Psychological Review

Likelihood-Based Parameter Estimation and Comparison of Dynamical Cognitive Models

Heiko H. Schütt, Lars O. M. Rothkegel, Hans A. Trukenbrod, Sebastian Reich, Felix A. Wichmann, and Ralf Engbert

Online First Publication, April 27, 2017. http://dx.doi.org/10.1037/rev0000068

Psychological Review

© 2017 American Psychological Association
0033-295X/17/$12.00 http://dx.doi.org/10.1037/rev0000068

THEORETICAL NOTE

Likelihood-Based Parameter Estimation and Comparison of Dynamical Cognitive Models

Heiko H. Schütt
University of Tübingen and University of Potsdam

Lars O. M. Rothkegel, Hans A. Trukenbrod, and Sebastian Reich
University of Potsdam

Felix A. Wichmann
University of Tübingen

Ralf Engbert
University of Potsdam

Dynamical models of cognition play an increasingly important role in driving theoretical and experimental research in psychology. Therefore, parameter estimation, model analysis and comparison of dynamical models are of essential importance. In this article, we propose a maximum likelihood approach for model analysis in a fully dynamical framework that includes time-ordered experimental data. Our methods can be applied to dynamical models for the prediction of discrete behavior (e.g., movement onsets); in particular, we use a dynamical model of saccade generation in scene viewing as a case study for our approach. For this model, the likelihood function can be computed directly by numerical simulation, which enables more efficient parameter estimation including Bayesian inference to obtain reliable estimates and corresponding credible intervals. Using hierarchical models inference is even possible for individual observers. Furthermore, our likelihood approach can be used to compare different models. In our example, the dynamical framework is shown to outperform nondynamical statistical models. Additionally, the likelihood based evaluation differentiates model variants, which produced indistinguishable predictions on hitherto used statistics. Our results indicate that the likelihood approach is a promising framework for dynamical cognitive models.

Keywords: likelihood, model fitting, dynamical model, eye movements, model comparison

The broad class of dynamical cognitive models (Van Gelder, 1998) provides a powerful framework for explaining behavioral data. This modeling approach has been particularly successful in sensorimotor control. For example, an early paradigmatic model was proposed by Haken, Kelso, and Bunz (1985) who introduced coupled nonlinear oscillators as a mathematical model for phase transitions in human finger movements. Another general theory was proposed by Erlhagen and Schöner (2002), who introduced a flexible framework of movement preparation based on dynamical equations for the temporal evolution of neural fields that specify motor actions in space and time. With their decision field theory, Busemeyer and Townsend (1993) developed a dynamical framework for decision making in uncertain environments. These representative examples indicate the broad range of dynamical models in cognitive science.

A strength of the dynamical approach is to generate specific predictions, including the dependencies between different datapoints over time. This, however, implies that the statistical treatment of dynamical models requires the comparison of model predictions for time-ordered and interdependent data, which complicates parameter identification and model comparison. As a result, dynamical models are often handled with heuristic and approximate methods. In this article, we discuss

Heiko H. Schütt, Neural Information Processing Group, University of Tübingen, and Department of Experimental and Biological Psychology, University of Potsdam; Lars O. M. Rothkegel and Hans A. Trukenbrod, Department of Experimental and Biological Psychology, University of Potsdam; Sebastian Reich, Institute for Mathematics, University of Potsdam; Felix A. Wichmann, Neural Information Processing Group, University of Tübingen; Ralf Engbert, Department of Experimental and Biological Psychology, University of Potsdam.

We thank Simon Barthelmé, Grenoble, for valuable discussions. This work was supported by grants from Deutsche Forschungsgemeinschaft to Ralf Engbert (Grant EN 471/13–1) and to Felix A. Wichmann (Grant WI 2103/4–1).

Correspondence concerning this article should be addressed to Heiko H. Schütt, Neural Information Processing Group, University of Tübingen, Sand 6, 72076 Tübingen, Germany. E-mail: heiko.schuett@uni-tuebingen.de

an alternative to these heuristic approaches, namely a statistically well-founded analysis based on the likelihood framework.

