Hierarchical structure learning for perceptual decision making in visual motion perception

Johannes Bill
Department of Neurobiology
Harvard Medical School
Boston, MA 02115
johannes_bill@hms.harvard.edu

Samuel J. Gershman
Department of Psychology
Harvard University
Cambridge, MA 02138
gershman@fas.harvard.edu

Jan Drugowitsch
Department of Neurobiology
Harvard Medical School
Boston, MA 02115
jan_drugowitsch@hms.harvard.edu

Abstract

Successful behavior in the real world critically depends on discovering the latent structure behind the volatile inputs reaching our sensory system. Our brains face the online task of discovering structure at multiple timescales ranging from short-lived correlations, to the structure underlying a scene, to life-time learning of causal relations. Little is known about the mental and neural computations driving the brain's ability of online, multi-timescale structure inference. We studied these computations by the example of visual motion perception owing to the importance of structured motion for behavior. We propose online hierarchical Bayesian inference as a principled solution for how the brain might solve multi-timescale structure inference. We derive an online Expectation-Maximization algorithm that continually updates an estimate of a visual scene's underlying structure while using this inferred structure to organize incoming noisy velocity observations into meaningful, stable percepts. We show that the algorithm explains human percepts qualitatively and quantitatively for a diverse set of stimuli, covering classical psychophysics experiments, ambiguous motion scenes, and illusory motion displays. It explains experimental results of human motion structure classification with higher fidelity than a previous ideal observer-based model, and provides normative explanations for the origin of biased perception in motion direction repulsion experiments. To identify a scene's structure the algorithm recruits motion components from a set of frequently occurring features, such as global translation or grouping of stimuli. We demonstrate in computer simulations how these features can be learned online from experience. Finally, the algorithm affords a neural network implementation which shares properties with motion-sensitive cortical areas MT and MSTd and motivates a novel class of neuroscientific experiments to reveal the neural representations of latent structure.

Keywords: hierarchical structure, online learning, Bayesian inference, visual perception, biological neural network

