed only after all RF assignments have been made, the resulting global assignment process is independent of the order of iterating over the RFs and could, in principle, be performed in parallel and continuous time. The fact that all computations are spatially confined to information within each RF further improves the process's biological plausibility.

For obtaining the switching distribution in Fig. 6e, we performed a 10,000 s long simulation and followed ideas from ref. 90: first we identified a perceptual threshold as the mode of $\{|\mu_t^{\mathrm{rot}}|\,\forall\,t\}$ (the exact value is actually not important). Then we defined two possible percepts which correspond to positive (negative) values of μ_t^{rot}. A perceptual switch occurred whenever μ_t^{rot} crossed the negative (positive) threshold of the other percept. The Gamma distribution was fitted by maximum likelihood.

Network implementation (Fig. 7)

A detailed derivation of how to implement the online hierarchical inference model in a neural network model is provided in Supplementary Note 4. In the following, we will focus on the specific model parameters used in the simulations of Fig. 7.

For both simulations (the demonstration in Fig. 7e–j and the proposed experiment in Fig. 7k–m), there were $K = 6$ location-indexed input variables in $D = 2$ spatial dimensions. Input was encoded according to the model of area MT presented in Supplementary Note 4, Section 6. Each velocity, $\boldsymbol{v}_k$, was encoded by a population of 192 neurons, with tuning centers organized on a polar grid with $N_\alpha = 16$ preferred directions, and $N_\rho = 12$ preferred speeds (sketched in Fig. 7c for smaller values of N_α and N_ρ). Each neuron in each of the K populations thus has coordinates (n_α, n_ρ) describing its preferred direction and speed. To account for the reported bias of MT tuning toward slow speed[67], the density of preferred speeds became sparser for higher speeds, which we modeled in Supplementary Note 4, Eq. (70) by $\mu_\rho(n_\rho) = \rho_{\min} + d_\rho\,n_\rho^{1.25}$, with $d_\rho = (\rho_{\max} - \rho_{\min})/(N_\rho - 1)^{1.25}$, and $\rho_{\min} = 0.1$, $\rho_{\max} = 8.0$, for neurons $n_\rho = 0, \ldots, N_\rho - 1$. Preferred directions covered the circle equidistantly. The remaining parameters in the tuning function were $\kappa_\alpha = 1/0.35^2$ and $\sigma_\rho^2 = 0.35^2$ for the angular and radial tuning widths, respectively, and $\psi = 0.1$ Hz for the overall firing rate scaling factor. For the network simulations, we increased the frame rate to $\delta t = 1/120$ Hz for the sake of a higher sampling rate on the x-axis in Fig. 7h–j (the simulation software stores firing rates only at the time of frames).

The distributed population comprised 100 neurons. Readout vectors, $\boldsymbol{a}_x$, for all variables represented by this population were drawn i.i.d. from a standard normal distribution, $\mathcal{N}(0,1)$, for each vector element. Adjoint matrices were calculated numerically to fulfill the required orthonormality conditions (see Supplementary Note 4, Section 4). The low-pass filtering time constant of the prediction error was $\tau_\epsilon = \tau_s/2 = 0.050$ s, such that the prediction error could react to changes in $\boldsymbol{\mu}_t$.

The one-to-one population comprised M neurons (or small populations; $M = 8$ for the demo, and $M = 7$ for the proposed experiment), one per function value, $f_\Sigma(\lambda_m^2)$. The proportionality constant for the readout was $f_\Sigma(\lambda_m^2) = 0.001\,r_{1\text{-to-}1,m}$.

Given the parameters and decoding vectors, the simulation software automatically transforms the differential equations of the online inference model into the corresponding neural dynamics, according to the rules stated in Supplementary Note 4, Section 4. Numerical integration of neural dynamics was performed by the same RK45 method used in the previous simulations.

For the demonstration in Fig. 7e–j, inputs were arranged on a ring of radius $R_k = 1$ with angular location $\vartheta_k = 60° \cdot k$ (measured from the x-axis in counter-clockwise direction). Presented velocities were $(v_x, v_y) = (-2\,\sin(\vartheta_k), 2\,\cos(\vartheta_k))$ for $t \le 1$ s, $(2\,\sin(\vartheta_k), -2\,\cos(\vartheta_k))$ for $1\,\mathrm{s} < t \le 2$ s, and $(2 + 2\,\sin(\vartheta_k), -2\,\cos(\vartheta_k))$ for $2\,\mathrm{s} < t$. In the $\boldsymbol{C}$-matrix underlying the network construction, the shared polar visual component was constructed according to Supplementary Note 1, Eq. (6). The shared translational and the 6 individual components were Cartesian.

In the proposed experiment in Fig. 7k–m, all motion components and the input were Cartesian such that input location played no role (formally, we maintained the circular arrangement of the previous network simulation). Input was generated from the model's underlying generative model for a motion tree comprising 1 shared component and 6 individual components. For a given fraction of shared motion, q, we set $\lambda_{\mathrm{shared}}^2 = 2^2\,q$ and $\lambda_{\mathrm{ind}}^2 = 2^2\,(1-q)$. Maintaining constant total squared motion strength, $\lambda_{\mathrm{shared}}^2 + \lambda_{\mathrm{ind}}^2 = 4$, ensures that the (marginal) input velocity distributions are statistically identical across all input locations and all values of q. In total, seven fractions,

$q = 1/8, 2/8, \ldots, 7/8$, of shared motion were presented. Per simulation run, each fraction was presented for 10 s, and simulations were repeated for 10 runs. For the subsequent data analysis, the neural responses of only the 2nd half ($5\,s \le t$) of the stimulus presentation were considered to avoid potential initial transients. A standard linear regression model (with intercept; `class sklearn.linear_model.LinearRegression`) was trained to decode the correct q from the distributed population's response, r_{dis}, for the fractions $q = 1/8$ and $q = 7/8$. The resulting linear readout (with intercept) was employed to decode q from r_{dis} for the remaining stimuli in Fig. 7m.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

No new experiment data was produced for this study. The behavioral data from ref. 17 is available with the original publication. The behavioral data for ref. 36 has been digitized by the authors and is included in the software repository: https://github.com/DrugowitschLab/structure-in-motion/blob/main/data/data_Braddick_2002_Fig3C.txt. Source data are provided with this paper.

Code availability

Computer simulations, data analyses and visualization have been performed with custom Python code which has been released[88] under the BSD 3-clause license and is available online: https://github.com/DrugowitschLab/structure-in-motion.

References

1. Kaiser, D., Quek, G. L., Cichy, R. M. & Peelen, M. V. Object vision in a structured world. *Trends Cognit. Sci.* **23**, 672–685 (2019).
2. Yantis, S. Multielement visual tracking: attention and perceptual organization. *Cognit. Psychol.* **24**, 295–340 (1992).
3. Driver, J., McLeod, P. & Dienes, Z. Motion coherence and conjunction search: implications for guided search theory. *Percept. Psychophys.* **51**, 79–85 (1992).
4. Royden, C. S. & Hildreth, E. C. Human heading judgments in the presence of moving objects. *Percept. Psychophys.* **58**, 836–856 (1996).
5. Liu, G. et al. Multiple-object tracking is based on scene, not retinal, coordinates. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 235–247 (2005).
6. Xu, H., Tang, N., Zhou, J., Shen, M. & Gao, T. Seeing "what" through "why": evidence from probing the causal structure of hierarchical motion. *J. Exp. Psychol. General* **146**, 896–909 (2017).
7. Dokka, K., Park, H., Jansen, M., DeAngelis, G. C. & Angelaki, D. E. Causal inference accounts for heading perception in the presence of object motion. *Proc. Natl Acad. Sci.* **116**, 9060–9065 (2019).
8. Bolton, A. D. et al. Elements of a stochastic 3D prediction engine in larval zebrafish prey capture. *ELife* **8**, e51975 (2019).
9. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).
10. Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).
11. Stocker, A. A. & Simoncelli, E. P. Sensory adaptation within a Bayesian framework for perception. In *Advances in neural information processing systems* (NeurIPS, 2005).
12. Welchman, A. E., Lam, J. M. & Bülthoff, H. H. Bayesian motion estimation accounts for a surprising bias in 3D vision. *Proc. Natl Acad. Sci.* **105**, 12087–12092 (2008).
13. Vul, E., Frank, M. C., Tenenbaum, J. B. & Alvarez, G. A. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in neural information processing systems* (NeurIPS, 2009).
14. Hedges, J. H., Stocker, A. A. & Simoncelli, E. P. Optimal inference explains the perceptual coherence of visual motion stimuli. *J. Vis.* **11**, 14 (2011).
15. Gershman, S. J., Tenenbaum, J. B. & Jäkel, F. Discovering hierarchical motion structure. *Vis. Res.* **126**, 232–241 (2016).
16. Bill, J., Pailian, H., Gershman, S. J. & Drugowitsch, J. Hierarchical structure is employed by humans during visual motion perception. *Proc. Natl Acad. Sci.* **117**, 24581–24589 (2020).
17. Yang, S., Bill, J., Drugowitsch, J. & Gershman, S. J. Human visual motion perception shows hallmarks of Bayesian structural inference. *Sci. Rep.* **11**, 3714 (2021).
18. Barlow, H. & Levick, W. R. The mechanism of directionally selective units in rabbit's retina. *J. Physiol.* **178**, 477–504 (1965).
19. Graziano, M. S., Andersen, R. A. & Snowden, R. J. Tuning of MST neurons to spiral motions. *J. Neurosci.* **14**, 54–67 (1994).
20. Pack, C. C., Livingstone, M. S., Duffy, K. R. & Born, R. T. End-stopping and the aperture problem: two-dimensional motion signals in macaque V1. *Neuron* **39**, 671–680 (2003).
21. Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu. Rev. Neurosci.* **28**, 157–189 (2005).
22. Mineault, P. J., Khawaja, F. A., Butts, D. A. & Pack, C. C. Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Natl Acad. Sci.* **109**, E972–E980 (2012).
23. Li, K. et al. Neurons in primate visual cortex alternate between responses to multiple stimuli in their receptive field. *Front. Comput. Neurosci.* **10**, 141 (2016).
24. Wertheimer, M. Laws of organization in perceptual forms. In *A sourcebook of gestalt psychology* (ed. Ellis, W.) 71–88 (Harcourt, Brace, 1938).
25. Johansson, G. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211 (1973).
26. Gogel, W. C. Relative motion and the adjacency principle. *Q. J. Exp. Psychol.* **26**, 425–437 (1974).
27. Grossberg, S., Léveillé, J. & Versace, M. How do object reference frames and motion vector decomposition emerge in laminar cortical circuits? *Atten. Percept. Psychophys.* **73**, 1147–1170 (2011).
28. Spelke, E. S. Principles of object perception. *Cognit. Sci.* **14**, 29–56 (1990).
29. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38 (1977).
30. Bishop, C. M. *Pattern recognition and machine learning* (Springer, 2006).
31. Cappé, O. & Moulines, E. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B* **71**, 593–613 (2009).
32. Tanaka, K., Fukada, Y. & Saito, H. Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey. *J. Neurophysiol.* **62**, 642–656 (1989).
33. Flombaum, J. I. & Scholl, B. J. A temporal same-object advantage in the tunnel effect: facilitated change detection for persisting objects. *J. Exp. Psychol. Hum. Perception Perform.* **32**, 840–853 (2006).
34. Gardiner, C. *Stochastic methods*, vol. 4 (Springer Berlin, 2009).
35. Duncker, K. Über induzierte bewegung. *Psychologische Forschung* **12**, 180–259 (1929).
36. Braddick, O. J., Wishart, K. A. & Curran, W. Directional performance in motion transparency. *Vis. Res.* **42**, 1237–1248 (2002).
37. Chen, Y., Meng, X., Matthews, N. & Qian, N. Effects of attention on motion repulsion. *Vis. Res.* **45**, 1329–1339 (2005).

38. Benton, C. P. & Curran, W. Direction repulsion goes global. *Curr. Biol.* **13**, 767–771 (2003).
39. Takemura, H., Tajima, S. & Murakami, I. Whether dots moving in two directions appear coherent or transparent depends on directional biases induced by surrounding motion. *J. Vis.* **11**, 17 (2011).
40. Marshak, W. & Sekuler, R. Mutual repulsion between moving visual targets. *Science* **205**, 1399–1401 (1979).
41. Kim, J. & Wilson, H. R. Direction repulsion between components in motion transparency. *Vis. Res.* **36**, 1177–1187 (1996).
42. Lorenceau, J. Motion integration with dot patterns: effects of motion noise and structural information. *Vis. Res.* **36**, 3415–3427 (1996).
43. Cali, J. N., Bennett, P. J. & Sekuler, A. B. Phase integration bias in a motion grouping task. *J. Vis.* **20**, 31 (2020).
44. Brandt, T., Dichgans, J. & Koenig, E. Differential effects of central versus peripheral vision on egocentric and exocentric motion perception. *Exp. Brain Res.* **16**, 476–491 (1973).
45. Angelaki, D. E., Gu, Y. & DeAngelis, G. C. Visual and vestibular cue integration for heading perception in extrastriate visual cortex. *J. Physiol.* **589**, 825–833 (2011).
46. Shivkumar, S., DeAngelis, G. C. & Haefner, R. M. A causal inference model for the perception of complex motion in the presence of self-motion. *J. Vis.* **20**, 1631 (2020).
47. Amano, K., Wandell, B. A. & Dumoulin, S. O. Visual field maps, population receptive field sizes, and visual field coverage in the human MT+ complex. *J. Neurophysiol.* **102**, 2704–2718 (2009).
48. Wallach, H. & O'connell, D. The kinetic depth effect. *J. Exp. Psychol.* **45**, 205 (1953).
49. Ullman, S. The interpretation of structure from motion. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **203**, 405–426 (1979).
50. Husain, M., Treue, S. & Andersen, R. A. Surface interpolation in three-dimensional structure-from-motion perception. *Neural Comput.* **1**, 324–333 (1989).
51. Treue, S., Husain, M. & Andersen, R. A. Human perception of structure from motion. *Vis. Res.* **31**, 59–75 (1991).
52. Treue, S., Andersen, R. A., Ando, H. & Hildreth, E. C. Structure-from-motion: perceptual evidence for surface interpolation. *Vis. Res.* **35**, 139–148 (1995).
53. Brouwer, G. J. & van Ee, R. Endogenous influences on perceptual bistability depend on exogenous stimulus characteristics. *Vis. Res.* **46**, 3393–3402 (2006).
54. Eby, D. W., Loomis, J. M. & Solomon, E. M. Perceptual linkage of multiple objects rotating in depth. *Perception* **18**, 427–444 (1989).
55. Bradley, D. C., Chang, G. C. & Andersen, R. A. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* **392**, 714–717 (1998).
56. Dodd, J. V., Krug, K., Cumming, B. G. & Parker, A. J. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J. Neurosci.* **21**, 4809–4821 (2001).
57. Brouwer, G. J. & van Ee, R. Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion. *J. Neurosci.* **27**, 1015–1023 (2007).
58. Wasmuht, D., Parker, A. & Krug, K. Interneuronal correlations at longer time scales predict decision signals for bistable structure-from-motion perception. *Sci. Rep.* **9**, 1–15 (2019).
59. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
60. Salinas, E. & Abbott, L. F. A model of multiplicative neural responses in parietal cortex. *Proc. Natl Acad. Sci.* **93**, 11956–11961 (1996).
61. Dayan, P. & Abbott, L. F. *Theoretical neuroscience: computational and mathematical modeling of neural systems* (Computational Neuroscience Series, 2001).
62. Groschner, L. N., Malis, J. G., Zuidinga, B. & Borst, A. A biophysical account of multiplication by a single neuron. *Nature* **603**, 119–123 (2022).
63. Gerstner, W. & Kistler, W. M. *Spiking neuron models: single neurons, populations, plasticity* (Cambridge University Press, 2002).
64. Komatsu, H. & Wurtz, R. H. Relation of cortical areas MT and MST to pursuit eye movements. I. Localization and visual properties of neurons. *J. Neurophysiol.* **60**, 580–603 (1988).
65. Duffy, C. J. & Wurtz, R. H. Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *J. Neurophysiol.* **65**, 1329–1345 (1991).
66. DeAngelis, G. C. & Uka, T. Coding of horizontal disparity and velocity by MT neurons in the alert macaque. *J. Neurophysiol.* **89**, 1094–1111 (2003).
67. Nover, H., Anderson, C. H. & DeAngelis, G. C. A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *J. Neurosci.* **25**, 10049–10060 (2005).
68. Kohn, A. & Movshon, J. A. Adaptation changes the direction tuning of macaque MT neurons. *Nat. Neurosci.* **7**, 764–772 (2004).
69. Krekelberg, B., Van Wezel, R. J. & Albright, T. D. Interactions between speed and contrast tuning in the middle temporal area: implications for the neural code for speed. *J. Neurosci.* **26**, 8988–8998 (2006).
70. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
71. Friston, K. Learning and inference in the brain. *Neural Netw.* **16**, 1325–1352 (2003).
72. Walsh, K. S., McGovern, D. P., Clark, A. & O'Connell, R. G. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann. N.Y. Acad. Sci.* **1464**, 242–268 (2020).
73. Millidge, B., Seth, A. & Buckley, C. L. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979* (2022).
74. Bastos, A. M. et al. Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
75. Mikulasch, F. A., Rudelt, L., Wibral, M. & Priesemann, V. Dendritic predictive coding: A theory of cortical computation with spiking neurons. *arXiv preprint arXiv:2205.05303* (2022).
76. Castet, E., Lorenceau, J., Shiffrar, M. & Bonnet, C. Perceived speed of moving lines depends on orientation, length, speed and luminance. *Vis. Res.* **33**, 1921–1936 (1993).
77. Allman, J., Miezin, F. & McGuinness, E. Direction-and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception* **14**, 105–126 (1985).
78. Huang, X., Albright, T. D. & Stoner, G. R. Stimulus dependency and mechanisms of surround modulation in cortical area MT. *J. Neurosci.* **28**, 13889–13906 (2008).
79. Nawrot, M. & Sekuler, R. Assimilation and contrast in motion perception: explorations in cooperativity. *Vis. Res.* **30**, 1439–1451 (1990).
80. Pastukhov, A. First, you need a Gestalt: an interaction of bottom-up and top-down streams during the perception of the ambiguously rotating human walker. *Sci. Rep.* **7**, 1158 (2017).
81. Angelaki, D. E., Gu, Y. & DeAngelis, G. C. Multisensory integration: psychophysics, neurophysiology, and computation. *Curr. Opin. Neurobiol.* **19**, 452–458 (2009).
82. Takahashi, K. et al. Multimodal coding of three-dimensional rotation and translation in area MSTd: comparison of visual and vestibular selectivity. *J. Neurosci.* **27**, 9742–9756 (2007).
83. Ventre-Dominey, J. Vestibular function in the temporal and parietal cortex: distinct velocity and inertial processing pathways. *Front. Integr. Neurosci.* **8**, 53 (2014).
84. Chowdhury, S. A., Takahashi, K., DeAngelis, G. C. & Angelaki, D. E. Does the middle temporal area carry vestibular signals related to self-motion? *Journal of Neuroscience* **29**, 12020–12030 (2009).

85. Rideaux, R. & Welchman, A. E. But still it moves: static image statistics underlie how we see motion. *J. Neurosci.* **40**, 2538–2552 (2020).
86. Kalman, R. E. & Bucy, R. S. New results in linear filtering and prediction theory. *J. Basic Eng.* **83**, 95–108 (1961).
87. Kutschireiter, A., Surace, S. C. & Pfister, J.-P. The hitchhiker's guide to nonlinear filtering. *J. Math. Psychol.* **94**, 102307 (2020).
88. Bill, J., Gershman, S. J. & Drugowitsch, J. Code for the publication: visual motion perception as online hierarchical inference. *GitHub*, https://doi.org/10.5281/zenodo.7152982 (2022).
89. Qian, N., Andersen, R. A. & Adelson, E. H. Transparent motion perception as detection of unbalanced motion signals. I. Psychophysics. *J. Neurosci.* **14**, 7357–7366 (1994).
90. Gershman, S. J., Vul, E. & Tenenbaum, J. Perceptual multistability as Markov chain Monte Carlo inference. In *Advances in neural information processing systems* (NeurIPS, 2009).

Acknowledgements

We thank Anna Kutschireiter for valuable discussions and feedback on the theory. This research was supported by grants from the NIH (NINDS U19NS118246, J.D.), the James S. McDonnell Foundation (Scholar Award for Understanding Human Cognition, Grant 220020462, J.D.), the Harvard Brain Science Initiative (Collaborative Seed Grant, J.D. & S.J.G.), and the Center for Brains, Minds, and Machines (CBMM; funded by NSF STC award CCF-1231216, S.J.G.).