An important application of the dynamical framework is the modeling of eye movements. Human observers move their eyes three to four times per second to shift gaze to regions of interest within a given visual scene (Henderson, 2003; Yarbus, 1967;). Eye movements are important because high-acuity vision is limited to the fovea, a small region with a spatial extension of about 2 degrees of visual angle (Nicholls et al., 2012; von Helmholtz, 1924). The analysis of fixated regions permits conclusions on the type of features that attract our gaze. For eye movements in natural scenes, saliency models concentrate on predicting the fixation density for large data sets (Itti & Koch, 2001). The density of fixations provides only information where people look regardless of serial order and durations of fixations. This research strategy turned out to be very successful and a range of saliency models was developed to predict fixation density for a given input image (Borji & Itti, 2013; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Kümmerer, Wallis, & Bethge, 2015).

Recently, there has been an increasing interest in cognitive models that produce sequences of fixations, that is, a scanpath, on a natural scene (Borji, Sihite, & Itti, 2014; Engbert, Trukenbrod, Barthelmé, & Wichmann, 2015; Le Meur & Liu, 2015; Zelinsky, 2008). Related models aim at a more complete explanation of the cognitive principles underlying the control of attention and eye movements during exploration of natural scenes. Statistical measures include simple statistics like the distribution of saccade lengths and angles between subsequent saccades (Klein & MacInnes, 1999; Smith & Henderson, 2009), but also more complex spatial statistics that relate image properties to fixation density (Barthelmé, Trukenbrod, Engbert, & Wichmann, 2013) or to spatial correlation functions (Engbert et al., 2015).

In the traditional approach for the evaluation of scanpath models, researchers typically simulate scanpaths from their models and compare simulated data to experimentally observed scanpaths using a broad range of statistics (Le Meur & Baccino, 2013). The most common statistics are those associated with the observed experimental data (e.g., distributions of saccade angle and saccade amplitudes). Alternative methods are based on comparisons of scanpaths that include string comparison methods based on the Levenshtein distance (Levenshtein, 1966; von der Malsburg & Vasishth, 2011) or vector-based methods (Jarodzka, Holmqvist, & Nyström, 2010). However, each effect and each discriminating statistic for scanpaths evaluates different aspects of the models. Thus, ranking of model performance depends critically on which effects are investigated and which statistics are applied. None of the statistics used so far quantify the general agreement between models and experimental data in a dynamical framework.

For saccade generation in dynamical cognitive models, a spatiotemporal map of activations (Erlhagen & Schöner, 2002) is built-up according to dynamical evolution equations (e.g., Jackson, 1992). When a saccade target is needed, the activation map is read out to generate a target with a probability that equals the relative activation as determined by the map at the time of saccadic selection. We study a dynamical model of scanpath generation for eye movements in scene viewing (Engbert et al., 2015). Although we focus on this concrete example to illustrate the procedures of model parameter identification and model comparison, the model only serves as a representative example for the broad class of dynamical cognitive models that are developed for the prediction of sequences of discrete motor actions.

In the current study, we investigate the application of the likelihood function as a statistical measure of model performance. The likelihood function of a model M is the probability that a given set of experimental data was generated by the model and a corresponding set of model parameters $\boldsymbol{\theta}$. Therefore, the likelihood function for a given model depends on the data set and the set of model parameter values that specify the model's behavior. The likelihood is the most widely used measure of model performance in mathematical statistics (Bickel & Doksum, 1977; Cox, 2006). However, because its numerical computation is believed to be difficult, the likelihood is not yet part of the standard toolbox for dynamical models of cognition. Solving likelihood computation for dynamical models of cognition is potentially very important because likelihood is the starting point for many additional concepts of statistical inference about model parameters and comparisons between different models, including Bayesian inference (Gelman, Carlin, Stern, & Rubin, 2014).

The likelihood can be computed whenever the model can generate the observed data with a certain probability that is nonzero. This is already guaranteed, if the probability for the next datum can be calculated given the previous data and is greater than zero for any observed datum. This means that the likelihood approach can be applied to an extremely broad class of models.

To investigate how the analysis of dynamical models can benefit from the likelihood approach, we demonstrate numerical computations for the recently published SceneWalk model of scanpath generation in natural scene viewing (Engbert et al., 2015). The general motivation for modeling human scanpaths is to derive the rules for the sequential deployment of overt attention (i.e., gaze position) in a natural scene-viewing task. The SceneWalk model starts from a given spatial distribution of fixation positions (an empirical saliency map). Thus, we assume to have perfect knowledge about saliency (up to differences between observers). This is not a strong limitation because the model could easily be combined with one of the successful saliency models (see Borji & Itti, 2013, for an overview). Thus, our modeling goal is to reproduce the key statistics of human scanpaths (e.g., distribution of saccade lengths and spatial correlations) for a given image, when the time-independent, two-dimensional distribution of fixation positions is known to a good approximation.