Acknowledgements

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

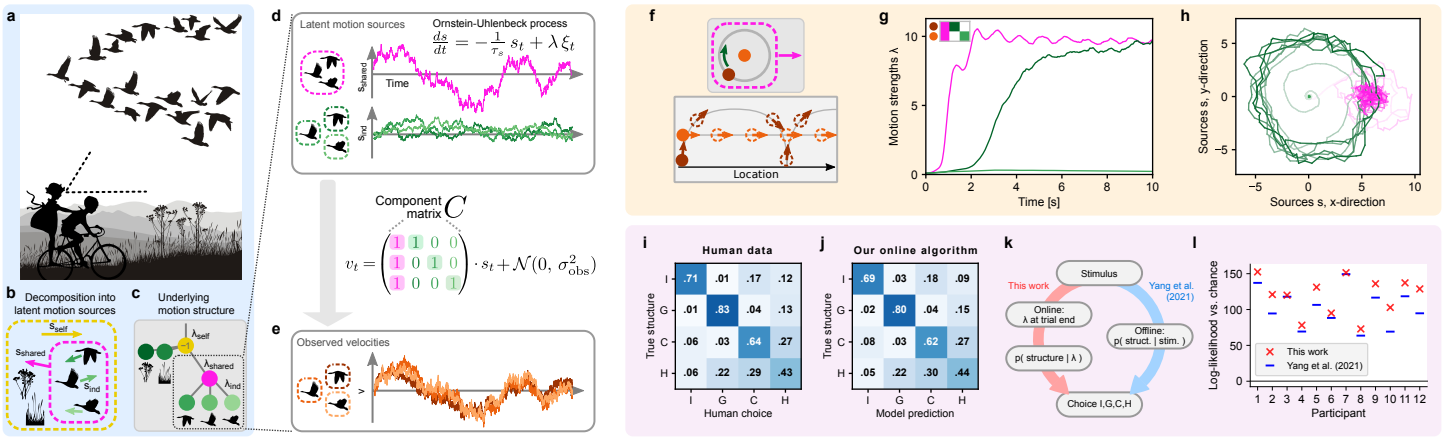


Figure 1: Explaining perception of structured motion as online hierarchical inference. ***(a – e)*** *Generative model of structured motion. Our online algorithm inverts the generative model to simultaneously identify the underlying structure and to decompose observed velocities into (latent) motion sources. See main text for details.* ***(f – h)*** *Demonstration of the algorithm by the example of the Duncker wheel.* ***(f)*** *The Duncker wheel is a rolling wheel with only two visible point lights at the hub and rim.* ***(g)*** *Like humans, the algorithm discovers a shared component for both lights (pink) and an individual component for the rim light (dark green).* ***(h)*** *The lights' velocities are decomposed into joint rightward motion and rim light-rotation. Brightness = time.* ***(i – l)*** *Classification task of ambiguous motion scenes, analyzing behavioral data from [9].* ***(i)*** *Human confusion matrix when classifying ambiguous motion scenes as Independent, Global, Clustered, and Hierarchically nested motion.* ***(j)*** *Confusion matrix of the algorithm on the same trials as [9]. The algorithm quantitatively explains human percepts and even captures the fine-structure in the confusion matrix.* ***(k)*** *For this, we fed the value of $\boldsymbol{\lambda}_t$ at trial end into a logistic regression classifier (trained on the ground truth, not human responses), and fitted the same choice model as [9], who had employed the ideal posterior on the full input trajectory.* ***(l)*** *Log-likelihood of human responses under both models. Our algorithm explains human responses better for every participant. The results in panels (j) and (l) are cross-validated.*

Introduction

Real-world scenes feature rich spatial and temporal structure. Understanding this structure allows humans and animals to make sense of their environment by organizing complex and often ambiguous sensory input streams into stable, meaningful percepts. The emerging compressed representations benefit goal-directed actions and decision making. While machine learning algorithms have been developed to infer the structure of large datasets offline [1], an understanding of how biological agents discover structure online is only beginning to emerge [2].

We studied online structure inference across multiple timescales by the example of visual motion perception. Motion structure, that is, statistical relations in velocities, carries essential information about the spatial and temporal evolution of the environment. For instance, the features composing an object typically move coherently, or self-motion adds optic flow to all observable velocities in a scene. To benefit behaviors such as navigation, tracking, prediction, and pursuit, the visual system must decompose observable velocities, $\boldsymbol{v}_t$, (see **Fig. 1a**) into their putative latent origins, e.g., the observer's self-, shared flock-, and each bird's individual motion (**Fig. 1b**).

Bayesian inference has provided a successful normative framework for understanding human visual motion perception in spatially constrained (local) patches [3, 4] and for simple structures [5, 6]. For structured motion spanning multiple objects, larger areas of the visual field, and longer timescales, however, a comprehensive theoretical description is only beginning to emerge. Recent work [7] has introduced tree structures for the mental organization of observed velocities into nested hierarchies, yet the inference process over structures had to be performed by a biologically unrealistic offline sampling algorithm. Theory-driven experiments have revealed that the human visual system makes use of hierarchical structure when solving visual tasks [8], and that aspects of motion structure perception can be explained by Bayesian structural inference [9], yet the algorithms underlying the structure discovery remained elusive.

We address the questions of how the visual system solves the chicken-and-egg problem of parsing motion in a scene in real time while simultaneously inferring the scene's underlying structure, and how the involved probabilistic computations can be performed by neural circuits. We propose an online Expectation-Maximization (EM) algorithm which leverages the fact that instantaneous motion (e.g., the speed and direction of flocking birds) and a scene's structure (e.g., the presence of a flock) evolve on different timescales. We demonstrate that the derived algorithm replicates a range of psychophysics experiments both qualitatively and quantitatively. Further, we present an implementation of the algorithm by recurrent neural networks, which feature connection and response properties of cortical areas implicated in visual motion processing, and propose a targeted experiment to test model predictions in neural recordings. Finally, we explore how the set of

typical motion components, such as shared motion or grouping, which the algorithm draws upon when explaining the structure of a scene, could be learned from experience on long timescales, rather than being given to it as a parameter.

Part of this work has been posted as a pre-print, `https://www.biorxiv.org/content/10.1101/2021.10.21.465346v1`, which presents the derivations and results in detail. The simulation results on online learning of the motion components are a new contribution.

Online hierarchical inference in a generative model of structured motion.