Author contributions

J.B., S.J.G., and J.D. conceived the study; J.B. developed the theory; J.B. performed the computer simulations; J.B. and J.D. analyzed and discussed the data; J.B., S.J.G., and J.D. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34805-5.

Correspondence and requests for materials should be addressed to Johannes Bill.

Peer review information *Nature Communications* thanks Alexander Pastukhov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

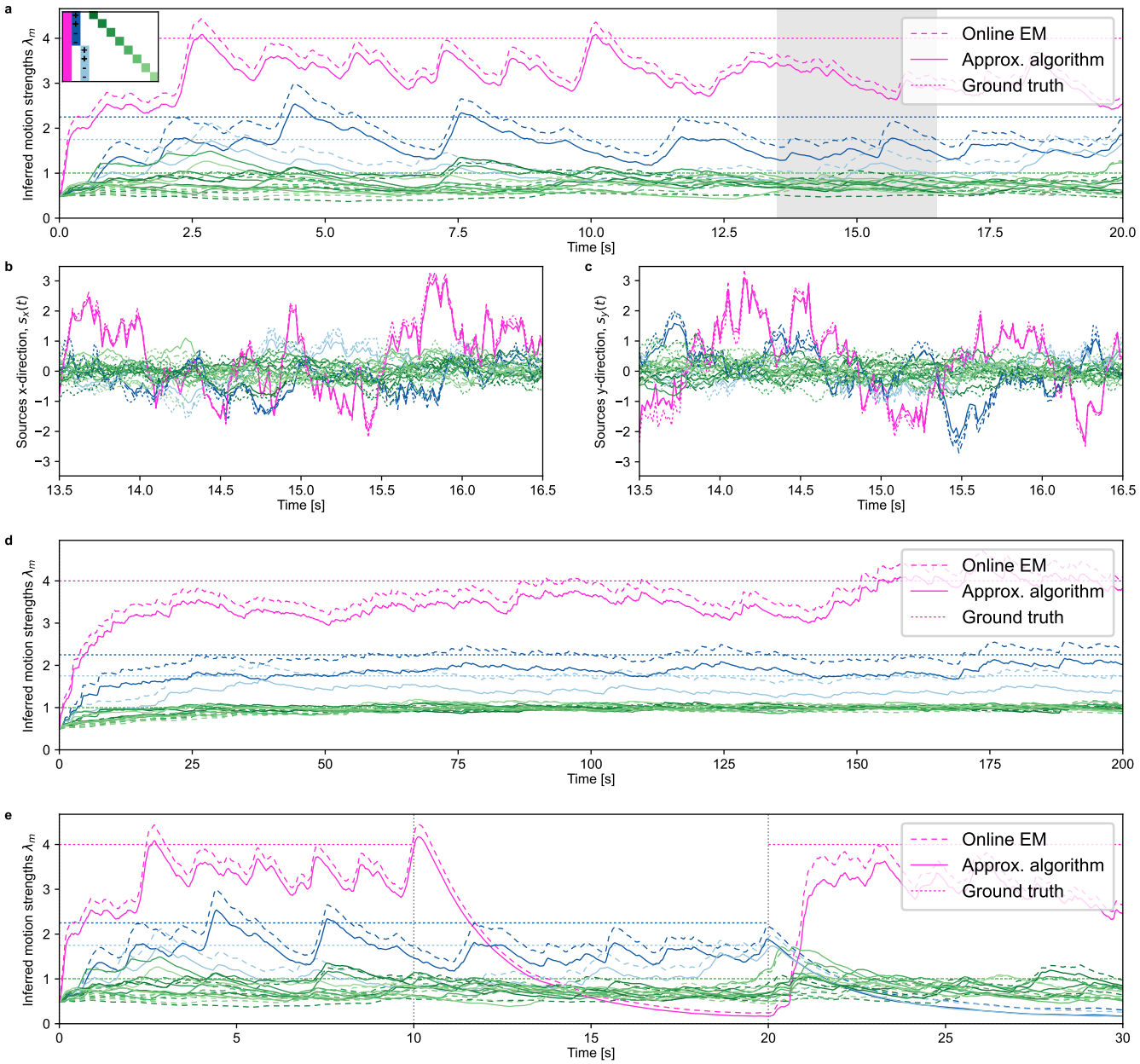
Supplementary Information

Visual motion perception as online hierarchical inference

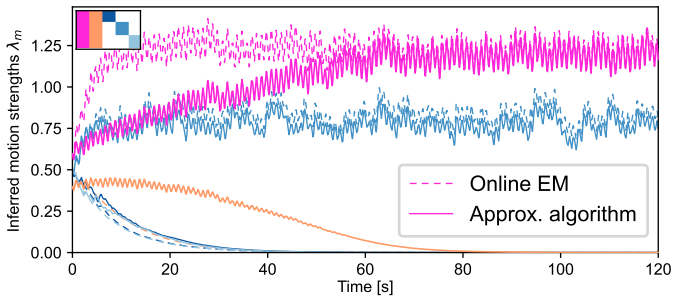
Johannes Bill, Samuel J. Gershman, Jan Drugowitsch

Contents

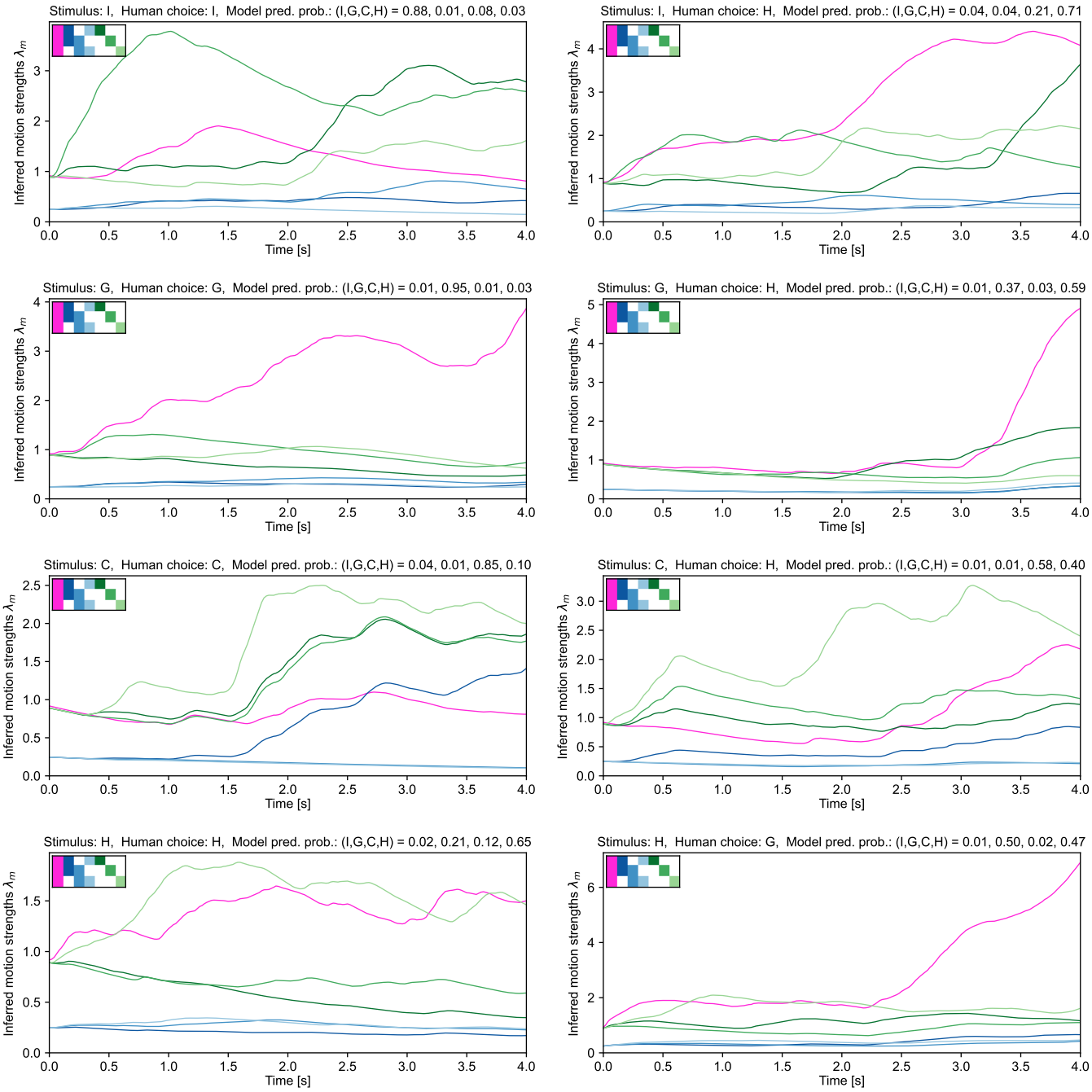
Supplementary figures



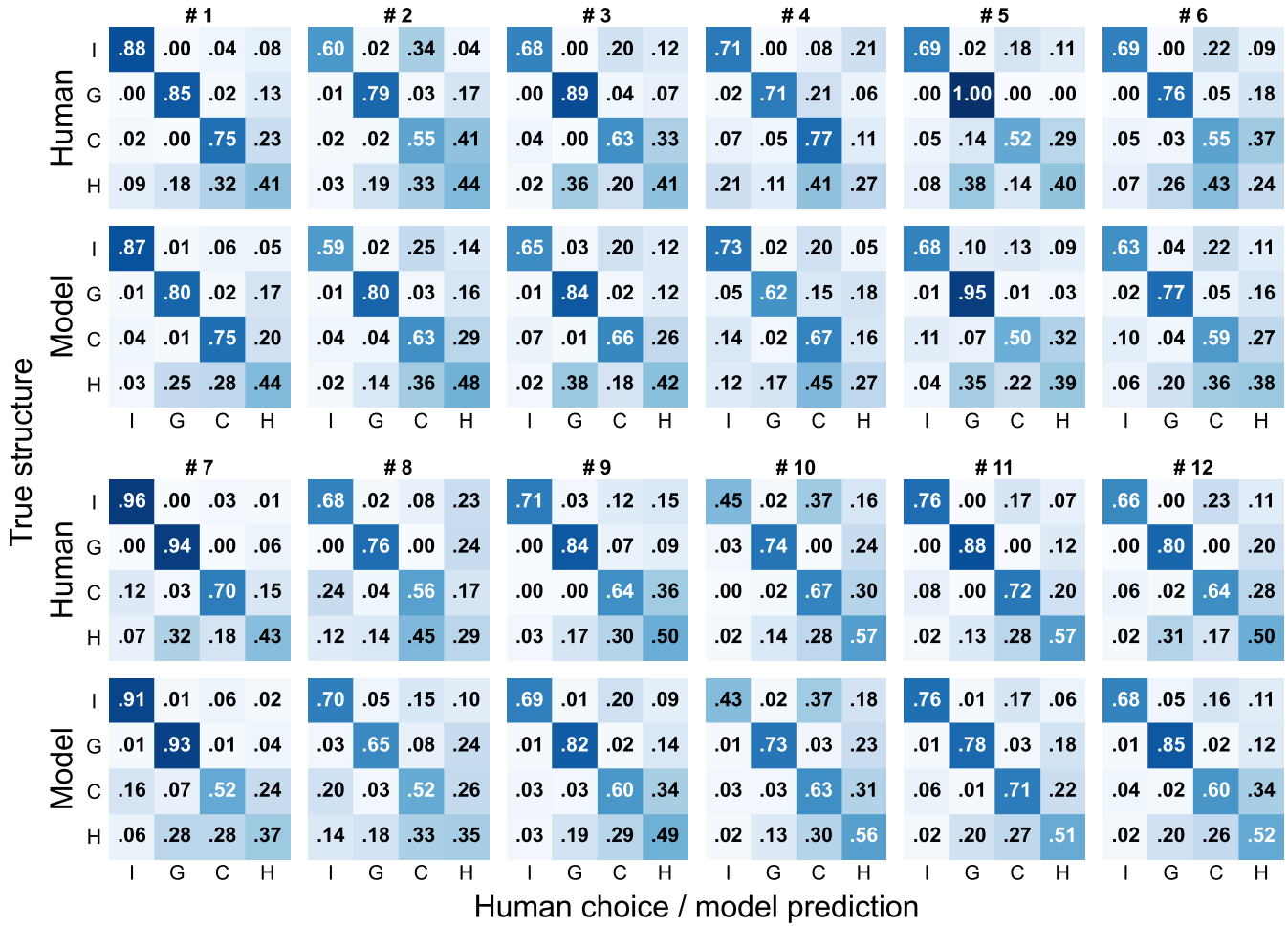
Supplementary Fig. 1 | The online inference model correctly recovers the structure and motion sources of presented input. (a) Inferred motion strengths by the model. Two-dimensional input was generated from the generative model for a deeply nested structure with shared motion (pink; $\lambda=4$), two separate groups of counter-rotating sub-groups (dark- and light-blue; $+/-$ in the inset indicates $C_{km}=+1/-1$; $\lambda=2.25$ and 1.75), and eight individual motions (greens; $\lambda=1$). Other parameters are the default parameters for object-indexed experiments (see table in the *Methods* section of the main paper). Shown are the inferred strengths for the online inference model (solid lines; labeled "Approx. algorithm", given by eqn. (1)−(3) of the main text), the more accurate, but computationally also more complex online EM algorithm (dashed lines; given by eqn. (29), (30), and (35) of the Supplementary Information), and the ground truth (dotted lines). The approximate algorithm yields results similar to the reference online EM algorithm. Both algorithms underestimate the motion strengths due to the sparsity prior $p(\lambda^2)$. **(b)** Inferred motion sources, x-direction for the highlighted duration of the simulation in panel a. Same color key as in panel a. **(c)** Same as panel b, but for the y-direction. **(d)** Repetition of the simulation in panel a, but with 10x longer time constant τ_λ, longer run time, and uniform prior over the motion strengths. The underestimation in the reference algorithm vanishes; the approximate algorithm maintains its approximation quality. **(e)** Repetition of the simulation in panel a, but with a temporally changing structure. After 10 s, the shared component is switched off in the input. After 20 s, the shared component is re-introduced, but the groups are switched off. Both inference algorithms successfully detect these changes. Source data are provided as a Source Data file.



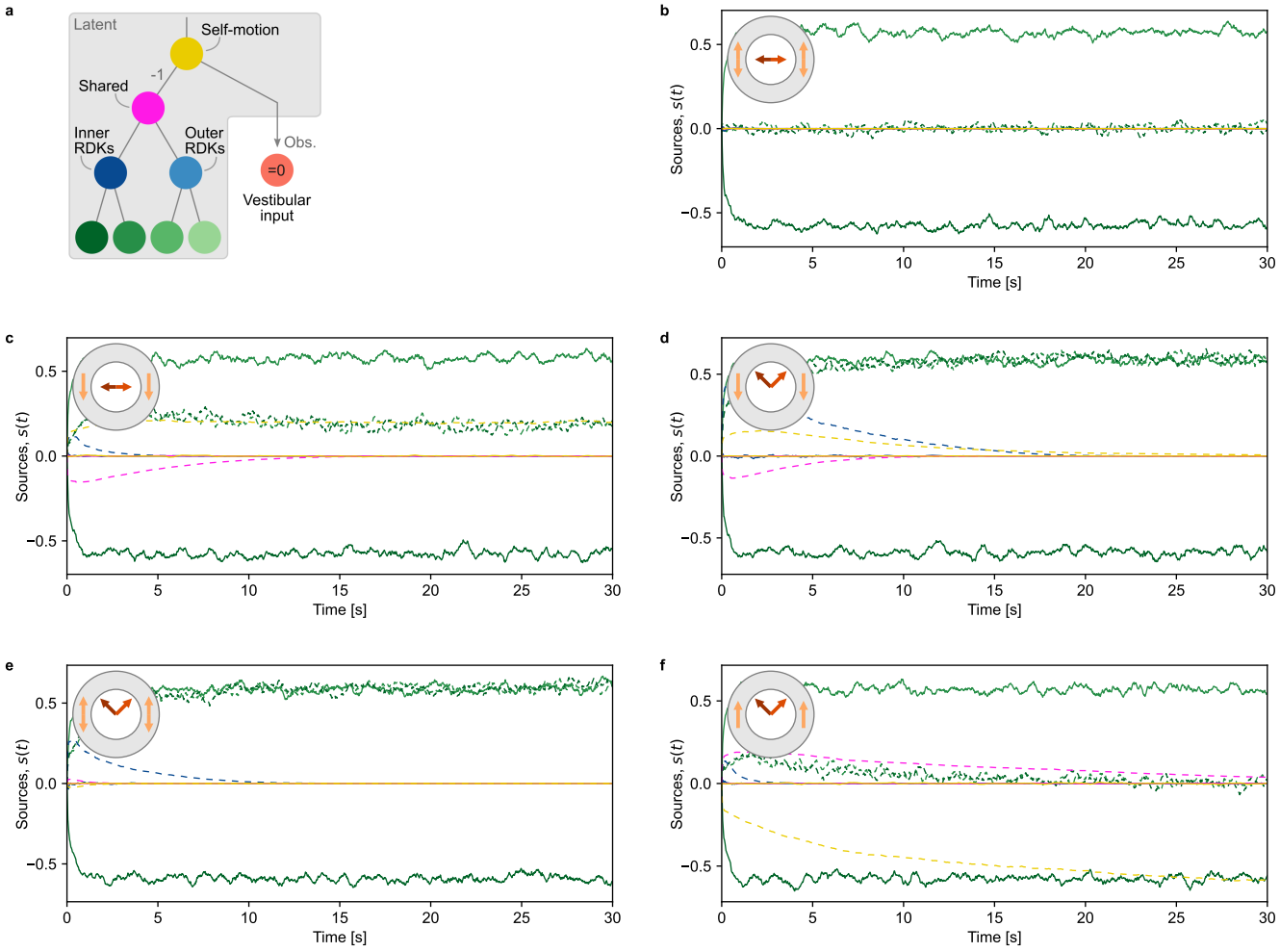
Supplementary Fig. 2 | The online model prefers simple structures, recruiting only necessary components from the reservoir. Shown is a repetition of the Johansson experiment from Figure 2c, yet with a duplicated shared motion component in the observer model (pink and orange, see inset in the top-left). A small difference at initialization ($t=0$) between the two components widens, such that eventually only one component is recruited and the other one is dismissed. This preference for simpler structures is a direct consequence of the sparsity-inducing Jeffreys prior. If a uniform prior had been used, both shared components would have been maintained (not shown). Furthermore, we notice that the reference online EM algorithm converges more rapidly than the approximate algorithm. The reason is found in the posterior covariance matrix, $\mathbf{\Sigma}$, which is fully computed for online EM according to eqn. (29) and in which the off-diagonal element between the two shared components introduces competition during the credit assignment in eqn. (39) (uncertainty in the two sources is negatively correlated, leading to a negative matrix element in $\mathbf{\Sigma}$). Source data are provided as a Source Data file.