Likelihood Computation for Dynamical Models

Definition of Likelihood Function

The fundamental theoretical concept for our approach is the likelihood $L_M(\boldsymbol{\theta}\,|\,\text{data})$ of a model M with parameters $\boldsymbol{\theta}$ given a specific set of experimental data, which is defined as the conditional probability density f_M for observing the data in the context of model M specified by parameters $\boldsymbol{\theta}$, such that

$$L_M(\boldsymbol{\theta}\,|\,\text{data}) = f_M(\text{data}\,|\,\boldsymbol{\theta}) \approx \frac{P_M(\text{data}\,|\,\boldsymbol{\theta})}{(\Delta A)^N}. \qquad (1)$$

In our case, data are given by a sequence of fixations, for which our models shall predict a density one after another. Each of these densities can be approximated by the probabilities to observe the fixations exactly on a discrete grid, divided by the area each gridpoint represents resulting in a denominator of $(\Delta A)^N$ for N fixations. We stay with this grid approximation to all likelihoods in this article, as many models are themselves defined on grids, including saliency models and the SceneWalk model that we investigate in the current study. The grid approximation simplifies numerical computations, because this probability is always defined and all integrals reduce to summations over grid points.

Furthermore we set $\Delta A = 1$, measuring area in grid points, which works, because all models that we aim to compare to each other make predictions on the same grid of possible fixation locations. Measuring the area in grid independent units (cm, pixels, degrees of visual angle, etc.) in principle enables comparisons between models, which are defined on different grids. However, the use of a coarser grid implicitly blurs model predictions for eye movement models and a blurring of the final predictions may change performance considerably (Judd, Ehinger, Durand, & Torralba, 2009). Thus, we think it is preferable to convert all model predictions to the same grid making all necessary conversions explicit.

The likelihood quantifies how well a model describes the data and is the most common criterion for model evaluation in mathematical statistics. Therefore, maximizing the likelihood of a given dataset by optimizing model parameters is a straightforward approach to model fitting.[1] Applicability of the likelihood approach depends on both the structure and complexity of a model M, that is, whether the likelihood can be computed exactly (analytically or via numerical simulation of the model) or whether we need to introduce further approximations. If it is not practical to compute the likelihood, likelihood-free strategies for parameter estimation and model comparison have been proposed as an alternative (see Discussion).

The Likelihood for Dynamical Models Based on Discrete Observations

To calculate the likelihood for dynamical models based on time-ordered experimental data and, specifically, for the SceneWalk model of eye movements in scene viewing (Engbert et al., 2015), we split the likelihood into a product of probabilities for all fixations $f_i = (x_{f_i}, y_{f_i})$ given the previous fixations $f_1 \ldots f_{i-1}$ in the sequence, that is

$$L_M(\boldsymbol{\theta} \,|\, \text{data}) = L_M(\boldsymbol{\theta} \,|\, f_1, f_2, \ldots, f_n)$$

$$= P_M(f_1) \prod_{i=2}^{n} P_M(f_i \,|\, f_1, \ldots, f_{i-1}, \boldsymbol{\theta}), \qquad (2)$$

where $P_M(f_1)$ is the probability of the initial fixation starting at time $t = 0$, which can be given by the experimental design or the model. The conditional probabilities $P_M(f_i | f_1 \ldots f_{i-1}, \boldsymbol{\theta})$ can be computed by enforcing the model to generate the sequence of fixations $f_1, \ldots f_{i-1}$ to obtain the probability for the ith fixation f_i. This is possible in dynamical models which generate a continuous-time activation map u that translates into a fixation probability π to place the next fixation at position f_i at time t. During numerical simulation, we force the model to generate a particular scanpath prescribed by the data $f_1, f_2, \ldots,$ which translates into a certain probability at each iteration and reduces the necessary computations to a single model run for a given scanpath. This procedure is illustrated for the first fixations on an image in Figure 1.

For practical purposes, it is advantageous to use the logarithm of the likelihood (log-likelihood):

$$l_M(\boldsymbol{\theta} \,|\, \text{data}) = \log(L_M(\boldsymbol{\theta} \,|\, \text{data})) \qquad (3)$$

$$= \sum_{i=1}^{N} \log(P_M(f_i | f_1 \ldots f_{i-1}, \boldsymbol{\theta})) \qquad (4)$$

The log-likelihood can be calculated and optimized more easily because it transforms the products over observations into sums of terms and scales numerical values to a more feasible range.