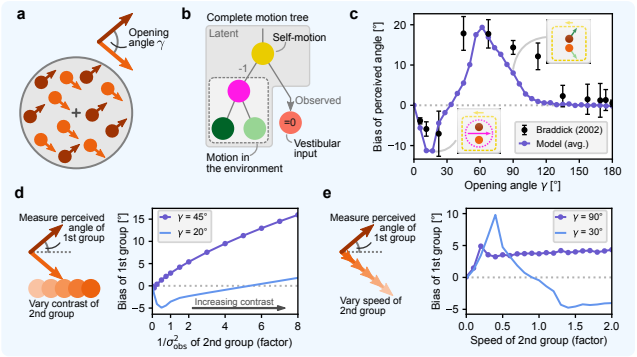
We build on the generative model of structured motion from [8] in which observable velocities, $\boldsymbol{v}_t$, are generated from latent causes, $\boldsymbol{s}_t$. The so-called *motion sources*, $\boldsymbol{s}_t$, can have volatile speed and direction, and usually respect a temporally more robust, tree-shaped *motion structure*: in **Fig. 1c**, graph connectivity defines how sources (nodes in the graph) affect observable objects (black object outlines), and vertical edge length, called *motion strength*, λ_m, indicates the long-term average speed of the associated source, s_m. Sources are a-priori assumed to evolve as Ornstein-Uhlenbeck processes in each spatial dimension (**Fig. 1d**, only 1 dim. shown) leading to stationary distributions, $s_m \sim \mathcal{N}\left(0,\, \frac{\tau_s}{2}\lambda_m^2\right)$. Observed velocities, $\boldsymbol{v}_t$, are noisy versions of the sum of all ancestral sources in the graph with graph connectivity represented by the *component matrix*, $\boldsymbol{C}$ (see **Fig. 1d & e** for illustration of 3 flocking birds). For most of the following, we assume that the component matrix, $\boldsymbol{C}$, is given and fixed, e.g., because it has been learned from experience.

We derived an online EM algorithm for simultaneously inferring estimates of the sources, $\boldsymbol{s}_t$, and of the underlying structure, $\boldsymbol{\lambda}_t$, from a stream of observations, $\boldsymbol{v}_t$ (assuming that time-constants and observation noise are fixed parameters). Note that, with the component matrix, $\boldsymbol{C}$, given, the motion structure is fully defined by $\boldsymbol{\lambda}$. The derivation exploits the different timescales of typical changes in $\boldsymbol{s}_t$ (volatile, E-step) and $\boldsymbol{\lambda}$ (stable, M-step). With mild approximations, we obtained:

$$\partial_t \boldsymbol{\lambda}_t^2 = -\tfrac{1}{\tau_\lambda}\,\boldsymbol{\lambda}_t^2 + \boldsymbol{\alpha} \odot \left(\boldsymbol{\mu}_t^2 + \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right) + \boldsymbol{\beta} \quad \text{with} \quad f_\Sigma(\lambda_m^2) = \frac{\sigma_{\text{obs}}^2}{\tau_s \,\|\boldsymbol{c}_m\|^2}\left(-1 + \sqrt{1 + \frac{{\tau_s}^2\,\|\boldsymbol{c}_m\|^2}{\sigma_{\text{obs}}^2}\,\lambda_m^2}\right), \tag{1}$$

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

Here, $\boldsymbol{\mu}$ and $\boldsymbol{f}_\Sigma$ are the posterior parameters of $p(\,\boldsymbol{s}_t \,|\, \boldsymbol{v}_{0:t}\,) = \mathcal{N}\left(\boldsymbol{\mu}_t,\, \mathrm{diag}\left[\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)\right]\right)$, $\odot$ denotes elementwise multiplication, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ stem from a sparsity-inducing prior on $\boldsymbol{\lambda}^2$, $\|\boldsymbol{c}_m\|^2$ is the norm of $\boldsymbol{C}$'s mth column, and we require that $\tau_\lambda > \tau_s$. As common for online EM, eqn. (1) + (2) define a coupled dynamical system. Intuitively, the algorithm measures the range in which the motion sources vary, $\langle \boldsymbol{s}_t^2 \rangle$, to estimate the structure parameters, $\boldsymbol{\lambda}_t^2$, during the M-step in eqn. (1). At the same time, during the E-step in eqn. (2), the system's expected input, $\boldsymbol{C}\boldsymbol{\mu}_t$, leads to prediction errors, $\boldsymbol{\epsilon}_t$, which are projected into the domain of sources, $\boldsymbol{C}^\mathsf{T}\,\boldsymbol{\epsilon}_t$, and gated by $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$ for credit assignment. The E-step performs the velocity decomposition *conditioned on* the scene's structure which is inferred during the M-step. This is, to our knowledge, the first online inference model of Bayesian motion structure perception.