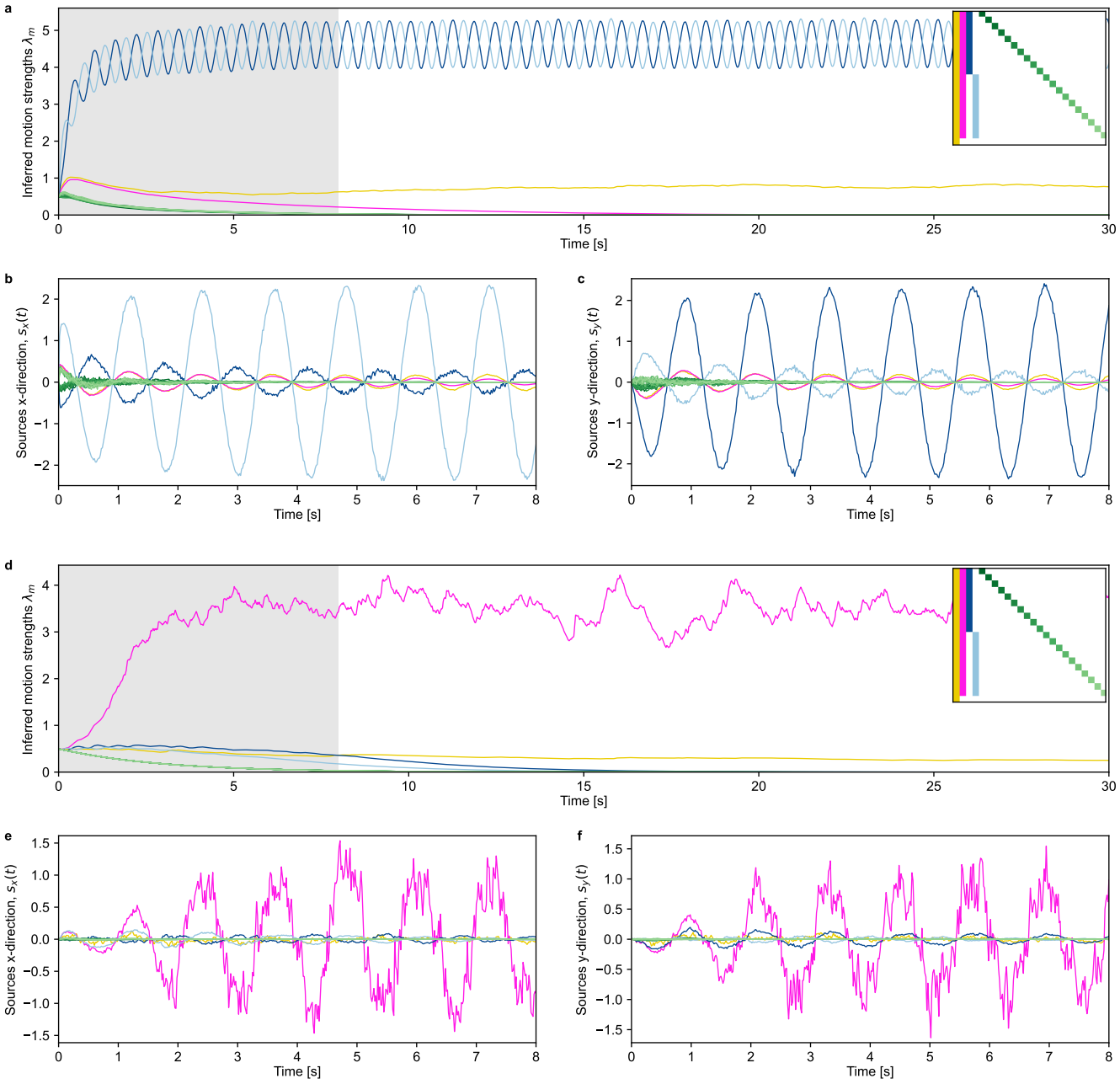
Supplementary Fig. 3 | Examples of motion structure inference for trials from (Yang et al., 2021). Shown are traces for $\lambda(t)$ for eight example trials of participant # 1. Axes titles state the ground truth, the participant's classification, and the predicted choice probabilities of the model. *Left column:* Trials of each structure which were correctly classified by the human participant. *Right column:* Trials of each structure which were incorrectly classified by the human participant. Source data are provided as a Source Data file.



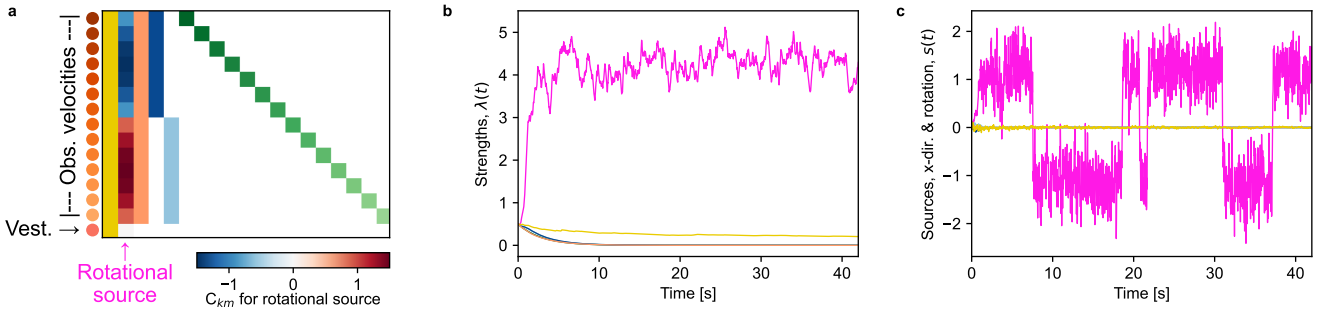
Supplementary Fig. 4 | The online model captures participant-specific error patterns in the data from (Yang et al., 2021). Shown are the confusion matrices for all 12 participants along with the cross-validated predictions of our model. The model captures participant-specific patterns, such as general performance levels; the preferential misclassification of hierarchical motion (H) as either global (G) or clustered (C); and the asymmetry between the I-C and C-I elements. Source data are provided as a Source Data file.



Supplementary Fig. 5 | Inferred motion sources for the center-surround interaction MDR experiment by (Takemura et al., 2011). (a) Used motion tree with four hierarchy levels. **(b)–(f)** Example trials for each of the experiment conditions from Figure 4h–l of the main text. Shown is the evolution of motion sources (solid lines: x-direction, dashed lines: y-direction, colors as in panel (a)). For visual clarity, only the "inner branch" (left half of the tree) is shown, and traces were smoothed with a 500 ms box filter for plotting (the histograms in Figure 4 are based on the non-smoothed data). The perceived directions in Figure 4 are the sum of the s_m's at trial end, without self-motion. Source data are provided as a Source Data file.



Supplementary Fig. 6 | Detailed evolution of all motion strengths and sources in the illusion from Lorenceau. (a) Motion strengths, $\lambda(t)$, in the low-noise condition. Motion is decomposed into a deep hierarchy of self-motion (yellow), shared motion (magenta), group motion (dark and light blue), and individual motion of all $K=20$ dots (greens). In the generative model, self-motion affects the velocities of all dots and of the zero-mean vestibular input (lowest row in the C-matrix inset). **(b)** Inferred x-direction, $\mu_x(t)$, of all motion sources in the low-noise condition, for the highlighted duration in panel a. **(c)** Same as panel b, but for the y-direction. **(d–f)** Same as panels a–c, but for the high-noise condition. Source data are provided as a Source Data file.



Supplementary Fig. 7 | A rotating cylinder is perceived for SfM also when more motion components are available. (a) Component matrix with additional motion components. From left to right: Translational self-motion, rotational motion around vertical axis, translational motion for all visual inputs, translational motion for all dots on the back of the cylinder, translational motion for all dots on the front of the cylinder, translational motions for each individual observed velocity. All translational motions have $\|C_{km}\|=1$ or 0, and their colors indicate the line colors in panels (b) and (c). For the rotational motion component, the color map shows the value for the v_x-direction, i.e., $v_{x,k}(t)=C_{k,2}\,s^{\text{rot}}(t)+$ contributions of other sources. The locations of the observed velocities lie on the cylinder starting on the right and then ascending in CCW direction when viewed from the top (cf. Figure 6b of the main text). Thus, $s^{\text{rot}}>0$, i.e., CCW rotation, leads to $v_x<0$ for locations at the back of the cylinder. The rotational source, s^{rot}, is shown in magenta in panels (b) and (c) **(b)** Identified motion strengths. Even with the richer set of available components, the model identifies only the rotation (magenta). **(c)** Inferred motion sources. The bi-stability of the perceived rotation remains unaffected by the richer available structure (cf. Figure 6d of the main text; same random seed used for observation noise and stochastic time points of the assignment process). Source data are provided as a Source Data file.

List of used variables

Variable	Description	Variable	Description
$\boldsymbol{v} = (v_1,..,v_K)^\mathsf{T}$	Observable velocity	$\boldsymbol{s} = (s_1,..,s_M)^\mathsf{T}$	Latent motion source
$v_k = v_{k,d}(t)$	Dim. d and time t often suppressed	$s_m = s_{m,d}(t)$	Dim. d and time t often suppressed
$\hat{\boldsymbol{v}} = (\hat{v}_1,..,\hat{v}_K)^\mathsf{T}$	Noise-free velocity	$\boldsymbol{\lambda} = (\lambda_1,..,\lambda_M)^\mathsf{T}$	Motion strength
$\boldsymbol{\mu} = (\mu_1,..,\mu_M)^\mathsf{T}$	Mean vector in $\boldsymbol{s}$-posterior	$\boldsymbol{\epsilon} = (\epsilon_1,..,\epsilon_K)^\mathsf{T}$	Prediction error
$\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$	Covariance in $\boldsymbol{s}$-post. and inv. precision	f_Σ	Adiabatic approx. of σ^2
$\boldsymbol{\sigma}^2$	Vect. of diag. elements $(\Sigma_{11},..,\Sigma_{MM})^\mathsf{T}$	$\boldsymbol{C}$	Component matrix of shape $(K \times M)$
D	no. dimensions	$\boldsymbol{c}_m$	m-th column of $\boldsymbol{C}$
$\mathrm{diag}[\boldsymbol{x}]$	Diag. matrix over some vec. $\boldsymbol{x}$	$\langle f(x) \rangle_{p(x)}$	Expectation of $f(x)$ under $p(x)$
τ_s	Time const. for s-inference (OU process)	$\boldsymbol{J}$	Interaction prior on motion comp.
τ_λ	Time const. for λ-inference	δt	Obs. interval (inverse frame rate)
τ_ϵ	Time const. for pred. err.	$\delta(t)$, δ_{ij}	Dirac and Kronecker delta
t_{obs}	Obs. time point in cont.-time (CT)	∂_t	Partial derivative (here, w.r.t. t)
t	No. time steps (DT); Or: time (CT)	$\boldsymbol{\nu}$	No. of $\boldsymbol{\mu}$-pseudo obs. (hyperprior)
$\mathcal{I}\chi(\nu,\kappa^2)$	Scaled inverse chi-squared distribution	$\boldsymbol{\kappa}$	Val. of pseudo observations
σ_ρ	MT width of speed tuning	κ_α	MT precision of direction tuning
σ_{obs}	Observation noise (δt-independent)	$\mathcal{Q}(\boldsymbol{\lambda})$	EM expect. compl. data log-likelihood
$\widetilde{\sigma}_{\mathrm{obs}}$	Disc.-time obs. noise ($=\sigma_{\mathrm{obs}}/\sqrt{\delta t}$)	S	Structure I, G, C, H in (Yang et al., 2021)
R	Radial rec. field loc.	β	Inv. temp. in (Yang et al., 2021)
ϑ	Angular rec. field location	b_G, b_C, b_H	Biases in (Yang et al., 2021)
s_r, s_φ	Polar sources: radial and angular velocity	π_{L}	Lapse prob. in (Yang et al., 2021)
ρ	Speed in MT tuning func.	α	Direction (angle) in MT tuning
$n_\rho, n_\alpha, N_\rho, N_\alpha$	MT neuron indices, max. values	$\mu_\rho(n_\rho)$, $\mu_\alpha(n_\alpha)$	Tuning center of neuron (n_ρ, n_α)
ψ	MT max. firing rate multiplier	$I_n(\kappa)$	Modified Bessel function of order n
$f, f_\sigma, f_\rho, f_\alpha$	MT tuning func. and sub-functions	$\boldsymbol{I}$	Identity matrix
$\boldsymbol{W}, \overline{\boldsymbol{W}}$	Linear maps in algo. & neural domain	$\underline{\boldsymbol{Q}}, \overline{\underline{\boldsymbol{Q}}}$	Quad. maps in algo. & neural domain
$\boldsymbol{b}, \overline{\boldsymbol{b}}$	Add. constants in algo. & neural domain	$\boldsymbol{A}, \boldsymbol{A}^\dagger$	Linear decoding matrix & adjoint
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	Variables in generic network derivation	x_k, y_k	Spatial locations for 2-dim. stimuli
$\boldsymbol{r}_{\mathrm{inp}}, \boldsymbol{r}_{\mathrm{dis}}, \boldsymbol{r}_{\text{1-to-1}}$	Firing rates in the network model	γ	Direction repulsion opening angle
v_{vst}	Vestibular self-motion input	$\dot{\varphi}$	Angular (rotational) velocity
σ_{vst}	Obs. noise for vestibular input	$\dot{r}$	Radial velocity
η_C	Learning rate for $\boldsymbol{C}$-learning		

Supplementary Table 1 | List of used variables. In a few cases, we chose to accept notation clashes when confusion is ruled out and the dual use actually facilitates clarity. These cases are indicated with "Or:" in the description.

Theory of the online hierarchical inference model and the neural network model

In **Supplementary Note 1**, we introduce the generative model for structured motion. In **Supplementary Note 2**, we derive the online hierarchical inference model. In **Supplementary Note 3**, we present optional extensions to the inference model. In **Supplementary Note 4**, we derive the recurrent neural network model.

Supplementary Note 1. Generative model of structured motion

The following model of hierarchically structured motion is an adaptation of the generative model from Bill et al. [1]. We consider K observable velocities, $v_{k,d}(t)$, in D spatial dimensions. To prevent clutter, we will develop most of the theory for the one-dimensional case, $D=1$, and use the vector notation, $\boldsymbol{v}=(v_1,..,v_K)^\mathsf{T}$. The extension to $D>1$ is straightforward and will be covered in a dedicated subsection. We will often suppress the explicit time dependence when confusion is ruled out. In other cases, we may write time as an index, $\boldsymbol{v}_t$, when a compact notion is desirable. Observable velocities, $\boldsymbol{v}$, are noisy instantiations of noise-free velocities, $\hat{\boldsymbol{v}}$, which are generated by M latent motion sources, $s_{m,d}(t)$, as will be specified below. Similar to velocities, we abbreviate $\boldsymbol{s}=(s_1,..,s_M)^\mathsf{T}$.

1.1 Composition of observable velocity from motion motifs

For most of this work, we restrict the influence of motion source $s_m(t)$ on velocity $\hat{v}_k(t)$ to three particularly basic relations: the motion of s_m can affect $\hat{v}_k$ either positively (e.g., the latent flock motion on an observed bird's velocity), negatively (e.g., the effect of self-motion on the observed scene), or not at all (e.g., the flock motion's effect on a tree). Formally, we describe these influences in a $K\times M$ component matrix $\boldsymbol{C}$, with $C_{km}=+1$, -1, and 0 for positive, negative and absent influence, respectively. We will sometimes refer to the columns of $\boldsymbol{C}$ as motion motif or motion component $\boldsymbol{c}_m$. Typically, we have $M > K$ because every observable can have its individual motion component which exclusively affects this observable.

The overall velocity $\hat{v}_k$ is the sum of all the motion sources' contributions:

$$\hat{\boldsymbol{v}} = \boldsymbol{C}\,\boldsymbol{s}\ . \tag{1}$$

It is easy to see that all tree-like hierarchies can be cast into this form. As a word of warning, the opposite is not true: not every possible $\boldsymbol{C}$ represents hierarchically structured motion.

Finally, it is worth mentioning that motion components are not necessarily limited to the simple $C_{km}=\pm 1$ values. For modeling gradual relationships, a motion source could affect some observables stronger than others. For instance, the current of a river could influence the water's velocity in the middle of the river stronger than close to the bank. All subsequent derivations likewise hold for such gradual relationships, as long as they influence the observed velocities linearly.

1.2 Generation of observable velocities from stochastic, latent motion sources

For the generative model, we assume that the motion sources, $s_m(t)$, evolve independently according to an Ornstein-Uhlenbeck process, see, e.g., [2],

$$\mathrm{d}s_m = -\tfrac{1}{\tau_s} s_m\,\mathrm{d}t + \lambda_m\,\mathrm{d}W_m\ , \tag{2}$$

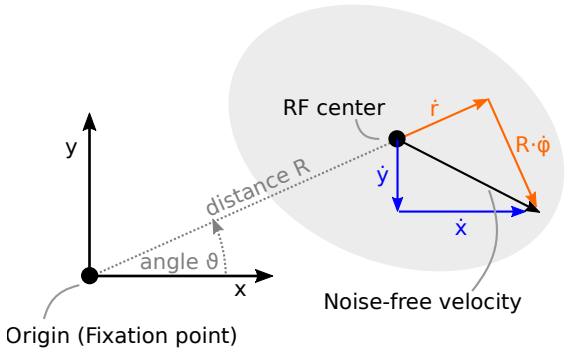
where Wiener process $W_m(t)$ drives changes in s_m via random forces, motion strength $\lambda_m \geq 0$ controls the magnitude of these forces, and $\tau_s > 0$ is the typical time constant of significant changes in s_m. The equilibrium distribution of this process is a normal distribution with zero mean and variance $\frac{\tau_s}{2}\lambda_m^2$:

$$\lim_{t\to\infty} p(s_m(t)) = \mathcal{N}\left(s_m; 0, \tfrac{\tau_s}{2}\lambda_m^2\right)\ . \tag{3}$$

The Ornstein-Uhlenbeck process is a suitable first-order approximation for modeling real-world motion as it (i) generates continuous trajectories $s_m(t)$, (ii) includes a notion of inertia/momentum via its temporal filtering with τ_s, (iii) offers an intuitive parameterization by scaling typical velocities linearly in λ_m, and (iv) implements a slow-velocity prior via a normal distribution like in Ref. [3].

Observable velocities, $\boldsymbol{v}(t)$, are generated by composing the latent motion sources to noise-free velocities according to eqn. (1) and then applying independent Gaussian noise:

$$\boldsymbol{v}(t) \sim \mathcal{N}\left(\boldsymbol{C}\,\boldsymbol{s}(t), \tilde{\sigma}_{\mathrm{obs}}^2\,\boldsymbol{I}\right)\ . \tag{4}$$



$$\dot{x} = \dot{r}\,\cos\vartheta - \dot{\varphi}\,R\sin\vartheta$$
$$\dot{y} = \dot{r}\,\sin\vartheta + \dot{\varphi}\,R\cos\vartheta$$

$$\dot{r} = \dot{x}\,\cos\vartheta + \dot{y}\,\sin\vartheta$$
$$\dot{\varphi} = \frac{1}{R}\left(-\dot{x}\,\sin\vartheta + \dot{y}\,\cos\vartheta\right)$$

Supplementary Fig. 8 | Coordinate transformation from polar motion sources to Cartesian velocities for supporting rotational and radial motion motifs. When receptive field (RF) centers are fixed, rotational motion, $\dot{\varphi}$, and radial motion, $\dot{r}$, feature fixed linear relations to the resulting Cartesian velocities in x- and y-direction. The angles ϑ and φ are measured relative to the x-axis in counter-clockwise direction by convention. Cartesian components are measured in rightward (x) and upward (y) direction.

In foresight of the continuous-time formulation, we denote the observation noise by $\widetilde{\sigma}_{\text{obs}}$ (with a tilde). It will later be adjusted to the frame rate of incoming observations such that the information per unit time remains constant. To keep the derivation tractable, we will ignore the reported velocity-dependence of observation noise in human perception (Weber's law) and treat $\widetilde{\sigma}_{\text{obs}}$ as a constant.

Marginalizing over the stationary distribution of the latent motion sources, eqn. (3), the observable velocities are jointly normally distributed due to their linear dependence on a Gaussian origin, with zero mean and covariance matrix $\frac{\tau_s}{2}\boldsymbol{C}\,\text{diag}\left[\boldsymbol{\lambda}^2\right]\boldsymbol{C}^\mathsf{T} + \widetilde{\sigma}^2_{\text{obs}}\boldsymbol{I}$, with $\text{diag}\left[\boldsymbol{\lambda}^2\right]$ denoting the diagonal matrix generated from vector $\boldsymbol{\lambda}^2 = (\lambda_1^2, .., \lambda_M^2)^\mathsf{T}$.

1.3 Motion structure

The motion strengths $\boldsymbol{\lambda}$ play a particularly important role in the generative process described by eqn. (2)+(4). For $\lambda_m = 0$, dependent motion sources decay to zero, i.e., $s_m \to 0$. Hence, motion strengths describe the presence ($\lambda_m > 0$) or absence ($\lambda_m = 0$) of motion components, as well as their typical magnitude ($\langle |s_m| \rangle \propto \lambda_m$). In other words, given a reservoir of components, $\boldsymbol{C}$, which have been learned to occur in visual scenes in general, the vector $\boldsymbol{\lambda}$ will describe the structural composition of motion relations in a specific visual scene. Knowing $\boldsymbol{\lambda}$ therefore is equivalent to knowing the motion structure of the scene.

1.4 Extension to multiple spatial dimensions

For $D > 1$, we will usually assume that all $s_{m,d}$ with the same index m share motion strength λ_m and motion component $\boldsymbol{c}_m$, yet each $s_{m,d}$ follows its own stochastic evolution, i.e., has its own Wiener process, $W_{m,d}$, in eqn. (2). This choice reflects that space is isotropic: due to the Gaussianity of each $s_{m,d}$, the joint distribution $p(s_{m,1}, .., s_{m,D})$ will be a multivariate Gaussian with covariance matrix $\frac{\tau_s}{2}\lambda_m^2\boldsymbol{I}$, with $\boldsymbol{I}$ denoting the identity matrix, and is, thus, invariant to rotations in the experimenter's choice of the coordinate system. Again, this is a useful first-order approximation, even though human motion perception has been reported to be not perfectly isotropic [4, 5].

In summary, the generative model of structured motion is characterized by the following set of parameters: number of spatial dimensions D, motion strengths $\boldsymbol{\lambda}$, component matrix $\boldsymbol{C}$, time constant τ_s, and observation noise $\widetilde{\sigma}_{\text{obs}}$. Some extensions of this model covering heterogeneous time constants, non-isotropic observation noise and missing observations will be discussed in dedicated subsections in the context of online inference, below.