The log-likelihood characterizes model performance on the whole dataset, in the current case the fixation sequence or scanpath. Therefore, the log-likelihood of a scanpath given a model depends on the length of the sequence or number of fixations. To obtain a number that is easier to compare between different realizations of scanpaths, it is more informative to compute the log-likelihood per fixation, which turns out to represent a sensitive measure of model performance as the log-likelihood is added up over all fixations in a given sequence.

Thus, effectively, we compute the average probability of an observed fixation, calculating the average as a geometric mean. However, we express all likelihoods on a logarithmic scale. When the $\log_2$ is used as we do in this article, the unit of the log-likelihoods is a *bit*. A difference of 1 bit between two log-likelihood values thus indicates that the corresponding likelihoods differ by a factor of 2.

A log-likelihood of zero indicates that the model predicted the observed data exactly and with probability one. This is a limiting case and certainly not a realistic scenario for typical cognitive models. Almost always models predict a distribution over multiple possible outcomes, which each have smaller probabilities than one. Therefore, log-likelihoods are almost always negative. Indeed the log-likelihoods we calculate subsequently will usually be in the range between $-10\frac{\text{bit}}{\text{fix}}$ and $-20\frac{\text{bit}}{\text{fix}}$.[2]

Model Details

For the analysis of the likelihood of the SceneWalk model, we need to compute the probability for the next fixation, given all previous fixations in a given trial. In this section, we describe how the SceneWalk model computes probability distributions. To explain this, we provide a short recap of the model internals and describe the details of some variants of the model used to exemplify the following model comparisons.

The SceneWalk model is based on two independent processing streams for excitatory and inhibitory aspects of saccade planning that are related to attentional deployment (Itti & Koch,

[1] We consider only finite dimensional parameters and models in this article. We know of no nonparametric models for scanpath generation. A nonparametric model increases the complexity of the analysis considerably. If the reader is interested in this, there is a broad literature on nonparametric statistics in both frequentist (Conover & Conover, 1980) and Bayesian statistics (Gershman & Blei, 2012).

[2] Note that these reference values are specific for our choice of grid and area unit, such that they cannot be compared with values obtained with a different grid or area unit. Especially, densities and thus likelihoods can be larger than 1 and log-likelihoods larger than 0, depending on the measure of area chosen.

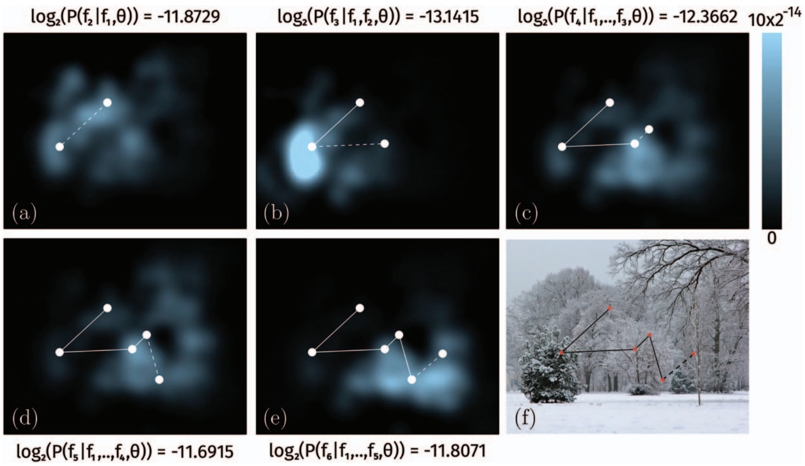

Figure 1. Numerical calculation of the likelihood for an example of a fixation sequence. Panels (a) through (e): Visualization of the probabilities of the first 5 fixations from a sequence as predicted from the SceneWalk model. We compute the probability $P(f_{\text{i}}|f_1 \ldots f_{\text{i}-1}, \boldsymbol{\theta})$ of the next fixation, which the human observer actually generated and force the model to choose the fixation location accordingly. With this new location we can calculate the probability distribution for the next saccade and can thus iterate through the observed scanpaths and calculate their probabilities given by the model and its parameter values. Panel (f): The presented image with the scanpath overlayed. See the online article for the color version of this figure.