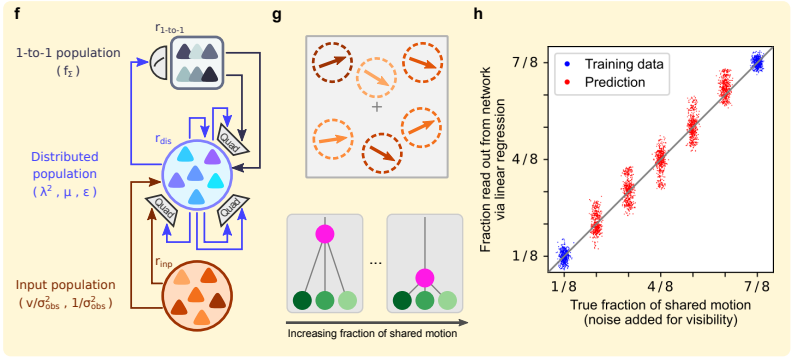


Figure 2: The model captures human motion direction repulsion and supports a neural implementation. *(**a–e**) Simulations of motion direction repulsion. **(a)** Motion direction repulsion is a systematic bias in the perceived angle between the motion of two groups of dots moving linearly in an aperture. **(b)** We endowed the model with self-, shared and group motion (yellow, pink, and green), as well as a noisy vestibular input signaling the observer's stationarity. **(c)** The algorithm replicates the biphasic opening angle-dependence of the bias measured by [10]. **(d & e)** We further make testable predictions for varying contrast (d) and speed (e) of the 2nd dot group. **(f)** Recurrent network model implementing the online algorithm. **(g)** Proposed experiment to measure neural representations of latent structure. Moving dots in several apertures follow our generative model (top). Different trials use different fractions of shared and individual motion (bottom). **(h)** The model predicts that the fraction of shared motion in the stimulus can be read out by a linear regression model (red points) which was only trained on a subset of trials (blue points).*

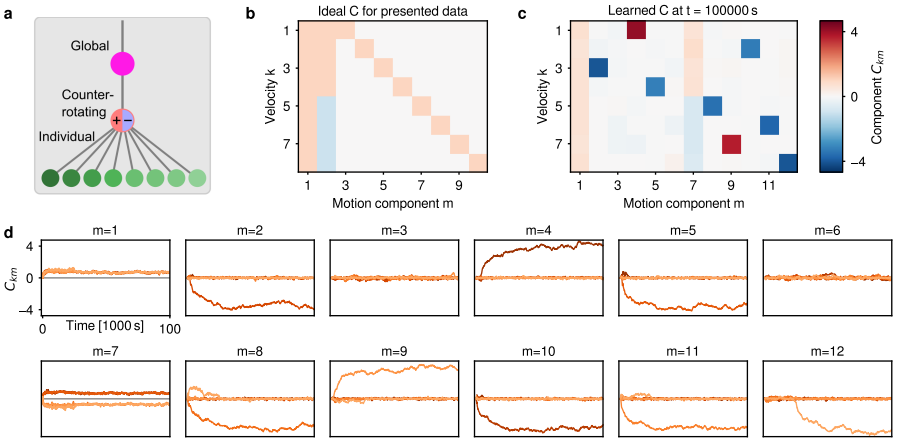


Figure 3: Learning of motion components from experience. ***(a)*** *Full structure of presented velocities. From the available features (global motion, counter-rotation, 8 individual motions) only a random subset is presented at any time.* ***(b)*** *Theoretically optimal $\boldsymbol{C}$-matrix for this stimulus.* ***(c)*** *Learned $\boldsymbol{C}$-matrix. All components have qualitatively been identified (note that the sign and m-order play no role). We gave the algorithm two more components in $\boldsymbol{C}$ than required. These extra components have, as expected, been left unused by the learning algorithm.* ***(d)*** *Time evolution of all 12 motion components during learning (one panel per component m; one line per velocity k).*

The algorithm explains experiments of human motion perception

Due to limited space, we here only present results on the Duncker wheel (**Fig. 1f – h**), motion structure classification of ambiguous scenes (**Fig. 1i – l**; data from [9]), and biased perception known as motion direction repulsion (**Fig. 2a – e**; data from [10]). How the algorithm replicates and explains these experiments is detailed in the figure captions.