1.5 Polar coordinates: rotational and radial motion

The discussion so far assumed that motion sources affect velocities in a translational (Cartesian) manner, that is, by adding the same vector to all dependent velocities. An important exception are rotational and radial motion in two dimensions ($D = 2$) which typically occur in flow field parsing (when moving forward, everything expands on the retina; when tilting your head, everything rotates in the opposite rotational direction).

It turns out that rotational and radial motion sources can be incorporated into our framework of linear velocity generation as per eqn. (1), when the spatial locations of all dependent velocities are fixed. This is, for example, fulfilled in experimental setups using drifting gratings: the drifting grating has motion energy, but its location stays fixed

at the same spot. Note how this setup is different from the perspective commonly taken in physics where velocity entails changes in location. The perspective we take here is geared towards studying brain computation where a group of neurons often processes motion in a certain, fixed area of the inputs. Specifically, this is satisfied in many visual areas including V1, MT, and MST, where neurons feature a fixed spatial receptive field in retinal coordinates. The index k in v_k then refers to a fixed location on the retina. We therefore call this condition "location-indexed".

To derive the generative model for rotational and radial motion in a location-indexed experiment, consider the spatial receptive fields illustrated in **Supplementary Fig. 8**. Each receptive field (RF) has a fixed center relative to the fovea and is characterized by radial distance R and angle ϑ, which is measured relative to the x-axis in counter-clockwise direction by convention. The generative model can use latent motion sources in radial direction, denoted by $\dot{r}$ in the figure, and in angular direction, denoted by $\dot{\varphi}$. These polar motion sources generate velocities $\dot{x}$ and $\dot{y}$ in Cartesian space. Importantly, the (generative) transformation from polar to Cartesian coordinates is linear for each receptive field:

$$\begin{aligned} \dot{x} &= \dot{r}\cos\vartheta - \dot{\varphi}\,R\sin\vartheta \\ \dot{y} &= \dot{r}\sin\vartheta + \dot{\varphi}\,R\cos\vartheta \ . \end{aligned} \tag{5}$$

Consequently, the generative process can be included into the component matrix $\boldsymbol{C}$ when each polar motion source affects dependent observables, $\hat{\boldsymbol{v}}_k$, in both spatial dimensions. In contrast to the previously discussed translational motion sources, rotational and radial motion sources may maintain separate motion components $\boldsymbol{c}_m$ and motion strengths λ_m:

$$\begin{aligned} c_{k,m=\text{rad},d=x} &= \cos\vartheta_k \ , & c_{k,m=\text{rot},d=x} &= -R_k\sin\vartheta_k \\ c_{k,m=\text{rad},d=y} &= \sin\vartheta_k \ , & c_{k,m=\text{rot},d=y} &= +R_k\cos\vartheta_k \ . \end{aligned} \tag{6}$$

As before, the model adds up the velocities from different sources and applies the observation noise only to the final velocity (in Cartesian space). The example of rotational and radial motion demonstrates how spatial dimensions can be mixed, as long as the coordinate transformation is linear for every observable $\hat{\boldsymbol{v}}_k$.

As a remark, the released Python code does not support separate motion strengths for rotational and radial motion, for technical reasons. Thus, in simulations containing shared rotation and shared expansion about the fovea, both motion sources share the same strength.

Rotation in 3D and structure-from-motion. The above addresses rotations around a pivot point on the fovea, i.e., the rotation happens in the same plane as the observations. A different form of rotational motion, which is equally covered by the generative model, is used for the structure-from-motion experiments in Figure 6 of the main text. Here, the rotation is around an axis that lies in the observation plane (e.g., the y-axis in the x-y-plane) while the motion happens in 3D, including a depth component, $\hat{v}_{k,z}$. As illustrated in Figure 6b of the main text, rotation (and also expansion around the axis) lead to a linear relation between the rotational motion source, s_t^{rot}, and the noise-free velocity vectors, $\hat{\boldsymbol{v}}_t$, at every fixed location in location-indexed experiments. As before, noise is added in Cartesian coordinates. Note that in the SfM displays of the main text, no depth information is presented in order to make the stimuli ambiguous with regard to their direction of rotation.

Supplementary Note 2. Online hierarchical inference

In **Supplementary Note 2, Section 1**, we develop the hierarchical inference algorithm in a discrete-time, batch formulation, using the Expectation-Maximization (EM) algorithm. In **Supplementary Note 2, Section 2**, we draw the continuous-time limit to obtain an online algorithm. Finally, in **Supplementary Note 2, Section 3**, we introduce an adiabatic approximation which reduces the required computations to neuron-friendly operations. This is the online model underlying all simulations of the main paper.

2.1 Inference via the Expectation-Maximization algorithm

Our goal for motion structure inference is to simultaneously infer the value of motion sources $\boldsymbol{s}(t)$ and the underlying structure $\boldsymbol{\lambda}$ from a stream of observations $\boldsymbol{v}_{1:t} = (\boldsymbol{v}_1, .., \boldsymbol{v}_t)$ with observations arriving at time steps of duration δt (inverse frame rate). The number of spatial dimensions D, components $\boldsymbol{C}$, time constant τ_s, and observation noise $\widetilde{\sigma}_{\text{obs}}$ are assumed to be known (although $\boldsymbol{C}$ could be learned on longer time scales, see **Supplementary Note 3, Section 5**). The challenge in this hierarchical inference task is that $\boldsymbol{s}_t$ and $\boldsymbol{\lambda}$ are mutually dependent on another: $\boldsymbol{\lambda}$ acts as a parameter in $p(\,\boldsymbol{s}_t\,|\,\boldsymbol{\lambda}\,)$ per eqn. (2), and will therefore affect the posterior $p(\,\boldsymbol{s}_t\,|\,\boldsymbol{v}_{1:t}\,)$. On the other hand, inferring the

presence of, say, flocking birds ($\lambda_{\text{flock}} > 0$) depends on "perceiving" an instantaneous flock motion, $s_{\text{flock}}(t)$, in the first place.

The EM algorithm [6] offers a solution to this chicken-and-egg problem. For its application, we leverage the fact that motion sources and strengths change on different time scales, τ_s for $\boldsymbol{s}_t$ and τ_λ for $\boldsymbol{\lambda}$. For $\tau_\lambda \gg \tau_s$, we can treat $\boldsymbol{\lambda}$ as a constant while inferring $\boldsymbol{s}(t)$—known as the E-step in EM—, and then, in alternation, optimize $\boldsymbol{\lambda}$ based on the inferred motion strengths—the M-step in EM. In **Supplementary Note 2, Section 1.1** and **Supplementary Note 2, Section 1.2**, we will address the E-step and the M-step separately. As before, we will develop the theory in one spatial dimension for notational clarity and then generalize to $D > 1$. Further, we will present the derivation in discrete time in a batch formulation. The continuous-time limit will be drawn in **Supplementary Note 2, Section 2**.

2.1.1 Inference of motion sources for a given structure (E-step)

For the E-step, we aim to infer $p(\boldsymbol{s}_{1:t} \,|\, \boldsymbol{v}_{1:t}; \boldsymbol{\lambda})$ for a given structure $\boldsymbol{\lambda}$, and then compute the expected value of the log-likelihood of the augmented data distribution $p(\boldsymbol{v}_{1:t}, \boldsymbol{s}_{1:t}; \boldsymbol{\lambda})$, see, e.g., Section 9.3. in [7]. For the remainder of this subsection, we will often suppress the explicit dependence on $\boldsymbol{\lambda}$ to avoid notational clutter. Since we are interested in an online algorithm, we will use the filtering solution which is obtained from iterative application of temporal propagation to the next observation time,

$$p(\boldsymbol{s}_{t-1} \,|\, \boldsymbol{v}_{1:t-1}) \;\longrightarrow\; p(\boldsymbol{s}_t \,|\, \boldsymbol{v}_{1:t-1}) \;, \tag{7}$$

and integration of the next observation,

$$p(\boldsymbol{s}_t \,|\, \boldsymbol{v}_{1:t}) \propto p(\boldsymbol{s}_t \,|\, \boldsymbol{v}_{1:t-1})\, p(\boldsymbol{v}_t \,|\, \boldsymbol{s}_t) \;. \tag{8}$$

Propagation, eqn. (7), is performed by propagating the density according to the stochastic process in eqn. (2). Mathematically, this is done by convolving $p(\boldsymbol{s}_{t-1} \,|\, \boldsymbol{v}_{1:t-1})$ with the Gaussian transition density $p(\boldsymbol{s}_t \,|\, \boldsymbol{s}_{t-1})$ of the Ornstein-Uhlenbeck process. Integration, eqn. (8), is the application of Bayes rule using the emission model of eqn. (4).

For linear stochastic dynamics with a Gaussian emission model, i.e., the present case, the posterior will always be a multivariate Gaussian with some mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$: $p(\boldsymbol{s}_t \,|\, \boldsymbol{v}_{1:t}) = \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Kalman filtering [8] is one possible algorithm for calculating the posterior moments $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, and we refer the interested reader to the Supporting Information of Ref. [1] for explicit forms of the Kalman filter's state transition matrix and process noise covariance matrix. For the present work, we will employ a more elegant, continuous-time solution that is equivalent to the Kalman-Bucy filter [9, 10] for calculating $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$. The derivation, which is provided in **Supplementary Note 2, Section 2**, will furthermore facilitate a neuro-friendly approximate implementation (**Supplementary Note 2, Section 3**). For now, let us assume that the posterior moments, $\boldsymbol{\mu}_{1:t}$ and $\boldsymbol{\Sigma}_{1:t}$, have been computed by whichever method of choice.

E-step. We compute the expected value of the log-likelihood of the augmented data distribution,[1]

$$\begin{aligned} Q(\boldsymbol{\lambda}) &= \langle \log p(\boldsymbol{v}_{1:t}, \boldsymbol{s}_{1:t}; \boldsymbol{\lambda}) \rangle_{p(\boldsymbol{s}_{1:t} \,|\, \boldsymbol{v}_{1:t})} && (9) \\ &= \sum_{j=1}^{t} \langle \log p(\boldsymbol{v}_j \,|\, \boldsymbol{s}_j; \boldsymbol{\lambda}) + \log p(\boldsymbol{s}_j \,|\, \boldsymbol{s}_{j-1}; \boldsymbol{\lambda}) \rangle_{p(\boldsymbol{s}_j, \boldsymbol{s}_{j-1} \,|\, \boldsymbol{v}_{1:t})} && (10) \\ &\overset{\text{ind.}}{\approx} \sum_{j=1}^{t} \langle \log p(\boldsymbol{v}_j \,|\, \boldsymbol{s}_j; \boldsymbol{\lambda}) + \log p(\boldsymbol{s}_j; \boldsymbol{\lambda}) \rangle_{p(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:j})} \overset{\text{def.}}{=} \sum_{j=1}^{t} Q_j(\boldsymbol{\lambda}) \;, && (11) \end{aligned}$$

where, at the third equality, we have made two approximations, namely, that (i) consecutive observations were independent, and (ii) we take the expectation w.r.t. the filtering posterior, $p(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:j})$, rather than the smoothing density, $p(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:t})$. The first approximation will be corrected for during the M-step in **Supplementary Note 2, Section 1.2**, where we will weigh the likelihood term as if it was comprised of only $t \cdot \delta t / \tau_s$ independent samples when combining it with a sparsity prior on $\boldsymbol{\lambda}$. This correction is justified because it is equivalent to considering only a sparse subsample of observations that each lie τ_s apart. Consecutive observations in this subsample are almost decoupled since the Ornstein-Uhlenbeck process in eqn. (2) decorrelates the motions sources at time scale τ_s. The second approximation—filtering rather than smoothing—is a modeling assumption for what information is employed by an online agent, e.g., a human observer.

[1] In the 2nd line, we implicitly assume, for mathematical rigor only, that there is an initial distribution, $p(\boldsymbol{s}_{t=0})$, which is absorbed again in the 3rd line.

The $Q_j(\boldsymbol{\lambda})$ can be calculated analytically:

$$Q_j(\boldsymbol{\lambda}) = \langle \log p\left(\boldsymbol{v}_j \,|\, \boldsymbol{s}_j; \boldsymbol{\lambda}\right) + \log p(\boldsymbol{s}_j; \boldsymbol{\lambda}) \rangle_{p\left(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:j}\right)} \qquad \leftarrow \text{ with } p(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:j}) = \mathcal{N}\left(\boldsymbol{s}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{12}$$

$$= \langle -\tfrac{1}{2}(\boldsymbol{v}_j - \boldsymbol{C}\boldsymbol{s}_j)^{\mathsf{T}} \widetilde{\sigma}_{\text{obs}}^{-2} \boldsymbol{I} (\boldsymbol{v}_j - \boldsymbol{C}\boldsymbol{s}_j) - \tfrac{1}{2}\boldsymbol{s}_j^{\mathsf{T}} \operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]^{-1} \boldsymbol{s}_j - \tfrac{1}{2}\log|\widetilde{\sigma}_{\text{obs}}^2 \boldsymbol{I}| - \tfrac{1}{2}\log|\operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]| \rangle_{p\left(\boldsymbol{s}_j \,|\, \boldsymbol{v}_{1:j}\right)} + \text{const.} \tag{13}$$

$$= \sum_{m=1}^{M} -\frac{1}{\tau_s}\frac{\mu_{j,m}^2 + \sigma_{j,m}^2}{\lambda_m^2} - \frac{1}{2}\log\lambda_m^2 \quad \text{[dropping } \boldsymbol{\lambda}\text{-independent terms and using } \langle s^2 \rangle_{\mathcal{N}(s;\mu,\sigma^2)} = \mu^2 + \sigma^2\text{]} \tag{14}$$

with $\sigma_{j,m}^2 := \Sigma_{j,mm}$, $|\cdot|$ denoting the determinant, and the λ-independent terms have been dropped because they will not play a role in the maximization w.r.t. $\boldsymbol{\lambda}$ in **Supplementary Note 2, Section 1.2**. For the full $Q(\boldsymbol{\lambda})$, we thus obtain:

$$Q(\boldsymbol{\lambda}) = \sum_{j=1}^{t} Q_j(\boldsymbol{\lambda}) = -\frac{t}{\tau_s}\sum_{m=1}^{M}\frac{\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j}{\lambda_m^2} + \frac{\tau_s}{2}\log\lambda_m^2 \ , \tag{15}$$

with $\langle\cdot\rangle_j$ being the time average over the batch.

2.1.2 Inference of motion strengths and sparsity prior (M-step)

For the M-step, $Q(\boldsymbol{\lambda})$ is maximized w.r.t. $\boldsymbol{\lambda}$ to obtain the maximum likelihood (ML) solution. Here, we will make use of the additional freedom to impose a prior distribution, $p(\boldsymbol{\lambda})$, on the motion strengths, see, e.g., Section 9.4. in [7]. First, we will introduce a family of prior distributions which reflect our knowledge that from a reservoir of motion components most components will be absent or small in any given scene (sparsity prior). Then, we will perform the M-step to derive the maximum a posteriori (MAP) solution.

Sparsity prior. In foresight of the M-step, we formulate the prior over $\boldsymbol{\lambda}^2$ (instead of $\boldsymbol{\lambda}$) and choose

$$p(\boldsymbol{\lambda}^2; \boldsymbol{\nu}, \boldsymbol{\kappa}^2) = \prod_{m=1}^{M} \mathcal{I}\chi(\lambda_m^2; \nu_m, \kappa_m^2) \qquad \text{with} \qquad \mathcal{I}\chi(\lambda^2; \nu, \kappa^2) = \frac{1}{\lambda^{(2+\nu)}}\exp\left[-\frac{\nu\kappa^2}{2\lambda^2} - A(\nu, \kappa^2)\right] \ . \tag{16}$$

$\mathcal{I}\chi(\lambda^2; \nu, \kappa^2)$ denotes the density of the scaled inverse chi-squared distribution which is the conjugate prior to a normal distribution with known mean and unknown variance. Conceptually, this is exactly our task at hand: we know that $\langle s_m \rangle = 0$ in the generative model, but we have to estimate its variance, $\langle s_m^2 \rangle$, which is controlled by λ_m^2. As will become obvious in the M-step, the (hyper-)parameters, $\boldsymbol{\nu}$ and $\boldsymbol{\kappa}^2$, will take the role of pseudo-counts and pseudo-observations, respectively. The log-partition, $A(\nu, \kappa^2) = \log\Gamma(\frac{\nu}{2}) - \frac{\nu}{2}\log\frac{\nu\kappa^2}{2}$, will have no effect on the MAP estimate.

Two choices of ν and κ^2 are of particular interest. For $\nu{=}\kappa^2{=}0$, we have $p(\lambda^2) \propto 1/\lambda^2$ which is a non-informative (Jeffreys) prior on the variance of s^2. For $\nu{=}-2$, $\kappa^2{=}0$, we have $p(\lambda^2) \propto 1$ which is a uniform prior. The latter choice of hyper-parameters will turn the MAP estimate into the ML estimate.

M-step. To maximize Q, we find the roots of its λ_m^2-derivatives:

$$0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}\lambda_m^2}\left(Q(\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}^2; \boldsymbol{\nu}, \boldsymbol{\kappa}^2)\right) = \frac{t}{\tau_s}\left(\frac{\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j}{(\lambda_m^2)^2} - \frac{\tau_s}{2}\frac{1}{\lambda_m^2}\right) + \frac{\nu_m\kappa_m^2}{2(\lambda_m^2)^2} - \frac{1 + \frac{\nu_m}{2}}{\lambda_m^2} \tag{17}$$

$$\Rightarrow\ t\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j + \tfrac{\tau_s}{2}\nu_m\kappa_m^2 = t\tfrac{\tau_s}{2}\lambda_m^2 + \tfrac{\tau_s}{2}(2 + \nu_m)\lambda_m^2 \tag{18}$$

$$\Rightarrow\ \tfrac{\tau_s}{2}\lambda_m^2 = \frac{t\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j + \frac{\tau_s}{2}\nu_m\kappa_m^2}{2 + \nu_m + t} \ . \tag{19}$$

Eqn. (19) highlights several intuitive properties of motion structure inference. First, the MAP value of motion strength λ_m^2 only depends on the inferred posterior moments of the corresponding motion source s_m, that is, there is no cross-talk between motion sources s_m and $s_{m'}$. Second, recalling that $\frac{\tau_s}{2}\lambda_m^2$ is the expected long-term variance of s_m according to eqn. (3), eqn. (19) tells us to match this expected variance, $\frac{\tau_s}{2}\lambda_m^2$, to the observed variance, $\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j$, of the inferred motion source *over time*.[2] Third, hyper-parameter κ_m plays the role of an average pseudo-observation,

[2] This is to be distinguished from the posterior's instantaneous uncertainty $\sigma_{t,m}^2$.

$\frac{\tau_s}{2}\kappa_m^2 = \langle \mu_m^2 + \sigma_m^2 \rangle$, which is then weighted as ν_m pseudo-samples against the t observed data samples. Thus, κ_m describes (a priori) typical values of λ_m. (The summand 2 in the denominator of eqn. (19) is a relict of the scale-invariant Jeffreys prior.) Finally, a uniform hyper-prior, $\nu_m = -2$, $\kappa_m = 0$, yields straightforward variance matching as the ML solution.

We conclude the M-step, by correcting eqn. (19) for the fact that the t data samples actually only represent $t \cdot \delta t / \tau_s$ independent samples, as promised in **Supplementary Note 2, Section 1.1**,

$$\frac{\tau_s}{2}\lambda_m^2 = \frac{\frac{t\cdot\delta t}{\tau_s}\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \rangle_j + \frac{\tau_s}{2}\nu_m\kappa_m^2}{2 + \nu_m + \frac{t\cdot\delta t}{\tau_s}} \ . \tag{20}$$

Eqn. (20) is the batch solution to motion structure inference which we will build on for the continuous-time formulation in **Supplementary Note 2, Section 2**.