2001; Itti, Koch, & Niebur, 1998) and inhibition-of-return (Klein, 2000; Klein & MacInnes, 1999;), respectively (see Figure 2). The excitatory pathway starts with a given fixation density (empirical saliency), which is multiplied with a Gaussian attention window around the current fixation location resulting in a local saliency map. This localization step serves as a first-order approximation to the peripheral loss in available information, cortical processing, and visual attention. For the inhibitory pathway we start with a simple Gaussian around the current fixation marking the currently visited area. The local saliency and the inhibitory Gaussian are both implicitly time-dependent through changes of gaze position.

For a current fixation position $\mathbf{x}_f = (x_f, y_f)$, we first compute the two Gaussian distributions centered at $\mathbf{x}_f$ on a grid of size $L \times L$. The attentional pathway uses a Gaussian aperture G_A with standard deviation σ_A to access the static empirical saliency map. The pathway for inhibitory tagging uses a Gaussian G_F with standard deviation σ_F to build-up inhibition that drives the model to new regions of the visual field. For a grid position (x, y), these Gaussians are given by

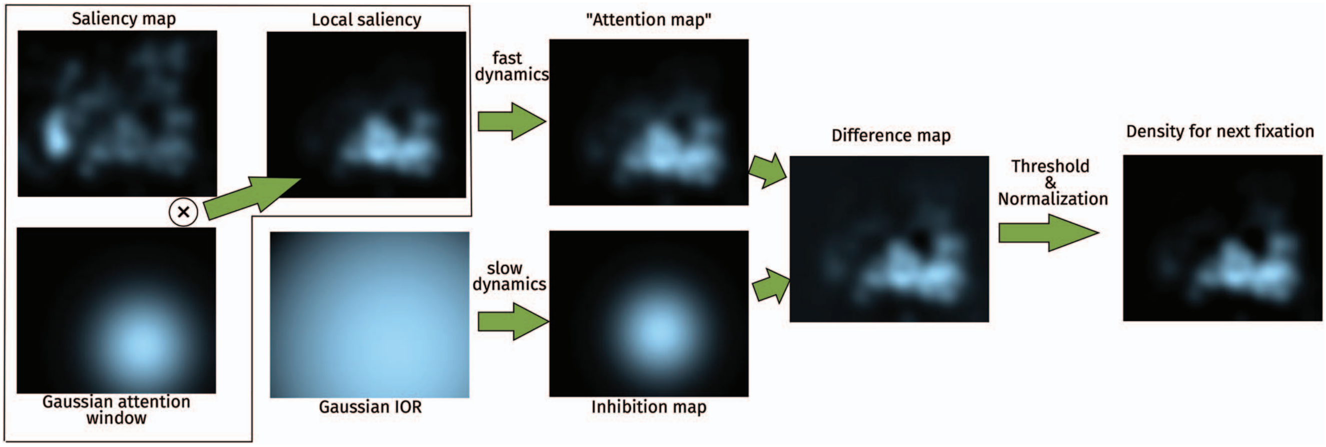

Figure 2. Schematic illustration of the SceneWalk model (Engbert et al., 2015). The temporal evolution of two independent processing streams for attention and inhibition-of-return is combined into the time-dependent potential $u(x, t)$ that determines the next saccade target. The saliency map is weighted by a Gaussian (attentional window) placed at the current fixation. The resulting local saliency map is used as the input for the build-up of activation in the attention map. An inhibition map is subtracted, which builds up more slowly using a constant-shape Gaussian around the current fixation as input. Finally, thresholding and normalization yield the final distribution $u(x, t)$ for the probabilistic selection of the next saccade target. See the online article for the color version of this figure.

$$G_{A/F}(x,y;x_f,y_f) = \frac{1}{2\pi\sigma^2_{A/F}}\exp\left(-\frac{(x-x_f)^2+(y-y_f)^2}{2\sigma^2_{A/F}}\right). \qquad (5)$$

Next, we define the change over time of the attention map $A(t) = \{A_{ij}(t)\}$ and the fixation map $F(t) = \{F_{ij}(t)\}$ with indices $1 \leq \{i,j\} \leq L$ running over the whole image. Two parameters ω_A and ω_f scale the rates of activation change in the two maps and we require the given time-independent salience map $S = \{S_{ij}\}$ and the Gaussians G_A and G_F from Equation 5:

$$\frac{\mathrm{d}A_{ij}(t)}{\mathrm{d}t} = -\omega_A A_{ij}(t) + \omega_A \frac{S_{ij}\cdot G_A(x_i,y_j;x_f,y_f)}{\sum_{kl} S_{kl}\cdot G_A(x_k,y_l;x_f,y_f)} \qquad (6)$$

$$\frac{\mathrm{d}F_{ij}(t)}{\mathrm{d}t} = -\omega_F F_{ij}(t) + \omega_F \frac{G_F(x_i,y_j;x_f,y_f)}{\sum_{kl} G_F(x_k,y_l;x_f,y_f)}, \qquad (7)$$

where the $\sum_{kl}$ symbol denotes the sum over all grid-points (k, l).