Neural network model and proposed neuroscience experiment

Eqs. (1) and (2) rely on only linear and quadratic operations (by adding $\boldsymbol{\epsilon}_t$ as a represented auxiliary variable). Following a derivation similar to [11], who showed how up-to-quadratic dynamics of computational variables can be implemented by recurrent rate-based networks, we devised a network model with biologically realistic neural interactions to implement the inference algorithm. The network architecture is shown in **Fig. 2f**. The network represents (inferred) motion strengths, $\boldsymbol{\lambda}_t^2$, and motion source means, $\boldsymbol{\mu}_t$, in a linear population code. The only variable not fitting into this framework is the square-root function in f_Σ which is thus represented by a dedicated neural population. The full derivation of the model and a demonstration of its ability to implement the inference algorithm are provided in the bioRxiv preprint.

Motivated from the network model, we propose a new class of theory-driven neuroscience experiments to probe the neural representations of latent structure. In each trial, moving dots, which follow the generative model from **Fig. 1d&e**, are presented in several apertures, see **Fig. 2g (top)**. Different trials use different fractions of shared and individual motion, such that the expected dot speed, $\langle v^2 \rangle \propto \lambda^2_{\text{shared}} + \lambda^2_{\text{ind}}$, is held constant across trials, see **Fig. 2g (bottom)**. The network model encodes $\boldsymbol{\lambda}_t^2$ linearly in its neural activity, and thus predicts that the fraction of shared motion, defined as $\lambda^2_{\text{shared}} \,/\, (\lambda^2_{\text{shared}} + \lambda^2_{\text{ind}})$, can be read out by a linear regression model. As shown in **Fig. 3h**, a linear regression model trained on two fractions (blue points) correctly reads out also other fractions (red points).

Learning of common motion components on long timescales

Could the motion components, $\boldsymbol{C}$, be learned from observations in an unsupervised manner? We derived a learning rule for the matrix elements, C_{km}, through gradient-based online EM. Since learning of the motion components, $\boldsymbol{C}$, aims to maximize the data likelihood beyond what could be explained by the sources, $\boldsymbol{s}_t$, and the scenes' structures, $\boldsymbol{\lambda}_t$, learning of $\boldsymbol{C}$ must evolve on long timescales. We think of it as lifetime learning of a set of features which commonly occur in natural scenes. To keep the emerging components sparse and interpretable, we further have the freedom to impose a regularizing prior, $p(\boldsymbol{C})$, during learning. The full learning rule reads:

$$\partial_t \boldsymbol{C} = \eta_C \left[\frac{1}{\sigma^2_{\text{obs}}} \left(\boldsymbol{v}_t \, \boldsymbol{\mu}_t^\mathsf{T} - \boldsymbol{C} \, \left(\boldsymbol{\mu}_t \, \boldsymbol{\mu}_t^\mathsf{T} + \boldsymbol{\Sigma}_t \right) \right) + \frac{1}{N_C} \nabla_{\boldsymbol{C}} \log p(\boldsymbol{C}) \right] \quad . \tag{3}$$

This rule establishes the intuition to compare the observed covariance between inputs and inferred motion sources against their expected covariance. Here, η_C is a small learning rate, N_C weights the prior vs. the likelihood, and $\boldsymbol{\Sigma}_t = \text{diag}\left[\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2) \right]$ is the variance of the $\boldsymbol{s}$-posterior.

In **Fig. 3**, we demonstrate, for the first time, online learning of hierarchically nested motion relations in a computer simulation. Observed velocities are generated according to the model in **Fig. 1d & e**, with the nested graph structure shown in **Fig. 3a**. The corresponding ground truth-component matrix is shown in **Fig. 3b**. At any time, a random subset of the available components from **Fig. 3b** is present (average: 3 active components), and a new random structure is drawn every 20 s. For the learning system, the component matrix is initially empty, i.e., $\boldsymbol{C}(t=0) = \boldsymbol{0}$, and we provide two

additional (empty) components to this matrix to test whether the algorithm finds a sparse, minimal solution. The system evolves according to eqn. (1)–(3) for $100{,}000\,\mathrm{s}$ using the following regularizing prior:

$$\log p(\boldsymbol{C}) \propto -\frac{1}{2l_2}\sum_{m=1}^{M}\left(\sum_{k=1}^{K}|C_{km}|\right)^2 - \frac{1}{l_1}\sum_{m=1}^{M}\sum_{k=1}^{K}|C_{km}| \quad . \tag{4}$$

This prior facilitates motion components to remain small (L2 regularization of full components), and individual matrix elements to be sparse (L1 regularization of C_{km}). Furthermore, we add small, zero-mean exploration noise in every time step during learning.