2.1.3 Extension to multiple spatial dimensions

When $D > 1$, we typically assume that each λ_m controls the variance of the $s_{m,d}$ in all spatial dimensions d (see **Supplementary Note 1, Section 4**). When going through the derivation of the EM algorithm, the decisive changes happen in eqn. (15): (i) the expectation now runs over all spatial dimensions, i.e., $\langle \sum_{d=1}^{D} \mu_{j,m,d}^2 + \sigma_{j,m,d}^2 \rangle_j$; (ii) the log-partition gets multiplied by D, due to each λ_m contributing with the power of D to $\log|\mathrm{diag}\left[\boldsymbol{\lambda}^2\right]|$. With these changes, the M-step in eqn. (20) finds its optimum when

$$D\,\frac{\tau_s}{2}\lambda_m^2 = \frac{\frac{t\cdot\delta t}{\tau_s}\langle \sum_{d=1}^{D} \mu_{j,m,d}^2 + \sigma_{j,m,d}^2 \rangle_j + \frac{\tau_s}{2}\nu_m\kappa_m^2}{\frac{2}{D} + \nu_m + \frac{t\cdot\delta t}{\tau_s}} \ . \tag{21}$$

Here, we have made a slight re-parameterization of ν_m and κ_m to preserve the developed intuition that κ_m describes typical values of λ_m, and that ν_m counts the number of pseudo-observations.[3]

2.2 Continuous-time, online inference

We now turn to a continuous-time formulation of the above motion stucture inference algorithm. While doing so, we will overload the notation of time t which previously denoted integer-valued time steps and now becomes real-valued. We will point out the respective locations where this transition happens, below, to preclude confusion. Concretely, we will first reformulate the generative model in the form of natural parameters in **Supplementary Note 2, Section 2.1**. Then, we derive continuous-time dynamics on these parameters in **Supplementary Note 2, Section 2.2**, as had been promised in **Supplementary Note 2, Section 1.1**, for solving the E-step. Finally, we cast eqn. (21) into a recursive equation for an online, continuous-time M-Step in **Supplementary Note 2, Section 2.3**. This results in the reference *online EM algorithm* for online hierarchical motion structure inference.

2.2.1 Equivalent formulation of the generative model and inference using natural parameters

Knowing that all distributions involved in propagation, eqn. (7), and integration, eqn. (8), are multivariate Gaussians, we write the emission model, eqn. (4), in terms of the sufficient statistics of $\boldsymbol{s}_t$,

$$p(\,\boldsymbol{v}_t \,|\, \boldsymbol{s}_t\,) = \mathcal{N}\left(\boldsymbol{v}_t; \boldsymbol{C}\,\boldsymbol{s}_t, \widetilde{\sigma}_{\mathrm{obs}}^2\,\boldsymbol{I}\right) \propto \exp\left[\begin{pmatrix} \boldsymbol{s}_t \\ \boldsymbol{s}_t\boldsymbol{s}_t^{\mathsf{T}} \end{pmatrix} \cdot \begin{pmatrix} \frac{\boldsymbol{C}^{\mathsf{T}}\boldsymbol{v}_t}{\widetilde{\sigma}_{\mathrm{obs}}^2} \\ -\frac{1}{2}\frac{\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}}{\widetilde{\sigma}_{\mathrm{obs}}^2} \end{pmatrix}\right] \ , \tag{22}$$

where we have dropped all $\boldsymbol{s}_t$-independent terms because they will play no role in the inference. Denoting the propagated, yet pre-integration, distribution by

$$p(\,\boldsymbol{s}_t \,|\, \boldsymbol{v}_{0:t-\delta t}\,) \propto \exp\left[\begin{pmatrix} \boldsymbol{s}_t \\ \boldsymbol{s}_t\boldsymbol{s}_t^{\mathsf{T}} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\Omega}\boldsymbol{\mu}_t \\ -\frac{1}{2}\boldsymbol{\Omega}_t \end{pmatrix}\right] \ , \tag{23}$$

[3] Specifically, we substituted $\frac{\nu_m}{D} \to \nu_m$ and $D\kappa_m^2 \to \kappa_m^2$. Then, $\nu_m = 1$ means one pseudo-observation in *each* spatial dimension. A uniform prior is imposed by $\nu_m = -2/D$ and $\kappa_m = 0$.

with yet-to-be-determined natural parameters $\mathbf{\Omega}_t := \mathbf{\Sigma}_t^{-1}$ and $\mathbf{\Omega}\boldsymbol{\mu}_t := \mathbf{\Omega}_t \boldsymbol{\mu}_t$, the integration of observations, eqn. (8), amounts to the following simple updates:

$$\mathbf{\Omega}_t \mapsto \mathbf{\Omega}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{C}}{\widetilde{\sigma}^2_{\text{obs}}} \qquad \text{and} \qquad \mathbf{\Omega}\boldsymbol{\mu}_t \mapsto \mathbf{\Omega}\boldsymbol{\mu}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{v}_t}{\widetilde{\sigma}^2_{\text{obs}}} \ . \tag{24}$$

In eqn. (23), we have made use already of the continuous-time notation, with time running from 0 to t and observations arriving at δt-intervals. We next address propagation, i.e., the continuous-time dynamics of eqn. (2) in terms of $\mathbf{\Omega}\boldsymbol{\mu}_t$ and $\mathbf{\Omega}_t$.

2.2.2 Continuous-time dynamics of natural parameters

For the OU process, eqn. (2), we know the evolution of the distribution of $\boldsymbol{s}_t$ between observations in closed form (see, e.g., [2]). Namely, the mean $\boldsymbol{\mu}_t$ decays towards $\mathbf{0}$ exponentially with time constant τ_s, and the covariance $\mathbf{\Sigma}_t$ decays towards its steady state value $\operatorname{diag}\left[\frac{\tau_s}{2}\boldsymbol{\lambda}^2\right]$ exponentially with time constant $\frac{\tau_s}{2}$. From these known dynamics of $\boldsymbol{\mu}_t$ and $\mathbf{\Sigma}_t$, we calculate the dynamics of the natural parameters:

$$\partial_t \mathbf{\Omega}_t = -\mathbf{\Omega}_t \left(\partial_t \mathbf{\Sigma}_t\right) \mathbf{\Omega}_t = \ldots = \left(\tfrac{\tau_s}{2}\right)^{-1} \underbrace{\left(\boldsymbol{I} - \mathbf{\Omega}_t \operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]\right)}_{\circledast} \mathbf{\Omega}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{C}}{\widetilde{\sigma}^2_{\text{obs}}}\, \delta(t - t_{\text{obs}}) \ . \tag{25}$$

For completeness, we have already included the integration of observations, eqn. (24), at observation time t_{obs} in the dynamics. In contrast to the dynamics of $\mathbf{\Sigma}_t$, the dynamics of $\mathbf{\Omega}_t$ are non-linear. Yet, we observe that, in the absence of observations, eqn. (25) leads to the desired fixed point since $\circledast = 0$ for $\mathbf{\Omega}^{-1} = \operatorname{diag}\left[\frac{\tau_s}{2}\boldsymbol{\lambda}^2\right]$. Likewise, we obtain for the other natural parameter $\mathbf{\Omega}\boldsymbol{\mu}_t$:

$$\partial_t (\mathbf{\Omega}\boldsymbol{\mu}_t) = \left(\partial_t \mathbf{\Omega}_t\right) \boldsymbol{\mu}_t + \mathbf{\Omega}_t \left(\partial_t \boldsymbol{\mu}_t\right) = \ldots = \tau_s^{-1} \underbrace{\left(\boldsymbol{I} - 2\mathbf{\Omega}_t \operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]\right)}_{\circledast\circledast} \mathbf{\Omega}\boldsymbol{\mu}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{v}_t}{\widetilde{\sigma}^2_{\text{obs}}}\, \delta(t - t_{\text{obs}}) \ . \tag{26}$$

Again, as a sanity check, we observe the desired decay to zero because of $\circledast\circledast \to -1$, in the absence of observations.

Continuous stream of observations. So far, we have treated observations $\boldsymbol{v}_t$ as point observations which arrive only at distinct time points t_{obs} and, then, lead to "jumps" via integrating over the Dirac delta. For a complete continuous-time formulation, we choose to turn observations into a continuous input stream. When observations arrive every δt-interval and are corrupted by i.i.d. Gaussian noise of variance $\widetilde{\sigma}^2_{\text{obs}}$, we can render their information content δt-independent by setting

$$\widetilde{\sigma}^2_{\text{obs}} = \sigma^2_{\text{obs}} / \delta t \tag{27}$$

with the alternative parameter σ^2_{obs} being independent of the observation frame rate (see, e.g., [10, 11]). Furthermore, for $\delta t \to 0$, we use that

$$\delta t\, \delta(t - t_{\text{obs}}) \to 1 \tag{28}$$

because we get one Dirac delta-integration per δt in eqn. (25) and (26) while all other variables stay (almost) constant. With these two substitutions, we obtain:

$$\partial_t \mathbf{\Omega}_t = \left(\tfrac{\tau_s}{2}\right)^{-1} \left(\boldsymbol{I} - \mathbf{\Omega}_t \operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]\right) \mathbf{\Omega}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{C}}{\sigma^2_{\text{obs}}} \ , \tag{29}$$

$$\partial_t (\mathbf{\Omega}\boldsymbol{\mu}_t) = \tau_s^{-1} \left(\boldsymbol{I} - 2\mathbf{\Omega}_t \operatorname{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right]\right) \mathbf{\Omega}\boldsymbol{\mu}_t + \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{v}_t}{\sigma^2_{\text{obs}}} \ , \tag{30}$$

with a continuous stream of observations, $\boldsymbol{v}_t$. Together, eqn. (29) and (30) solve the E-step (from **Supplementary Note 2, Section 1.1**) by re-transforming the parameters via $\mathbf{\Sigma}_t = \mathbf{\Omega}_t^{-1}$ and $\boldsymbol{\mu}_t = \mathbf{\Sigma}_t \mathbf{\Omega}\boldsymbol{\mu}_t$. These moments are used in the filtering posterior $p(\boldsymbol{s}_t \,|\, \boldsymbol{v}_{0:t}; \boldsymbol{\lambda})$ of the reference online EM algorithm. The solution in terms of natural parameters is equivalent to the Kalman-Bucy filter [9] which is derived directly in terms of $\boldsymbol{\mu}_t$ and $\mathbf{\Sigma}_t$, as we will see in **Supplementary Note 2, Section 3.1**.

2.2.3 Simultaneous online inference of motion sources and structure

We complete our derivation of online, hierarchical inference of motion sources and motion structure by casting the M-step, eqn. (20), into a recursive form and drawing the continuous-time limit. We restate eqn. (20) for reference,

$$\tfrac{\tau_s}{2}\lambda_m^2 = \frac{\frac{t\cdot\delta t}{\tau_s}\left\langle \mu_{j,m}^2 + \sigma_{j,m}^2 \right\rangle_j + \frac{\tau_s}{2}\nu_m\kappa_m^2}{2+\nu_m+\frac{t\cdot\delta t}{\tau_s}} \ , \tag{31}$$

where $\mu_{j,m}$ and $\sigma_{j,m}^2 = \Sigma_{j,mm}$ are the posterior parameters obtained from the E-step. Recalling that $\boldsymbol{\lambda}_t^2$ is based on the time average of t discrete-time samples, we formulate $\boldsymbol{\lambda}_{t+1}^2$ as a sliding window average,

$$\boldsymbol{\lambda}_{t+1}^2 = \left(1-\frac{1}{t}\right)\cdot\boldsymbol{\lambda}_t^2 + \frac{1}{t}\cdot\left(\frac{\tau_s}{2}\right)^{-1}\frac{\frac{t\cdot\delta t}{\tau_s}\left(\boldsymbol{\mu}_{t+1}^2+\boldsymbol{\sigma}_{t+1}^2\right)+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{2+\boldsymbol{\nu}+\frac{t\cdot\delta t}{\tau_s}} \ , \tag{32}$$

where $(1-\frac{1}{t})$ and $\frac{1}{t}$ are the weights for convex combination, and all vector operations are elementwise. Subtracting $\boldsymbol{\lambda}_t^2$ and dividing by inter-observation interval δt, we obtain:

$$\frac{\boldsymbol{\lambda}_{t+1}^2-\boldsymbol{\lambda}_t^2}{\delta t} = -\frac{1}{t\,\delta t}\left(\boldsymbol{\lambda}_t^2-\left(\frac{\tau_s}{2}\right)^{-1}\frac{\frac{t\cdot\delta t}{\tau_s}\left(\boldsymbol{\mu}_{t+1}^2+\boldsymbol{\sigma}_{t+1}^2\right)+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{2+\boldsymbol{\nu}+\frac{t\cdot\delta t}{\tau_s}}\right) \ . \tag{33}$$

In this form, drawing the continuous-time limit is straight-forward. We let $\delta t\to 0$ while keeping $\tau_\lambda := t\,\delta t$ constant:

$$\partial_t\boldsymbol{\lambda}_t^2 = -\frac{1}{\tau_\lambda}\left(\boldsymbol{\lambda}_t^2-\left(\frac{\tau_s}{2}\right)^{-1}\frac{\frac{\tau_\lambda}{\tau_s}\left(\boldsymbol{\mu}_t^2+\boldsymbol{\sigma}_t^2\right)+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{2+\boldsymbol{\nu}+\frac{\tau_\lambda}{\tau_s}}\right) \ . \tag{34}$$

Time t is now in continuous-time. The time constant τ_λ is the (continuous-time) width of the integration window and defines the minimum time scale at which significant changes of $\boldsymbol{\lambda}_t^2$ are expected to occur. From a strict algorithmic perspective of EM, we require that $\tau_\lambda \gg \tau_s$. However, we observe in computer simulations that for practical applications even small values, $\tau_\lambda \gtrapprox \tau_s$, work reliably.

Extension to multiple spatial dimensions. We conclude by generalizing eqn. (34) to multiple spatial dimensions via comparison of eqn. (20) and eqn. (21):

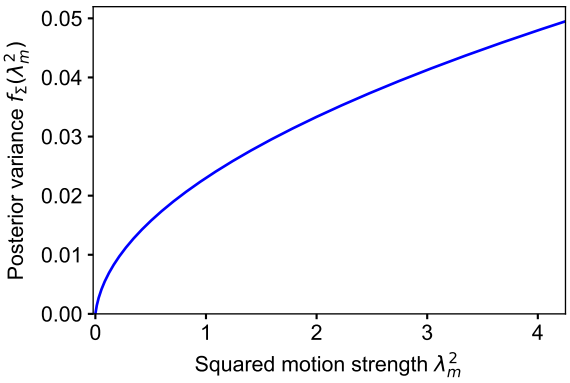
$$\partial_t\boldsymbol{\lambda}_t^2 = -\frac{1}{\tau_\lambda}\left(\boldsymbol{\lambda}_t^2-\left(\frac{D\,\tau_s}{2}\right)^{-1}\frac{\frac{\tau_\lambda}{\tau_s}\left(\sum_{d=1}^D\boldsymbol{\mu}_{t,d}^2+\boldsymbol{\sigma}_{t,d}^2\right)+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{\frac{2}{D}+\boldsymbol{\nu}+\frac{\tau_\lambda}{\tau_s}}\right) \ . \tag{35}$$

Eqn. (35) is used for inferring the motion strengths $\boldsymbol{\lambda}_t^2$ in the reference online EM algorithm. The interactions of eqn. (29), (30) and (35) confirm and particularize our earlier chicken-and-egg intuition that motion sources, $\boldsymbol{s}_t$, and motion structure, $\boldsymbol{\lambda}_t$, are mutually coupled. The exact interactions are rather convoluted, and we will identify in **Supplementary Note 2, Section 3** an approximate interaction that is intuitively more accessible.

2.3 Adiabatic approximation for prediction error-based inference

The inference algorithm described by eqn. (29), (30), and (35) is a nice solution from a mathematical perspective. Yet, for a theory of brain computation, it is questionable whether neural dynamics could reliably calculate 3rd-order polynomials in the variables $\boldsymbol{\Omega}\boldsymbol{\mu}_t$, $\boldsymbol{\Omega}_t$, and $\boldsymbol{\lambda}_t^2$ as demanded, for instance, by eqn. (30).

In the following, we therefore explore an alternative, approximate solution which, as we will see, considerably simplifies the involved computations while leading to almost identical results during motion structure inference. We will proceed in three steps. First, in **Supplementary Note 2, Section 3.1**, we transfer eqn. (29) + (30) back into the domain of moments, $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, and recover a prediction error-based update equation known as the Kalman-Bucy filter [9, 10]. Then, in **Supplementary Note 2, Section 3.2**, we introduce an adiabatic approximation for the posterior covariance, $\boldsymbol{\Sigma}_t$, assuming that it has always converged to stationarity. Finally, in **Supplementary Note 2, Section 3.3**, we derive an analytical solution for the converged uncertainty for the special case of diagonal covariance. The resulting algorithm, which we term the *adiabatic observer model* in the Supporting Information and which is the online hierarchical inference model used in the main text, features interpretable dynamics on the behaviorally relevant quantities $\boldsymbol{\mu}_t$ and $\boldsymbol{\lambda}_t^2$ and relies on only quadratic computations of vector-valued variables, rather than 3rd-order computations on matrix-valued variables.



Supplementary Fig. 9 | Adiabatic, diagonal solution for the posterior variance. Shown is the function $f_\Sigma(\lambda_m^2)$, given by eqn. (42). Parameters: $\tau_s = 300\,\text{ms}$, $\sigma_{\text{obs}} = 0.05$, $\|\boldsymbol{c}_m\|^2 = 4$.

2.3.1 Prediction error-based formulation

We first transform eqn. (29) into dynamics of $\boldsymbol{\Sigma}_t$:

$$\partial_t \boldsymbol{\Sigma}_t = -\boldsymbol{\Sigma}_t \left(\partial_t \boldsymbol{\Omega}_t\right) \boldsymbol{\Sigma}_t \tag{36}$$

$$= -\left(\tfrac{\tau_s}{2}\right)^{-1} \boldsymbol{\Sigma}_t + \text{diag}\left[\boldsymbol{\lambda}^2\right] - \boldsymbol{\Sigma}_t \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{C}}{\sigma_{\text{obs}}^2} \boldsymbol{\Sigma}_t \ , \tag{37}$$

and use this result for transforming eqn. (30) into dynamics of $\boldsymbol{\mu}_t$:

$$\partial_t \boldsymbol{\mu}_t = \partial_t (\boldsymbol{\Sigma}_t \boldsymbol{\Omega} \boldsymbol{\mu}_t) = (\partial_t \boldsymbol{\Sigma}_t) \boldsymbol{\Omega} \boldsymbol{\mu}_t + \boldsymbol{\Sigma}_t \partial_t \boldsymbol{\Omega} \boldsymbol{\mu}_t \tag{38}$$

$$= -\frac{\boldsymbol{\mu}_t}{\tau_s} + \boldsymbol{\Sigma}_t \boldsymbol{C}^\mathsf{T} \left(\frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} - \frac{\boldsymbol{C}\boldsymbol{\mu}_t}{\sigma_{\text{obs}}^2} \right) \ . \tag{39}$$

Eqn. (39) is a pretty neat equation as it reveals the "inner working" of inference as updating $\boldsymbol{\mu}_t$ with the help of prediction errors $(\boldsymbol{v}_t - \boldsymbol{C}\boldsymbol{\mu}_t)$, which are projected "up" into motion source space via $\boldsymbol{C}^\mathsf{T}$, and then weighted by the relative uncertainty of internal estimates vs. the uncertainty of observations, $\boldsymbol{\Sigma}_t \,/\, \sigma_{\text{obs}}^2$. In particular, in the absence of observations ($\sigma_{\text{obs}}^2 \to \infty$), the estimated mean $\boldsymbol{\mu}_t = \langle \boldsymbol{s}_t \rangle$ decays to zero with time constant τ_s, as expected from the OU process. At first glance, it may seem that the inferred structure, $\boldsymbol{\lambda}^2$, plays no role in inferring $\boldsymbol{s}_t$ anymore. But actually, $\boldsymbol{\lambda}^2$ still is present in eqn. (39) indirectly through its effect on $\boldsymbol{\Sigma}_t$. We will study this indirect effect in the following.

2.3.2 Convergence approximation on the posterior precision

For $\tau_\lambda > \tau_s$, we observe that $\boldsymbol{\Sigma}_t$ can be calculated directly as a function of $\boldsymbol{\lambda}_t^2$, instead of going through the hassle of integrating eqn. (37). This can be see as follows. We know that the posterior covariance $\boldsymbol{\Sigma}_t$ decays towards its stationary value with a time constant in the order of $\tau_s/2$. The stationary value itself is a dynamic equilibrium between increasing uncertainty due to diffusion (the underlying Wiener process in eqn. (2)) and decreasing uncertainty due to incoming observations (corresponds to the term $\boldsymbol{C}^\mathsf{T}\boldsymbol{C}/\sigma_{\text{obs}}^2$ in eqn. (29)). Notably, the stationary value does not depend on the observations, $\boldsymbol{v}_t$. This is a peculiarity of the inference task at hand which is known from Kalman filtering. This leaves $\boldsymbol{\lambda}_t^2$ as the only dynamic variable in eqn. (37) to influence the stationary point of $\boldsymbol{\Sigma}_t$. Since $\boldsymbol{\lambda}_t^2$ changes on time scales $\tau_\lambda > \tau_s$, the covariance, $\boldsymbol{\Sigma}_t$, will always have enough time to react to any change in its stationary value. This justifies treating $\boldsymbol{\Sigma}_t$ as having converged at any time, a method known in physics as adiabatic approximation.