These evolution equations were formulated as difference equations in Engbert et al. (2015). However, we moved to differential equations here, as they can be solved analytically. By solving Equations 6 and 7, we can exploit the fact that the input $G_{A/F}$ changes only as a result of saccadic gaze shifts $\mathrm{x}_f \mapsto \mathrm{x}'_f$. The solution of the differential equations for initial maps A_0 and F_0 at the start of the fixation at time t_0 are given as

$$A(t) = \frac{G_A S}{\sum G_A S} + e^{-\omega_A(t-t_0)}\left(A_0 - \frac{G_A S}{\sum G_A S}\right) \qquad (8)$$

and

$$F(t) = \frac{G_F}{\sum G_F} + e^{-\omega_F(t-t_0)}\left(F_0 - \frac{G_F}{\sum G_F}\right), \qquad (9)$$

where indices have been dropped to simplify the presentation. As a consequence of the linear dynamics of the maps, the solutions describe exponential change from the map represented at the beginning of the fixation toward the input map. Using these equations, we can calculate the activities at the end of the fixation directly. Another advantage is that this formulation prevents temporal discretization errors (in the original model, a 10-ms temporal discretization was used, see Engbert et al., 2015, for details).

At the first fixation the maps in the model need to be initialized. The original model was initialized with zero activities of the maps for attention and inhibitory tagging. For short durations of the first fixation, however, this led to unintended behavior, as the maps are normalized. Small activations on the maps are amplified by the normalization which introduces unwanted starting effects. To prevent this problem of the model's initial conditions, we prepared the maps with a uniform distribution of sum one and adjusted the magnitude of the input such that the equilibrium size of the maps was normalized to one as well. Thus, the sum of activation of the attention map and of the map for inhibitory tagging remains at a constant value of one throughout each simulated trial.

Finally, the two independent activation maps $A(x, t)$ and $F(x, t)$ are combined into a map $u(x, t)$, which is defined as the difference of the attention and inhibition maps after thresholding and normalization. To obtain a flexible relative weighting within each map, numerical values of activations are raised to power λ for the attention map A and to power γ for the fixation map F, respectively. Next, each map is normalized to unit sum. Finally, the map for inhibitory tagging is multiplied by a factor c_F and subtracted from the attention map. As a result, we obtain a time-dependent potential $u_{ij}(t)$ for target selection:

$$u_{ij}(t) = \frac{[A_{ij}(t)]^\lambda}{\sum_{kl}[A_{kl}(t)]^\lambda} - c_F\frac{[F_{ij}(t)]^\gamma}{\sum_{kl}[F_{kl}(t)]^\gamma}. \qquad (10)$$

Note that we introduced the factor c_F as an additional parameter, which was not present in the original model (Engbert et al., 2015).

Taking a power of the map at each point changes not only the weighting between different peaks, but also shrinks or widens the individual peaks. Therefore, to obtain parameters which represent the size of the final influence and are thus easier to interpret, we reparametrized the model using the following equations:

$$\lambda\sigma_A^{'2} = \sigma_A^2 \qquad \gamma\sigma_F^{'2} = \sigma_F^2 \qquad (11)$$

Thus σ'_A and σ'_F are the standard deviations the Gaussians would have if they were mapped through the nonlinearity directly.