At the end of the simulation, all components were correctly identified, see **Fig. 3c**. The time evolution during learning is shown in **Fig. 3d** for all $M=12$ components. The orange lines denote the $K=8$ velocities ($k=1$: darkest). A horizontal gray line marks $C_{km}=0$ (if visible). Global motion ($m=1$) and counter-rotation ($m=7$) are discovered quickly. The individual components take more time.

Interaction of priors across timescales

All dynamic variables, that is, $\boldsymbol{s}$, $\boldsymbol{\lambda}$, and $\boldsymbol{C}$, are softly constrained by prior distributions. We briefly discuss the modeling assumptions arising from those priors as well as how the priors interact with another. On the fasted timescale of the generative model, the motion sources, $\boldsymbol{s}_t$, obey a Gaussian prior, $p(s_m) = \mathcal{N}\left(0,\, \frac{\tau_s}{2}\lambda_m^2\right)$, which favors slow velocities and is supported for modeling visual motion perception by experimental work [3]. On an intermediate timescale, motion strengths, $\boldsymbol{\lambda}$, are governed by priors from the family of scaled inverse chi-squared distributions leading to the constants $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in eqn. (1). For this work, we employed two members of this family: for all allocentric motion of objects, that is, every λ_m except self-motion, we chose the scale-free Jeffreys prior, $p(\lambda_m^2) \propto \lambda_m^{-2}$. This prior induces sparsity of inferred motion structures by preferring vanishing motion components ($\lambda_m^2 = 0$) and interacts with $p(s_m)$ by setting its variance. For self-motion, we chose a uniform distribution, $p(\lambda_{\text{self}}^2) = const.$, capturing that frequent saccades and other eye movements give rise to fast retinal velocities for all observables, $\boldsymbol{v}_t$. On the longest timescale, motion components, $\boldsymbol{C}$, follow the regularizing prior of eqn. (4) which has been discussed earlier. Notably, in interaction with $p(\boldsymbol{\lambda}^2)$, this prior resolves an invariance in the generative model: multiplying $\boldsymbol{C}$ with any constant can be fully compensated by dividing $\boldsymbol{\lambda}$ by the same constant, thereby leaving $p(\,\boldsymbol{v}\,|\,\boldsymbol{\lambda},\,\boldsymbol{C}\,)$ unchanged. It is the exponential penalty on large components in $p(\boldsymbol{C})$ that counteracts the preference of $\boldsymbol{\lambda}$ for small values, thereby leading to the (now unique) equilibria observed in **Fig. 3d**.

Conclusion

We have proposed a comprehensive theory of online hierarchical inference for structured visual motion perception. The derived algorithm 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 and made concrete predictions for new experiments. The algorithm afforded a recurrent neural network model motivating targeted neuroscience experiments to reveal the neural representations of latent structure. Finally, we demonstrated in a computer simulation that also the set of motion components could be learned online from experience.

References

[1] Charles Kemp and Joshua B Tenenbaum. "The discovery of structural form". In: *Proceedings of the National Academy of Sciences* (2008).

[2] Andrew M Saxe, James L McClelland, and Surya Ganguli. "A mathematical theory of semantic development in deep neural networks". In: *Proceedings of the National Academy of Sciences* (2019).

[3] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. "Motion illusions as optimal percepts". In: *Nature neuroscience* (2002).

[4] Alan A. Stocker and Eero P. Simoncelli. "Noise characteristics and prior expectations in human visual speed perception". In: *Nature Neuroscience* (Apr. 2006).

[5] Andrew E. Welchman, Judith M. Lam, and Heinrich H. Bülthoff. "Bayesian motion estimation accounts for a surprising bias in 3D vision". In: *Proceedings of the National Academy of Sciences* (2008).

[6] James H Hedges, Alan A Stocker, and Eero P Simoncelli. "Optimal inference explains the perceptual coherence of visual motion stimuli". In: *Journal of vision* (2011).

[7] Samuel J Gershman, Joshua B Tenenbaum, and Frank Jäkel. "Discovering hierarchical motion structure". In: *Vision research* (2016).

[8] Johannes Bill et al. "Hierarchical structure is employed by humans during visual motion perception". In: *Proceedings of the National Academy of Sciences* (2020).

[9] Sichao Yang et al. "Human visual motion perception shows hallmarks of Bayesian structural inference". In: *Scientific reports* (2021).

[10] Oliver J Braddick, Keith A Wishart, and William Curran. "Directional performance in motion transparency". In: *Vision research* (2002).

[11] Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. "Marginalization in neural circuits with divisive normalization". In: *Journal of Neuroscience* (2011).