For stationary $\boldsymbol{\Sigma}_t$, it follows from setting $\partial_t \boldsymbol{\Sigma}_t = \boldsymbol{0}$ in eqn. (37) that

$$\frac{\tau_s}{2} \boldsymbol{\Sigma} \frac{\boldsymbol{C}^\mathsf{T}\boldsymbol{C}}{\sigma_{\text{obs}}^2} \boldsymbol{\Sigma} + \boldsymbol{\Sigma} - \text{diag}\left[\tfrac{\tau_s}{2}\boldsymbol{\lambda}^2\right] = \boldsymbol{0} \ . \tag{40}$$

2.3.3 Analytic solution for diagonal covariance matrices

Eqn. (40) is a continuous-time algebraic Riccati equation which can, in general, be solved using eigendecompositions of an extended matrix. However, for a neural implementation, we will follow a simpler route by assuming that

$\boldsymbol{\Sigma}$ is diagonal. This amounts to ignoring correlations in uncertainty about latent motion sources in the posterior distribution, for instance, during reasoning of the type: *"I have correctly decomposed all velocities in expectation, and I know my uncertainty about each motion component. But if I underestimated the flock velocity, then I likely overestimated the birds' individual velocities."* Only the last step of this reasoning will be ignored by dropping off-diagonal elements in $\boldsymbol{\Sigma}$. We observe in computer simulations that neglecting these posterior correlations typically has little impact on the solution.

For diagonal $\boldsymbol{\Sigma}$, eqn. (40) can be solved for each element $\sigma_m^2 := \Sigma_{mm}$ separately:

$$\frac{\tau_s}{2}\frac{\|\boldsymbol{c}_m\|^2}{\sigma_{\text{obs}}^2}(\sigma_m^2)^2 + \sigma_m^2 - \frac{\tau_s}{2}\lambda_m^2 = 0 \ , \tag{41}$$

where we have defined $\|\boldsymbol{c}_m\|^2 = \sum_{k=1}^K C_{km}^2$ to denote the vector-norm of the m-th column of $\boldsymbol{C}$, that is, the squared Euclidean length of the m-th motion component. Solving eqn. (41) is straightforward, and we denote with $f_\Sigma(\lambda_m^2)$ the resulting function for calculating σ_m^2 as a function of λ_m^2:

$$\sigma_m^2 = f_\Sigma(\lambda_m^2) = \frac{\sigma_{\text{obs}}^2}{\tau_s\|\boldsymbol{c}_m\|^2}\left(-1 + \sqrt{1 + \frac{{\tau_s}^2\|\boldsymbol{c}_m\|^2}{\sigma_{\text{obs}}^2}\lambda_m^2}\right) \ . \tag{42}$$

f_Σ is a monotonically increasing, non-negative function. Its graph is shown in **Supplementary Fig. 9** for typical parameter values. In the limit of small motion strengths, $\lambda_m \to 0$, the variance grows quadratically in λ_m (non-squared): $f_\Sigma(\lambda_m^2) \approx \frac{\tau_s}{2}\lambda_m^2$. For large strengths, $\lambda_m \to \infty$, the variance becomes linear in λ_m: $f_\Sigma(\lambda_m^2) \approx \frac{\sigma_{\text{obs}}}{\|\boldsymbol{c}_m\|}\lambda_m$.

2.3.4 Putting it together: neuro-friendly algorithm for online structure inference

To obtain the neuro-friendly *adiabatic observer model*, we simply plug eqn. (42) into the dynamics of $\boldsymbol{\lambda}_t^2$, given by eqn. (35), and $\boldsymbol{\mu}_t$, given by eqn. (39):

$$\partial_t\boldsymbol{\lambda}_t^2 = -\frac{1}{\tau_\lambda}\left(\boldsymbol{\lambda}_t^2 - \left(\frac{D\,\tau_s}{2}\right)^{-1}\frac{\frac{\tau_\lambda}{\tau_s}\left(\sum_{d=1}^D \boldsymbol{\mu}_{t,d}^2 + \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right) + \frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{\frac{2}{D} + \boldsymbol{\nu} + \frac{\tau_\lambda}{\tau_s}}\right) \ , \tag{43}$$

$$\partial_t\boldsymbol{\mu}_t = -\frac{\boldsymbol{\mu}_t}{\tau_s} + \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\,\boldsymbol{C}^\mathsf{T}\left(\frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} - \frac{\boldsymbol{C}\boldsymbol{\mu}_t}{\sigma_{\text{obs}}^2}\right) \ , \tag{44}$$

$$\text{with} \quad f_\Sigma(\lambda_m^2) = \frac{\sigma_{\text{obs}}^2}{\tau_s\|\boldsymbol{c}_m\|^2}\left(-1 + \sqrt{1 + \frac{{\tau_s}^2\|\boldsymbol{c}_m\|^2}{\sigma_{\text{obs}}^2}\lambda_m^2}\right) \ .$$

In this vector notation, $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$ is evaluated elementwise, and the sum in eqn. (43) includes one evaluation of $\boldsymbol{f}_\Sigma$ for each spatial dimension $d = 1..D$. In eqn. (44), $f_\Sigma(\boldsymbol{\lambda}_t^2)$ is multiplied elementwise with the "up-projected" prediction error.

We recognize how the motion structure, $\boldsymbol{\lambda}_t$, controls the gating function, $f_\Sigma(\boldsymbol{\lambda}_t^2)$, for performing the credit assignment of the prediction errors, in eqn. (44). Furthermore, we note that also the posterior uncertainty can be recovered at any time since $\sigma_{t,m}^2 = f_\Sigma(\lambda_{t,m}^2)$.

For $D = 1$, eqn. (43) is eqn. (1) from the main text, eqn. (44) is eqn. (2), and eqn. (42) is eqn. (3).

2.3.5 A pleasant note on inference of rotational and radial motion

The approximations introduced for the adiabatic observer model hold, remarkably and importantly, also for the biologically relevant case of radial and rotational motion (cf. **Supplementary Note 1, Section 5** and **Supplementary Fig. 8** for the generative model).

To illustrate this, consider the case of two motion sources, radial $s_{\text{rad}}(t)$ and rotational $s_{\text{rot}}(t)$, with motion features given by eqn. (6). The motion features $\boldsymbol{C}$ depend on the receptive field locations with parameters R_k and ϑ_k. For inferring the radial component, $s_{\text{rad}}(t)$, all receptive fields are equally informative, irrespective of their eccentricity, R_k. Accordingly, eqn. (44) weighs all radial prediction errors equally, as expressed by the R_k-independence of the radial row in $\boldsymbol{C}^\mathsf{T}$. This is different when estimating rotational motion, $s_{\text{rot}}(t)$. Here, prediction errors measured near the fovea (small R_k) contribute only little to $\partial_t\mu_{\text{rot}}$: for small R_k, we expect to observe a small rotational motion energy via $\boldsymbol{C}\boldsymbol{\mu}_t$, such that the "Cartesian" observation noise of size σ_{obs} makes the stimulus virtually uninformative about the rotational velocity $\dot{\varphi}$. The scaling of $\boldsymbol{C}^\mathsf{T}$ with R_k accounts for that. Receptive fields far away from the fovea, in

contrast, predict a strong rotational velocity via $\boldsymbol{C}\boldsymbol{\mu}_t$, such that noise of size σ_{obs} (in Cartesian space) has only a minor impact on estimating μ_{rot}. Accordingly, the scaling of $\boldsymbol{C}^\mathsf{T}$ with R_k assigns a higher weight to peripheral receptive fields for estimating rotational motion.

As a final remark, the above example assumed the observation noise σ_{obs} to be constant across all receptive fields. Eqn. (44) naturally supports extensions to heterogeneous observation noise because all local prediction errors are measured in units of their local noise.

Supplementary Note 3. Extensions of the online model

3.1 Non-isotropic observation noise and missing observations

For the main manuscript, we have assumed the observation noise, σ_{obs}, to be a constant across time, t, observed features, k, and spatial dimensions, d. In real-world scenes, the observation noise could change along all those indices. An object could be occluded, or otherwise temporarily invisible, leading to $1/\sigma^2_{\text{obs}}=0$. Different objects might have different observation noise, e.g., due to different visual contrast. The aperture problem could render local velocity signals ambiguous: e.g., what is the direction of motion for a straight line that is larger than the aperture? This could be modeled by small observation noise perpendicular to the line, and large noise parallel to the line. The so-constructed diagonal covariance matrix, $\boldsymbol{\Sigma}_{\text{diag}}$, is then rotated as per $\boldsymbol{\Sigma}_{\text{xy}}=\boldsymbol{Q}^\mathsf{T}\boldsymbol{\Sigma}_{\text{diag}}\boldsymbol{Q}$ with rotation matrix, $\boldsymbol{Q}$, into the canonical x-y-coordinate frame.

The above extensions of the observation noise are supported by our online model, as long as changes in σ^2_{obs} occur slower than τ_s (so the adiabatic approximation remains valid). In eqn. (44), $\frac{1}{\sigma^2_{\text{obs}}}$ is extended to have different elements, $\frac{1}{\sigma^2_{\text{obs},k,d}}$, and is multiplied elementwise with $v_{t,k,d}$ and $(\boldsymbol{C}\boldsymbol{\mu})_{t,k,d}$. When calculating f_Σ, replace

$$\frac{\|\boldsymbol{c}_m\|^2}{\sigma^2_{\text{obs}}} \qquad \text{by} \qquad \left(\boldsymbol{C}^\mathsf{T}\,\text{diag}\left[\frac{\mathbf{1}}{\boldsymbol{\sigma}^2_{\text{obs}}}\right]\boldsymbol{C}\right)_{mm} \tag{45}$$

for each spatial dimension. In eqn. (43), the summation then runs over the different spatial dimensions of f_Σ. The idea that $\frac{1}{\sigma^2_{\text{obs}}}$ is vector-valued is also used in the network implementation in **Supplementary Note 4**.

A note on the provided Python code package: In the code for the network, $\boldsymbol{\sigma}^2_{\text{obs}}$ can currently have different values for every input velocity v_k, but is assumed to be (a) identical in both spatial dimensions, and (b) not to change over time. The code for the algorithm is less restrictive by supporting temporary masking of inputs (presented via `class ObservationGeneratorVelo`; leading to posterior variance $\sigma^2_m=f_\Sigma(\lambda^2_m)=\frac{\tau_s}{2}\lambda^2_m$ if all dependent objects are invisible according to eqn. (41)) and different noise in spatial dimensions.

3.2 Heterogeneous time constants

Different motion components, s_m, might have different time-constants, τ_s, for typical changes in speed and direction to occur. An extension to a vector $\boldsymbol{\tau_s}=(\tau_{s,1},..,\tau_{s,M})$ is straightforward by replacing all occurences of τ_s and $1/\tau_s$ by $\text{diag}[\boldsymbol{\tau_s}]$ and $\text{diag}\left[\boldsymbol{\tau_s}^{-1}\right]$, respectively. The code package supports heterogeneous time constants for both the algorithm and the network.

3.3 Interaction priors capturing feature compatibility

Some motion components may be unlikely to occur together. Consider, for example, two cluster components in $\boldsymbol{C}$ which are overlapping but do not contain one another: the simultaneous occurrence of the two clusters would not be compatible with a tree structure. We can accommodate such incompatibility with the help of an *interaction prior*.

In the following, we outline how interaction priors can be included in the theory, and how they will affect the inference process. We endow the $\boldsymbol{\lambda}^2$-prior from eqn. (16) with an interaction term:

$$p(\boldsymbol{\lambda}^2;\boldsymbol{\nu},\boldsymbol{\kappa}^2) \propto \left[\prod_{m=1}^{M}\mathcal{I}\chi(\lambda^2_m;\nu_m,\kappa^2_m)\right]\cdot e^{-\frac{1}{2}(\boldsymbol{\lambda}^2)^\mathsf{T}\boldsymbol{J}(\boldsymbol{\lambda}^2)}\,, \tag{46}$$

where the interaction matrix $\boldsymbol{J}\in\mathbb{R}^{M\times M}$ is a symmetric, zero-diagonal matrix that models feature incompatibility. For instance, positive values, $J_{ml}=J_{lm}>0$, describe a (soft) incompatibility between the m^{th} and l^{th} motion component.

In the derivation of the M-step, the interaction prior leads to an additional term $(-\boldsymbol{J}\boldsymbol{\lambda}^2)_m$ on the right-hand side of eqn. (17).[4] For the optimum in eqn. (19), this leads to the following equation (using vector notation and covering multiple spatial dimensions):

$$D\,\tfrac{\tau_s}{2}\left(\boldsymbol{I}+\frac{2}{2/D+\boldsymbol{\nu}+t}\mathrm{diag}\big[(\boldsymbol{\lambda}^2)^2\big]\,\boldsymbol{J}\right)\boldsymbol{\lambda}^2=\frac{t\sum_d\langle\boldsymbol{\mu}_j^2+\sigma_j^2\rangle_j+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{2/D+\boldsymbol{\nu}+t}\ . \tag{47}$$

For small values of $\|\boldsymbol{J}\|$ or, similarly, large values of t ($=\tau_\lambda/\tau_s$ in continuous-time), the matrix $\widetilde{\boldsymbol{J}}:=\big(\boldsymbol{I}+\frac{2}{2/D+\boldsymbol{\nu}+t}\mathrm{diag}\big[(\boldsymbol{\lambda}^2)^2\big]\,\boldsymbol{J}\big)$ is invertible with the approximate inverse $\widetilde{\boldsymbol{J}}^{-1}\approx\big(\boldsymbol{I}-\frac{2}{2/D+\boldsymbol{\nu}+t}\mathrm{diag}\big[(\boldsymbol{\lambda}^2)^2\big]\,\boldsymbol{J}\big)$. We can therefore move this matrix to the right-hand side and follow the derivation for online inference without interaction priors. This leads to the following equivalent of eqn. (43):

$$\partial_t\boldsymbol{\lambda}_t^2=-\frac{1}{\tau_\lambda}\left(\boldsymbol{\lambda}_t^2-\left(\frac{D\,\tau_s}{2}\right)^{-1}\left(\boldsymbol{I}-\frac{2}{\frac{2}{D}+\boldsymbol{\nu}+\frac{\tau_\lambda}{\tau_s}}\mathrm{diag}\big[(\boldsymbol{\lambda}^2)^2\big]\,\boldsymbol{J}\right)\frac{\frac{\tau_\lambda}{\tau_s}\left(\sum_{d=1}^D\boldsymbol{\mu}_{t,d}^2+\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right)+\frac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2}{\frac{2}{D}+\boldsymbol{\nu}+\frac{\tau_\lambda}{\tau_s}}\right)\ . \tag{48}$$

The only difference to eqn. (43) is that the target values on the right hand side are mixed together via $\widetilde{\boldsymbol{J}}^{-1}$. This gives rise to quite intuitive dynamics: if two motion components are incompatible, they mutually subtract their respective (independent) target values from another, thereby slightly changing the motion structure in which the E-step will interpret future input and, ultimately, leading to soft winner-takes-all competition. The term $\big(\mathrm{diag}\big[(\boldsymbol{\lambda}^2)^2\big]\big)$ limits the competition to those components that are significantly different from zero.

3.4 Detecting motion components that had decayed to baseline

If motion components are not detected in the structure for a longer time, the associated strength, λ_m, will decay to zero. Since $f_\Sigma(0)=0$, this will prohibit future detection of said motion component. This issue can be addressed in two ways. In a biological agent, noise in the nervous system will lead to small fluctuations in the encoded value of $\boldsymbol{\lambda}$, thereby "probing" the presence of motion components simply via noisy deviations from $\lambda_m=0$. Alternatively, a more principled solution exploits the hyperparameters, ν_m and κ_m, in eqn. (16) to prevent λ_m from decaying to zero, e.g., by choosing $\nu_m=1$ and $\kappa_m=0.1$. This follows the intuition that pseudo-observations in support of $\lambda_m=\kappa_m$ had been observed for a duration $\nu_m\,\tau_s$.

3.5 Learning the motion components on long time-scales

While not being the focus of this work, we briefly touch upon the question of how the motion components, $\boldsymbol{C}$, could be learned online from observations in an unsupervised manner. To this end, we follow a similar EM scheme as for inferring $\boldsymbol{\lambda}$ and note that in eqn. (13) only the quadratic term depends on $\boldsymbol{C}$. For the M-step, however, instead of maximizing Q directly, we perform gradient ascent with respect to $\boldsymbol{C}$:

$$\nabla_{\boldsymbol{C}}Q_t(\boldsymbol{C})=\big\langle\nabla_{\boldsymbol{C}}\big[-\frac{1}{2\sigma_{\text{obs}}^2}(\boldsymbol{v}_t-\boldsymbol{C}\boldsymbol{s}_t)^\mathsf{T}(\boldsymbol{v}_t-\boldsymbol{C}\boldsymbol{s}_t)\big]\big\rangle_{p(\boldsymbol{s}_t\,|\,\boldsymbol{v}_{0:t})} \tag{49}$$

$$=\frac{1}{\sigma_{\text{obs}}^2}\big\langle\boldsymbol{v}_t\,\boldsymbol{s}_t^\mathsf{T}-\boldsymbol{C}\,(\boldsymbol{s}_t\,\boldsymbol{s}_t^\mathsf{T})\big\rangle_{p(\boldsymbol{s}_t\,|\,\boldsymbol{v}_{0:t})} \tag{50}$$

$$=\frac{1}{\sigma_{\text{obs}}^2}\Big(\boldsymbol{v}_t\,\boldsymbol{\mu}_t^\mathsf{T}-\boldsymbol{C}\,\Big(\boldsymbol{\mu}_t\,\boldsymbol{\mu}_t^\mathsf{T}+\boldsymbol{\Sigma}_t\Big)\Big)\ . \tag{51}$$

This gradient establishes the intuition to compare the observed covariance between inputs and motion components against their expected covariance.

Furthermore, we note that the parameterization of motion structure via $\boldsymbol{C}$ and $\boldsymbol{\lambda}$ leaves an invariance: any scaling of $\boldsymbol{\lambda}$ can be compensated by an inverse scaling of $\boldsymbol{C}$. Due to the sparsity prior on $\boldsymbol{\lambda}$, which favors small values, this bears the risk of unbounded growth in $\boldsymbol{C}$. We can address this risk by imposing a regularizing prior on $\boldsymbol{C}$, e.g., a Laplace prior or a Gaussian prior, such that the system finds a balance between small $\boldsymbol{\lambda}$ and small $\|\boldsymbol{C}\|$. We incorporate the regularizer in the gradient-based update rule with the aim to balance prior and likelihood such that the likelihood is weighted to contribute N_C independent samples:

$$\partial_t\boldsymbol{C}=\eta_C\left[\frac{1}{\sigma_{\text{obs}}^2}\Big(\boldsymbol{v}_t\,\boldsymbol{\mu}_t^\mathsf{T}-\boldsymbol{C}\,\Big(\boldsymbol{\mu}_t\,\boldsymbol{\mu}_t^\mathsf{T}+\boldsymbol{\Sigma}_t\Big)\Big)+\frac{1}{N_C}\nabla_{\boldsymbol{C}}\log p(\boldsymbol{C})\right]\ . \tag{52}$$

[4] Note that the normalization of the distribution in eqn. (46) will not depend on λ^2, and thus will play no role in the M-step, which uses λ^2-derivatives of $\log p(\lambda^2)$.

Examples for the regularizer are $\nabla_{\boldsymbol{C}} \log p(\boldsymbol{C}) = -\text{sign}(\boldsymbol{C})/b$ for a Laplace prior and $\nabla_{\boldsymbol{C}} \log p(\boldsymbol{C}) = -\boldsymbol{C}/b$ for a Gaussian prior. For the online M-step, both $\boldsymbol{\mu}_t$ and $\boldsymbol{\lambda}_t$ must have time to adapt to any changes in $\boldsymbol{C}$. This is ensured by requiring that the learning rate, η_C, is small enough to average over a large number of independent samples, i.e., $\eta_C \ll 1/\tau_\lambda$. The small learning rate emphasizes how we think of $\boldsymbol{C}$-optimization as a long-term learning process of a "feature dictionary".

Eqn. (52) offers a path to learning $\boldsymbol{C}$ via an online EM algorithm. In a network implementation with linear population codes, however, it remains unclear how the update could be communicated to synapses: $\boldsymbol{C}$ determines many synaptic weights in the network. Finding encoding and decoding vectors of the variables to support simple plasticity rules is future work.

Supplementary Note 4. Neural network implementation

How could biological recurrent neural networks implement online motion structure inference? In light of the theory developed in **Supplementary Note 2, Section 3**, we will operationalize this question by implementing eqn. (43) and eqn. (44).

While we will strive to incorporate salient properties from motion sensitive brain areas, the exact computational mechanisms underlying many experimental findings are still elusive. Thus, inevitably, several modeling assumptions have to be made. These assumptions are presented in **Supplementary Note 4, Section 1**. We then discuss, in **Supplementary Note 4, Section 2**, which variables—input and latent—we choose to be linearly decodable by downstream populations, and, in **Supplementary Note 4, Section 3**, express the adiabatic observer model in terms of these variables. For performing the required computations on these variables, we will extend the ideas developed in Beck et al. [12], in **Supplementary Note 4, Section 4**, to a systematic theory of neural integration of high-dimensional linear and quadratic differential equations. In **Supplementary Note 4, Section 5**, we will apply the theory to derive a rate-based recurrent neural network model for online motion structure inference. Finally, in **Supplementary Note 4, Section 6**, we introduce—as an example for the computer simulations—a neural encoding model of input variables that captures many properties of middle temporal visual area (MT) while staying mathematically tractable with regard to its computational function.

4.1 Aims and assumptions

We view our network model as a starting point for an experiment–theory loop. Some neural response properties will be rather general and could be tested in experiments directly. Others will be more specific and could guide targeted experiments. In any case, we expect that many aspects of this initial model will be revised and refined in the process.

For the model, we make three assumptions:

- **Rate-based network.** We assume that all information is conveyed in the neuronal firing rates. Thus, no exact spike-timing is considered. Further, we will allow negative firing rates—think of them as negative deviations from a baseline value.

- **Linear and quadratic operations.** We assume that neurons can integrate their synaptic inputs in two ways: linearly and quadratically. Specifically, we assume that the dynamics of the firing rate of a neuron (or small population) i takes the form,

$$\tau_i \, \partial_t r_i = -r_i + f_i(\boldsymbol{w}_i^\mathsf{T} \boldsymbol{r} + \boldsymbol{r}^\mathsf{T} \boldsymbol{Q}^{(i)} \boldsymbol{r} + b_i) \ , \tag{53}$$

  with time constant τ_i, a potentially non-linear activation function f_i, weight vector $\boldsymbol{w}_i^\mathsf{T}$, quadratic interaction matrix $\boldsymbol{Q}^{(i)}$, and bias b_i. In the main text, we had omitted the (per-neuron) bias, b_i, because it can be absorbed in f_i. For the following formal derivation, we make the bias explicit for clarity, and, as we will see, this leads to activation functions, f_i, which are different only on a per-population basis. Eqn. (53) is a standard form for rate-based network models [13], and quadratic interactions are commonly used in neural network modeling [12, 14].

- **Linear decoding of task-relevant variables.** A subset of variables, especially those which are relevant for actions and decision making, can be read out linearly by downstream populations. We will specify the subset of variables in the next section.

4.2 Linearly decodable variables

Inspecting eqn. (43) and eqn. (44), an elegant decomposition into basic operations (that is, addition, linear and quadratic multiplication) employs the following variables:

$$\boldsymbol{\mu}_t\,, \qquad \boldsymbol{\lambda}_t^2\,, \qquad \frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2}\,, \qquad \frac{1}{\sigma_{\text{obs}}^2}\,, \qquad \underbrace{\left(\frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} - \frac{\boldsymbol{C}\boldsymbol{\mu}_t}{\sigma_{\text{obs}}^2}\right)}_{\text{Pred. err. } \boldsymbol{\epsilon}_t}\,, \qquad f_\Sigma(\boldsymbol{\lambda}_t^2)\ . \tag{54}$$

The first three variables, $\boldsymbol{\mu}_t$, $\boldsymbol{\lambda}_t^2$, $\boldsymbol{v}_t$, are directly related to the task of decomposing visual scenes. Further, we include the observation noise, $1/\sigma_{\text{obs}}^2$, as an input variable (rather than treating it as a constant) to accommodate the extended theory presented in **Supplementary Note 3, Section 1**, permitting the network to handle, for instance, transient occlusion of objects. Of course, the value of σ_{obs}^2 could also be a constant in the network. The prediction error, $\boldsymbol{\epsilon}_t$, is an auxiliary variable to restrict the complexity of operations to being at most quadratic. (We will see in **Supplementary Note 4, Section 4** that the computational complexity is directly inherited by the neural dynamics.) Finally, the posterior variance, $f_\Sigma(\boldsymbol{\lambda}_t^2)$, which is a hallmark of Bayesian computation, is required for motion structure decomposition and, potentially, for Bayesian decision making.

The variables listed in eqn. (54) are assumed to be linearly decodable, that is, they can be read out from neural activity (at the example of $\boldsymbol{\mu}_t$) via

$$\boldsymbol{\mu}_t = \boldsymbol{A}^\mu \boldsymbol{r}_t\ , \tag{55}$$

with some readout matrix $\boldsymbol{A}^\mu$. The other variables maintain corresponding matrices $\boldsymbol{A}^\lambda$, $\boldsymbol{A}^v$, $\boldsymbol{A}^\sigma$, $\boldsymbol{A}^\epsilon$, and $\boldsymbol{A}^\Sigma$, respectively. Here, $\boldsymbol{r}_t$ are the instantaneous firing rates of a population of neurons that encode $\boldsymbol{\mu}_t$. Multiple variables can be encoded by the same neural population.

4.3 Motion structure inference via at-most quadratic operations

First, we observe that eqn. (43) and eqn. (44) almost exclusively contain linear and quadratic terms when expressed in the variables of eqn. (54):

$$\partial_t \boldsymbol{\lambda}_t^2 = -\frac{1}{\tau_\lambda}\boldsymbol{\lambda}_t^2 + \frac{2}{D\,\tau_s\,\tau_\lambda\,(\frac{2}{D} + \boldsymbol{\nu} + \frac{\tau_\lambda}{\tau_s})}\left(\frac{\tau_\lambda}{\tau_s}\sum_{d=1}^{D}\boldsymbol{\mu}_{t,d}\odot\boldsymbol{\mu}_{t,d} + \frac{D\,\tau_\lambda}{\tau_s}f_\Sigma(\boldsymbol{\lambda}_t^2) + \tfrac{\tau_s}{2}\boldsymbol{\nu}\boldsymbol{\kappa}^2\right)\ , \tag{56}$$

$$\partial_t \boldsymbol{\mu}_t = -\frac{1}{\tau_s}\boldsymbol{\mu}_t + f_\Sigma(\boldsymbol{\lambda}_t^2)\odot\boldsymbol{C}^\mathsf{T}\boldsymbol{\epsilon}_t\ , \tag{57}$$

$$\partial_t \boldsymbol{\epsilon}_t = -\frac{1}{\tau_\epsilon}\left(\boldsymbol{\epsilon}_t - \frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} + \frac{1}{\sigma_{\text{obs}}^2}\odot\boldsymbol{C}\boldsymbol{\mu}_t\right) = -\frac{1}{\tau_\epsilon}\boldsymbol{\epsilon}_t + \frac{1}{\tau_\epsilon}\frac{\boldsymbol{v}_t}{\sigma_{\text{obs}}^2} - \frac{1}{\tau_\epsilon}\frac{1}{\sigma_{\text{obs}}^2}\odot\boldsymbol{C}\boldsymbol{\mu}_t\ , \tag{58}$$

where we have made elementwise multiplication explicit via the $\odot$-operator, and moved the prediction error, $\boldsymbol{\epsilon}_t$, into a separate dynamic equation with time constant τ_ϵ. For this separation to maintain faithful results, we require that $\tau_\epsilon < \tau_s$, such that the prediction error can react to changes in $\boldsymbol{\mu}_t$. The only variable in eqn. (54) that cannot be calculated within this scheme is $f_\Sigma(\boldsymbol{\lambda}^2)$, which contains a square root, and, thus, has to be addressed separately, below. The variables $\boldsymbol{v}_t/\sigma_{\text{obs}}^2$ and $1/\sigma_{\text{obs}}^2$ are the input variables that are fed into the system.

4.4 Neural dynamics for integrating linear and quadratic differential equations

We will now establish how linear and quadratic dynamics of latent variables, such as eqn. (56) – (58), can be integrated in neural space. What follows is basically a clearly structured generalization of the ideas presented in Ref. [12].

Notation in the algorithmic domain. Inevitably, some notation has to be introduced for addressing all of the above dynamics in both the *algorithmic domain* (i.e., dynamics of variables) and the *network domain* (i.e., dynamics of neuronal firing rates). To keep the presentation general, we will adopt variable-dynamics of the generic form

$$\partial_t \boldsymbol{z} = \boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \qquad \text{with} \qquad (\boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x})_i \overset{\text{def}}{=} \sum_{j,k}\underline{Q}_{ijk}y_j x_k\ . \tag{59}$$

Here, $\boldsymbol{z}$, $\boldsymbol{y}$, and $\boldsymbol{x}$ denote vector-valued variables. Further, $\boldsymbol{b}$ is a vector-valued additive constant, $\boldsymbol{W}$ a matrix, and $\underline{\boldsymbol{Q}}$ a 3rd-order tensor, for which we establish the notation of small, capital, and underlined capital letters, respectively.

Note that expressions with elementwise multiplication are covered by the tensors. For instance, using Einstein summation convention,

$$(\boldsymbol{y} \odot \boldsymbol{x})_i = y_i\, x_i = \delta_{ij}\delta_{ik} y_j\, x_k = (\boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x})_i \qquad \text{with} \qquad \underline{Q}_{ijk} = \delta_{ij}\delta_{ik}\ , \tag{60}$$

$$\left[\boldsymbol{W}^1 (\boldsymbol{y} \odot \boldsymbol{W}^2 \boldsymbol{x})\right]_i = W^1_{ij}(\boldsymbol{y} \odot \boldsymbol{W}^2\boldsymbol{x})_j = W^1_{ij}(y_j W^2_{jk} x_k) = (\boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x})_i \qquad \text{with} \qquad \underline{Q}_{ijk} = W^1_{ij} W^2_{jk}\ , \tag{61}$$

$$\left[(\boldsymbol{W}^1\boldsymbol{y}) \odot (\boldsymbol{W}^2\boldsymbol{x})\right]_i = W^1_{ij} y_j W^2_{ik} x_k = (\boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x})_i \qquad \text{with} \qquad \underline{Q}_{ijk} = W^1_{ij} W^2_{ik}\ . \tag{62}$$

Thus, all algorithmic dynamics in eqn. (56) – (58) are of form eqn. (59).

Notation in the network domain. We now turn to the question of how neuronal populations can calculate dynamics of the form in eqn. (59) when the involved variables are linearly decodable, that is, when $\boldsymbol{z} = \boldsymbol{A}^z \boldsymbol{r}^z$. Refining the notation in eqn. (55), we will make explicit which population $\boldsymbol{r}^z$ encodes variable $\boldsymbol{z}$ and suppress the time dependence in $\boldsymbol{r}^z_t$. Again, we emphasize that differently denoted populations, e.g., $\boldsymbol{r}^z$ and $\boldsymbol{r}^x$, can and often will refer to the same population—the refined notation simply gives us the flexibility to cover various cases.

Following [12], we will further make use of what is called the *adjoint matrix* $\boldsymbol{A}^{z\dagger}$ of matrix $\boldsymbol{A}^z$, which is characterized by $\boldsymbol{A}^z \boldsymbol{A}^{z\dagger} = \boldsymbol{I}$. Such right-inverse, albeit not unique, always exists if the rows of $\boldsymbol{A}^z$ are linearly independent, which is commonly fulfilled when the number of neurons exceeds the number of variables. If variables $\boldsymbol{z}$ and $\boldsymbol{x}$ are encoded by the same populations of neurons, we further require that $\boldsymbol{A}^z \boldsymbol{A}^{x\dagger} = \boldsymbol{A}^x \boldsymbol{A}^{z\dagger} = \boldsymbol{0}$. As long as these orthogonality conditions are satisfied, the exact form of the matrices $\boldsymbol{A}$ is arbitrary, from a mathematical point of view.

As we will show in the next paragraph, quadratic, linear, and constant terms in the algorithmic domain, eqn. (59), translate one-to-one into quadratic, linear and constant terms in the network domain. We therefore establish the notation $\overline{\underline{\boldsymbol{Q}}}$, $\overline{\boldsymbol{W}}$, and $\overline{\boldsymbol{b}}$ (with an overbar) to refer to 3rd-order tensors, matrices and biases in the neural domain, respectively. These are exactly the function arguments that we had deemed feasible in eqn. (53) (there introduced without the overbar).

Neural dynamics. The neural dynamics for implementing each of the computations in eqn. (59) are as follows.

Quadratic terms: $\partial_t \boldsymbol{z} = \boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x}$ is implemented via

$$\partial_t \boldsymbol{r}^z = \boldsymbol{r}^y \overline{\underline{\boldsymbol{Q}}}\, \boldsymbol{r}^x \qquad \text{with} \qquad \overline{\underline{Q}}_{ijk} \stackrel{\text{def}}{=} A^{z\dagger}_{i\alpha} \underline{Q}_{\alpha\beta\gamma} A^y_{\beta j} A^x_{\gamma k}\ , \qquad \text{in short:} \qquad \overline{\underline{\boldsymbol{Q}}} \stackrel{\text{def}}{=} \boldsymbol{A}^{z\dagger}(\underline{\boldsymbol{Q}} \boldsymbol{A}^y \boldsymbol{A}^x)\ . \tag{63}$$

Proof:

$$(\partial_t \boldsymbol{z})_i = (\partial_t \boldsymbol{A}^z \boldsymbol{r}^z)_i = (\boldsymbol{A}^z \partial_t \boldsymbol{r}^z)_i = A^z_{ij}(\boldsymbol{r}^y \overline{\underline{\boldsymbol{Q}}} \boldsymbol{r}^x)_j = A^z_{ij} \overline{\underline{Q}}_{jkl} r^y_k r^x_l = A^z_{ij} A^{z\dagger}_{j\alpha} \underline{Q}_{\alpha\beta\gamma} A^y_{\beta k} A^x_{\gamma l} r^y_k r^x_l = \delta_{i\alpha} \underline{Q}_{\alpha\beta\gamma} y_\beta x_\gamma = (\boldsymbol{y}\underline{\boldsymbol{Q}}\boldsymbol{x})_i\ . \tag{64}$$

Linear terms: $\partial_t \boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$ is implemented via

$$\partial_t \boldsymbol{r}^z = \overline{\boldsymbol{W}}\, \boldsymbol{r}^x \qquad \text{with} \qquad \overline{\boldsymbol{W}} = \boldsymbol{A}^{z\dagger} \boldsymbol{W} \boldsymbol{A}^x\ . \tag{65}$$

Proof:

$$\partial_t \boldsymbol{z} = \boldsymbol{A}^z \partial_t \boldsymbol{r}^z = \boldsymbol{A}^z \overline{\boldsymbol{W}} \boldsymbol{r}^x = \boldsymbol{A}^z \boldsymbol{A}^{z\dagger} \boldsymbol{W} \boldsymbol{A}^x \boldsymbol{r}^x = \boldsymbol{W}\boldsymbol{x}\ . \tag{66}$$

Constant terms: $\partial_t \boldsymbol{z} = \boldsymbol{b}$ is implemented via

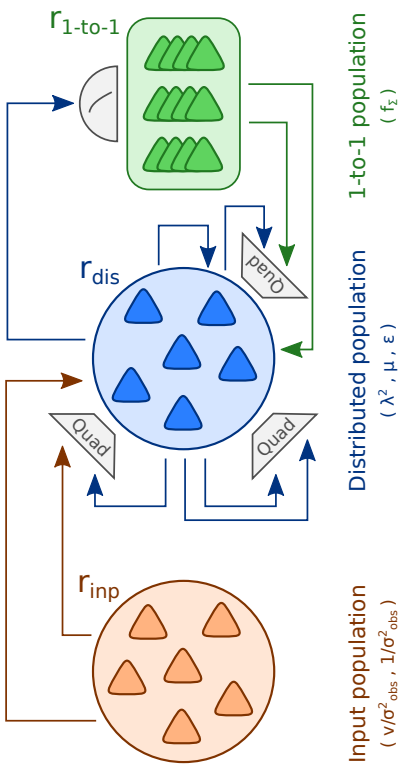
$$\partial_t \boldsymbol{r}^z = \overline{\boldsymbol{b}} \qquad \text{with} \qquad \overline{\boldsymbol{b}} = \boldsymbol{A}^{z\dagger} \boldsymbol{b}\ . \tag{67}$$

Proof:

$$\partial_t \boldsymbol{z} = \boldsymbol{A}^z \partial_t \boldsymbol{r}^z = \boldsymbol{A}^z \overline{\boldsymbol{b}} = \boldsymbol{A}^z \boldsymbol{A}^{z\dagger} \boldsymbol{b} = \boldsymbol{b}\ . \tag{68}$$

Linearity: Neural dynamics for linear combinations, e.g., $\partial_t \boldsymbol{z} = \boldsymbol{W}^1\boldsymbol{x} + \boldsymbol{W}^2\boldsymbol{y}$ are simply the sum of the individual terms, e.g., $\partial_t \boldsymbol{r}^z = \overline{\boldsymbol{W}^1}\boldsymbol{r}^x + \overline{\boldsymbol{W}^2}\boldsymbol{r}^y$. The proof follows directly from the linearity of $\boldsymbol{A}^z$.

Shared populations: Neural dynamics of variables encoded by the same population, e.g., $\boldsymbol{r} = \boldsymbol{r}^z = \boldsymbol{r}^x$, do not interfere. The proof follows from the orthogonality $\boldsymbol{A}^z \boldsymbol{A}^{x\dagger} = \boldsymbol{A}^x \boldsymbol{A}^{z\dagger} = \boldsymbol{0}$ and from observing that every term in the above neural dynamics is led by an adjoint matrix $\boldsymbol{A}^{z\dagger}$ or $\boldsymbol{A}^{x\dagger}$. Therefore, neural dynamics inducing changes in $\boldsymbol{z}$ do not convey any changes in $\boldsymbol{x}$, and vice versa.



Supplementary Fig. 10 | Network model for motion structure inference. The network is composed of three neuronal populations. The input population, r_{inp}, encodes the input variables, $1/\sigma^2_{\text{obs}}$ and v/σ^2_{obs}, as a distributed code. The distributed population, r_{dis}, encodes the latent variables, λ^2, $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$, as a distributed code. The one-to-one population, $r_{\text{1-to-1}}$, encodes the latent posterior uncertainty, $f_\Sigma(\lambda^2)$, as a one-to-one code. All of these variables can be read out linearly from the network firing rate, at any time. Synaptic connections within and between populations mediate linear (indicated as arrows) and quadratic (indicated as "Quad" boxes) interactions. The non-linear function f_Σ is implemented by a leaky integrate-and-fire type response (indicated by the half-circle).
Overall, the network implements eqn. (56) – (58) of the algorithmic domain and, thereby, emulates the adiabatic observer model given by eqn. (43) + (44).

4.5 Recurrent network model for online motion structure inference

Supplementary Note 4, Section 4 provides us with a straight-forward recipe for implementing eqn. (56) – (58) in a neural network. To keep the network as general as possible, the input variables, $1/\sigma^2_{\text{obs}}$ and v/σ^2_{obs}, are encoded by an *input population*, $\boldsymbol{r}_{\text{inp}}$. The latent variables, $\boldsymbol{\lambda}^2$, $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$, are encoded by a *distributed population*, $\boldsymbol{r}_{\text{dis}}$. Both the input and distributed population employ a distributed code with arbitrary readout matrices $\boldsymbol{A}$ obeying the orthonormality conditions stated in **Supplementary Note 4, Section 4**. For the distributed population, the activation function, f_i, in eqn. (53) is simply the identity function. We refrain from restating the exact neural dynamics here because they are obtained directly by translating the terms in eqn. (56) – (58) into their neural counterparts by means of eqn. (63), (65), and (67).

For a functioning network model, however, two pieces are missing: the input code, and handling of the function $f_\Sigma(\lambda_t^2)$ as was promised in **Supplementary Note 4, Section 3**. These two pieces are discussed next.

Connecting the input. The input is fed into the network externally and is thus by definition not controlled by internal dynamics of the network. Nonetheless, the activity $\boldsymbol{r}_{\text{inp}}(v/\sigma^2_{\text{obs}}, 1/\sigma^2_{\text{obs}})$, which is a function of the input variables, is required to support linearly decoding v/σ^2_{obs} and $1/\sigma^2_{\text{obs}}$ via known readout matrices $\boldsymbol{A}^v$ and $\boldsymbol{A}^\sigma$. Note that no adjoint matrices are required for the input. While any valid input code can be used in our generic network model, finding activation functions grounded in biological experiments together with matching readout matrices is typically non-trivial. We present one such input model, which resembles fundamental response properties of area MT, in **Supplementary Note 4, Section 6**.

Handling $f_\Sigma(\lambda_t^2)$. The non-linearity of the function $f_\Sigma(\lambda_t^2)$, given by eqn. (42), prohibits a direct incorporation of the effect of $\boldsymbol{\lambda}^2$ on its dependent variables in eqn. (56) and eqn. (57), within the computational framework of **Supplementary Note 4, Section 4**. The core reason is that the linear readout $\boldsymbol{A}^\lambda \boldsymbol{r}_{\text{dis}}$ does not commute with the square-root function. Yet, it turns out that $f_\Sigma(\lambda_t^2)$ can be incorporated into the network model quite easily owing to its simple functional form. Since f_Σ keeps all motion components separate, $f_\Sigma(\lambda_m^2) = \text{const}_m \cdot (-1 + \sqrt{1 + \text{const}_m\, \lambda_m^2})$, we can employ a dedicated population, $\boldsymbol{r}_{\text{1-to-1}}$, using a one-to-one coding model, which assigns one neuron (or small sub-population) to each component of the posterior variance:

$$r_{\text{1-to-1},m} = \frac{1}{A^\Sigma_{mm}} f_\Sigma(A^\lambda_{m\star} \boldsymbol{r}_{\text{dis}}) \ . \tag{69}$$

Here we have used that $\lambda_m^2 = \boldsymbol{A}^{\lambda}_{m\star}\boldsymbol{r}_{\text{dis}}$ can be read out linearly from the population's activity. The coefficient A^{Σ}_{mm} scales the firing rate of $r_{\text{1-to-1},m}$. This leads to a neurally plausible activation function of $r_{\text{1-to-1}}$: In **Supplementary Fig. 9**, replace λ_m^2 by the "input current" $\boldsymbol{A}^{\lambda}_{m\star}\boldsymbol{r}_{\text{dis}}$ on the x-axis, and f_Σ by $r_{\text{1-to-1}}$ on the y-axis. This is reminiscent of the firing response of leaky integrate-and-fire neurons or, more generally, Type I neurons, as a function of the input current. Finally, we can read out $\boldsymbol{f}_\Sigma$ linearly from $\boldsymbol{r}_{\text{1-to-1}}$ via readout matrix $\boldsymbol{A}^{\Sigma}=\text{diag}\big[(A^{\Sigma}_{11},..,A^{\Sigma}_{MM})\big]$, thereby allowing us to apply the formalism of **Supplementary Note 4, Section 4** to $f_\Sigma(\boldsymbol{\lambda}^2)$ (which acts as a variable).

In the above argumentation, we have made two simplifying assumption. First, we have assumed an instantaneous response for $\boldsymbol{r}_{\text{1-to-1}}$ instead of the low-pass filtered response of eqn. (53). Since f_Σ varies only on the long time scale τ_λ, eqn. (69) could easily be replaced by a low-pass filtered version with $f_i := f_\Sigma / A^{\Sigma}_{mm}$ being the neurons' activation function in eqn. (53). Secondly, we notice that, strictly, f_Σ depends not only on $\boldsymbol{\lambda}^2$, but also on $1/\sigma^2_{\text{obs}}$. While the quadratic interaction between these variables, as expressed by eqn. (42), is covered by the theory, we decided to reduce the complexity of the network model by assuming a fixed default value for σ^2_{obs} in the computer simulations. Again, an extension respecting the explicit σ^2_{obs}-dependence would be straight-forward.

The complete network model. Plugging all of the components together, we obtain the network model shown in **Supplementary Fig. 10**. This network emulates the adiabatic observer model given by eqn. (43) + (44).

4.6 Neural coding of the input: an example for area MT

While we aimed to leave the neural code for all *latent* variables as generic as possible in the network examples, we specify an *input* code that respects known response properties of area MT. In the following, we present the tuning functions for the input neurons which are derived from models and properties in the literature on area MT [15–17]. Their most important computational property is that they support linear readout of $\boldsymbol{v}/\sigma^2_{\text{obs}}$ and $1/\sigma^2_{\text{obs}}$ in Cartesian coordinates.

We will proceed in three steps. First, we define the tuning functions in polar coordinates since response properties are commonly presented in this coordinate system in the experimental literature. Second, we state some helpful mathematical properties of the proposed tuning functions. Third, we provide the readout matrices $\boldsymbol{A}^v$ and $\boldsymbol{A}^\sigma$ and demonstrate how they accurately decode the relevant variables, $\boldsymbol{v}/\sigma^2_{\text{obs}}$ and $1/\sigma^2_{\text{obs}}$.

The tuning function in polar coordinates. Commonly, MT tuning is characterized in polar coordinates because the tuning function becomes separable, i.e., it factorizes into a product of functions over different variables. Owing to the retinotopic organization of MT, we consider for the remainder of this subsection a local population of neurons sharing the same spacial receptive field. Those neurons are indexed, as shown in **Supplementary Fig. 11 (left)**, by $n_\alpha = 1..N_\alpha$ and $n_\rho = 1..N_\rho$ according to their preferred direction, μ_α, and speed (absolute value of velocity), μ_ρ, respectively. We use the following tuning function in response to a stimulus with direction $\alpha \in [0, 2\pi)$, speed $\rho \geq 0$, and observation noise σ^2_{obs}:

$$f(\alpha, \rho, \sigma^2_{\text{obs}}; n_\alpha, n_\rho) = f_\sigma(\sigma^2_{\text{obs}}) \cdot f_\alpha(\alpha; n_\alpha) \cdot f_\rho(\rho; n_\rho) \tag{70}$$

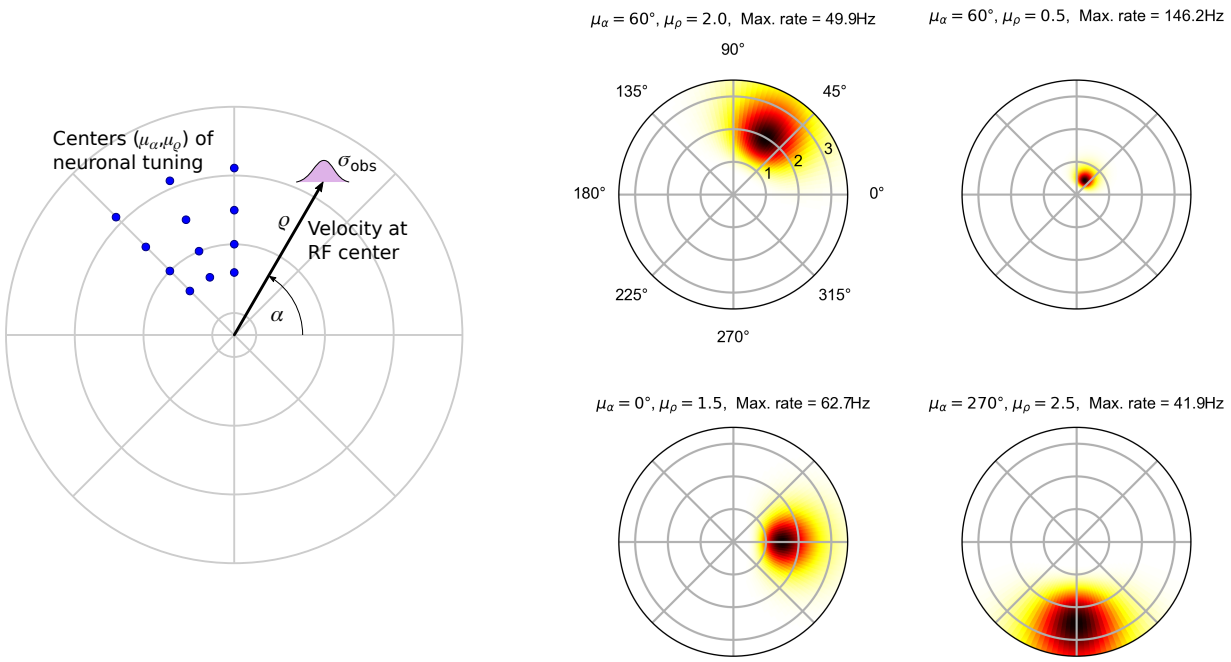
with

$$f_\sigma(\sigma^2_{\text{obs}}) = \frac{\psi}{\sigma^2_{\text{obs}}} \ , \tag{71}$$

$$f_\alpha(\alpha; n_\alpha) = \frac{d_\alpha}{2\pi I_0(\kappa_\alpha)}\, e^{\kappa_\alpha \cos(\alpha - d_\alpha n_\alpha)} \ , \tag{72}$$

$$f_\rho(\rho; n_\rho) = \frac{\mu'_\rho(n_\rho)}{\sqrt{2\pi\sigma^2_\rho}\,\mu_\rho(n_\rho)}\, e^{-\frac{\left(\log(\rho) - \log(\mu_\rho(n_\rho))\right)^2}{2\sigma^2_\rho}} \ . \tag{73}$$

Example tuning functions are shown in **Supplementary Fig. 11 (right)**. Eqn. (70) is composed of sub-functions for the noise f_σ, motion direction f_α, and motion speed f_ρ, which employ a range of parameters: The overall (maximum) firing rate is scaled by ψ. The angle between cells' preferred direction is $d_\alpha = 2\pi/N_\alpha$, such that neuron n_α's preferred direction is $d_\alpha n_\alpha$. The directional tuning width is described by κ_α (formally, κ_α is the precision parameter of a von-Mises-distribution density function, and $I_0(\kappa_\alpha)$ is the modified Bessel function of order 0 for normalization). Neuron n_ρ's preferred speed is given by function $\mu_\rho(n_\rho)$, with μ'_ρ denoting the function's derivative. Finally, the width of speed tuning is controlled by σ^2_ρ. Let us briefly discuss how eqn. (70) captures known properties of MT:



Supplementary Fig. 11 | Tuning functions of MT neurons. *Left:* The tuning of MT neurons, in response to a stimulus with direction α, speed ρ and observation noise σ_{obs}, is separable in polar coordinates. In a local population, the neurons' preferred velocity tuning covers directions, μ_α, uniformly, while the density of neurons tuned to speed, μ_ρ, decreases for higher speed. Note that all quantities refer to a local coordinate system centered at the receptive field (RF) center. So, the coordinates here are not to be confused with the coordinate system in **Supplementary Fig. 8** which describes RF locations. *Right:* Tuning function according to eqn. (70) for four example neurons. Tuning centers and maximum firing rates are given in the axes titles. Parameters (using Python indexing, i.e., $n_\rho{=}0,..,N_\rho{-}1$): $\psi{=}0.1$, $\sigma^2_{\text{obs}}{=}(0.05/3)^2$, $N_\alpha{=}16$, $N_\rho{=}12$, $\mu_\rho(n_\rho) = \rho_{\text{min}} + d_\rho\, n_\rho^{1.25}$, $d_\rho = (\rho_{\text{max}} - \rho_{\text{min}})/(N_\rho - 1)^{1.25}$, $\rho_{\text{min}}{=}0.1$, $\rho_{\text{max}}{=}8.0$, $\kappa_\alpha{=}1/0.35^2$, $\sigma^2_\rho{=}0.35^2$.

- Neurons are tuned to speed (absolute value of velocity) and direction (almost entirely into only one direction, not the opposite direction).
- Direction tuning is commonly described by a von Mises density function. Preferred directions cover the circle roughly isotropically, here via $d_\alpha\, n_\alpha$.
- Speed tuning can be described by a log-normal function of the speed ρ. The density of speed tuning centers in MT has been reported to decrease for larger speeds, which can be captured by the function $\mu_\rho(n_\rho)$.
- Activity is modulated by contrast (via σ^2_{obs}), with lower contrast (higher noise) attenuating the overall firing rate.

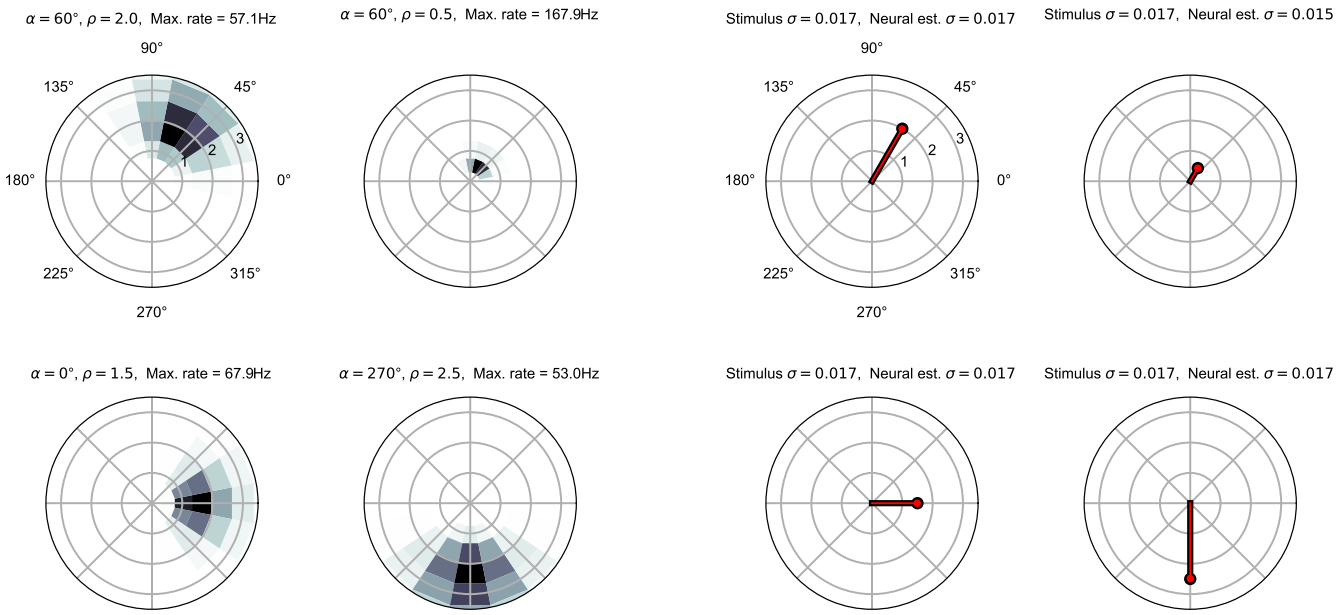
We make the simplifying assumptions that (i) all neurons have the same firing rate scaling factor ψ, and (ii) that the tuning widths, given by κ_α and σ^2_ρ, are "not too wide". The meaning of "not too wide" will become clear in the following mathematical consideration.

Mathematical properties of the tuning function. We next discuss some useful properties of the components of the above tuning function. First, we note that f_α and f_ρ have the form of known probability density functions over the neuron indices n_α and n_ρ, respectively. In particular, they integrate to one in the limit of many, narrowly spaced neurons:

$$\int_0^{\frac{2\pi}{d_\alpha}} f_\alpha(\alpha; n_\alpha)\, dn_\alpha = \int_0^{2\pi} e^{\kappa_\alpha \cos(\alpha - d_\alpha n_\alpha)} / (2\pi I_0(\kappa_\alpha))\, d(d_\alpha n_\alpha) = 1 \tag{74}$$

and

$$\int_0^{\infty} f_\rho(\rho; n_\rho)\, dn_\rho = \int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_\rho}\,\mu_\rho} e^{-\frac{(\log(\rho) - \log(\mu_\rho))^2}{2\sigma^2_\rho}}\, d\mu_\rho = 1\ . \tag{75}$$



Supplementary Fig. 12 | Linear readout of input statistics from population responses. *Left:* Population response when encoding four example stimuli. The population consists of 192 neurons with the parameters given in **Supplementary Fig. 11**. For clarity, only neurons with $\mu_\rho < 3.5$ are shown. *Right:* Linear readout of $1/\sigma^2_{\text{obs}}$, $v_x/\sigma^2_{\text{obs}}$ and $v_y/\sigma^2_{\text{obs}}$ from the population activities on the left via weights given by eqn. (78). Shown are, in polar coordinates, the stimulus ground truth (black) and the estimate by the linear readout (red). The estimated uncertainty is provided is the axes titles.

Furthermore, the distributional forms give rise to nice moments w.r.t. the tuning centers $d_\alpha\, n_\alpha$ and $\mu_\rho(n_\rho)$:

$$\langle e^{i d_\alpha n_\alpha} \rangle_{f_\alpha} = \frac{I_1(\kappa_\alpha)}{I_0(\kappa_\alpha)} e^{i\alpha} \overset{\text{large } \kappa_\alpha}{\approx} e^{i\alpha} = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix} \tag{76}$$

$$\text{and} \qquad \langle \mu_\rho(n_\rho) \rangle_{f_\rho} = \rho e^{\sigma_\rho^2/2} \overset{\text{small } \sigma_\rho}{\approx} \rho \ . \tag{77}$$

We now understand how narrow (i.e., "not too wide") tuning functions enable reading out the encoded direction, α, and speed, ρ: large κ_α and small σ_ρ^2 afford the approximations in eqn. (76) and (77). Further, the mathematical relations highlight that κ_α and σ_ρ^2 could be modulated by the observation noise σ^2_{obs} without changing the ability to encode/decode the input.

Linear readout of input statistics. With the above mathematical properties at hand, we identify matrices $\boldsymbol{A}^\sigma$ and $\boldsymbol{A}^v$ for linear readout:

$$A^\sigma_{k,(n_\alpha,n_\rho)} = \frac{1}{\psi} \qquad \text{and} \qquad A^v_{k,(n_\alpha,n_\rho)} = \frac{1}{\psi} \begin{pmatrix} \cos d_\alpha n_\alpha \\ \sin d_\alpha n_\alpha \end{pmatrix} \mu_\rho(n_\rho) \ , \tag{78}$$

because reading out with these matrices from the MT-population of the k-th observable yields:

$$\int\int A^\sigma_{k,(n_\alpha,n_\rho)} f(\alpha_k, \rho_k, \sigma^2_{\text{obs},k}; n_\alpha, n_\rho)\, dn_\alpha\, dn_\rho = \frac{1}{\sigma^2_{\text{obs},k}} \tag{79}$$

$$\int\int A^v_{k,(n_\alpha,n_\rho)} f(\alpha_k, \rho_k, \sigma^2_{\text{obs},k}; n_\alpha, n_\rho)\, dn_\alpha\, dn_\rho = \frac{\rho_k}{\sigma^2_{\text{obs},k}} \begin{pmatrix} \cos\alpha_k \\ \sin\alpha_k \end{pmatrix} = \begin{pmatrix} v_x/\sigma^2_{\text{obs},k} \\ v_y/\sigma^2_{\text{obs},k} \end{pmatrix} \ . \tag{80}$$

Examples of the population response to four motion stimuli is shown in **Supplementary Fig. 12 (left)** for a population of 192 neurons. The tuning centers span 12 radii ("speed") and 16 angles ("direction"). In **Supplementary Fig. 12 (right)**, the resulting linear readout (red) is shown alongside the ground truth (black). Even the coarse coverage of the stimulus space by 192 neurons is sufficient for a faithful reconstruction of the stimulus. As an interesting observation, while neurons tuned to higher speeds have wider tuning curves, this does not imply that their activity would encode heightened uncertainty: σ_{obs} is identical in all of the examples in **Supplementary Fig. 11** and **Supplementary Fig. 12**.

In summary, we have identified with eqn. (70) an MT tuning function that supports linear readout of the variables v/σ^2_{obs} and $1/\sigma^2_{\text{obs}}$ via matrices $\boldsymbol{A}^v$ and $\boldsymbol{A}^\sigma$.

Supplementary References

1. Bill, J., Pailian, H., Gershman, S. J. & Drugowitsch, J. Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences* **117,** 24581–24589 (2020).
2. Gardiner, C. *Stochastic methods* (Springer Berlin, 2009).
3. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nature neuroscience* **5,** 598–604 (2002).
4. Manning, C., Thomas, R. T. & Braddick, O. Can speed be judged independent of direction? *Journal of vision* **18,** 15–15 (2018).
5. Moscatelli, A., La Scaleia, B., Zago, M. & Lacquaniti, F. Motion direction, luminance contrast, and speed perception: an unexpected meeting. *Journal of vision* **19,** 16–16 (2019).
6. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22 (1977).
7. Bishop, C. M. *Pattern recognition and machine learning* **4** (Springer, 2006).
8. Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82,** 35–45 (1960).
9. Kalman, R. E. & Bucy, R. S. New results in linear filtering and prediction theory. *Journal of Basic Engineering* **83,** 95–108 (1961).
10. Jazwinski, A. H. *Stochastic processes and filtering theory* (Courier Corporation, 2007).
11. Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience* **32,** 3612–3628 (2012).
12. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience* **31,** 15310–15319 (2011).
13. Dayan, P. & Abbott, L. F. *Theoretical neuroscience: computational and mathematical modeling of neural systems* (Computational Neuroscience Series, 2001).
14. Salinas, E. & Abbott, L. F. A model of multiplicative neural responses in parietal cortex. *Proceedings of the national academy of sciences* **93,** 11956–11961 (1996).
15. Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu. Rev. Neurosci.* **28,** 157–189 (2005).
16. Nover, H., Anderson, C. H. & DeAngelis, G. C. A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *Journal of Neuroscience* **25,** 10049–10060 (2005).
17. Krekelberg, B., Van Wezel, R. J. & Albright, T. D. Interactions between speed and contrast tuning in the middle temporal area: implications for the neural code for speed. *Journal of Neuroscience* **26,** 8988–8998 (2006).