Normalization. To obtain a probability distribution from $u_{ij}(t)$, the potential is normalized to be positive and to have a unit integral over the whole image. In the normalization procedure of the original model, negative values of the potential $u_{ij}(t)$ implied probability zero to select position (i, j) as the next saccade target. However, this is an unrealistic assumption in the model because experimental data do not indicate regions which are never selected as a saccade target. We changed the model accordingly. First, we define a function which continuously maps u to an intermediate u^*, which is positive everywhere, that is

$$u^*(u) = \begin{cases} u & u > 0 \\ 0 & u \leq 0 \end{cases} \qquad (12)$$

In a second step we compute a mixture with a uniform distribution using a weighting factor ζ to obtain the probability $\pi(i, j)$ for each position on the lattice to be selected as the next fixation target,

$$\pi(i,j) = (1-\zeta)\frac{u^*_{ij}}{\sum_{kl} u^*_{kl}} + \zeta\frac{1}{\sum_{kl} 1}. \qquad (13)$$

This formulation maps the original function u to a probability on the map, which always returns a positive probability $(\geq \zeta\,/\,(\sum_{kl}1))$ for any next fixation. Furthermore, areas with high u ($u >> \zeta$) are not further distorted by this mapping, such that relative weightings from the original empirical saliency map are kept.

The distribution $\pi(i, j)$ directly represents the probability of a specific grid-point to be the next fixation target, given the previous fixations, that is, the map to be used in the likelihood calculation described in Equation 2 and illustrated in Figure 1 completing our description of the likelihood calculation for the SceneWalk model.

Competing Models

Nondynamic benchmarks. First, we compare the performance of our model to nondynamical models that represent limiting cases for saliency evaluation: An image independent spatial bias and empirical saliency. The image independent spatial bias mostly represents the central fixation bias (Buswell, 1935; Tatler, 2007)—the experimental observation that observers initially direct their gaze positions toward the image center. A corresponding model can be realized as an image-independent kernel density estimate of all fixations of the full set of images. The empirical saliency model represents the optimal prediction of fixation positions from other observers generated as a kernel density estimate as well, using fixations on the tested image only. Additionally, we implemented a model which generates a uniform distribution over the full image as a null model setting a zero point on our log-likelihood scale.

A model without inhibition. As a first dynamical model to compare with, we chose a model without inhibition, to test whether this part of the model is necessary as the influence of inhibition of return on scene viewing behavior has been challenged recently (Smith & Henderson, 2009). To implement this model, we simply set $c_F = 0$ in our original model removing the influence of the inhibitory pathway. As u then cannot become negative anymore, we also replaced the mapping from u to u^* with the identity. As a consequence, all parameters of the inhibitory pathway are superfluous in this model, such that we are left with only four parameters for this model: ω_A, σ_A, λ and ζ.

Divisive inhibition model. The original SceneWalk model implements a subtractive inhibition. However, there are no strong reasons for why this inhibition should be subtractive. An alternative and common model of interaction is divisive inhibition (Carandini & Heeger, 2012). To test this alternative form of combining the two maps, we changed the formula for u to

$$u_{ij}(t) = \frac{[A_{ij}(t)]^{\lambda}}{c_F^{\gamma} + [F_{ij}(t)]^{\gamma}} \qquad (14)$$

As for the model without inhibition, the variable u cannot become negative. Again, we replaced the mapping from u to u^* with the identity. This way to combine excitation and inhibition has the same number of parameters as the original subtractive formulas. Thus, we are left with eight parameters as for the original model.

Estimation of Model Parameters

As it is common practice our previous approach to the estimation of model parameters was based on minimization of an ad hoc loss function that included gaze positions and saccade lengths as measures of model performance (see Appendix in Engbert et al., 2015). First, we computed the squared differences between densities of gaze positions from experimental and simulated data using two-dimensional bins for discretization. Second, we compared experimentally observed and simulated saccade lengths via squared differences from bins of the distributions. The sum of both measures was minimized to obtain parameter estimates.

However, there were several problems associated with this approach that motivated us to develop an alternative framework. First, our earlier approach worked for a limited set of parameters only. Some of the parameters had to be fixed at plausible values. These fixed parameters included important parameters, for example, normalization exponents of the dynamic activation maps, which are critical for the spatial correlation functions we intended to reproduce. Second, the qualitative model analyses necessary to find useful and plausible values for the fixed parameters required time-consuming hand-selected model runs. Third, our earlier fitting approach based on a subset of hand-selected fixed parameters and estimates from minimization of an ad hoc loss-function could not guarantee reliable or consistent estimates and was missing a statistical justification. Moreover, confidence intervals of the model parameters were inaccessible and were, therefore, replaced by an ad hoc indicator of errors of parameter estimates derived from multiple runs of the minimization algorithm. Because of these shortcomings of the earlier approach, we set out to develop an improved strategy for parameter estimation that would be statistically well-founded, reliable, and efficient in terms of computer time, while working for all parameters.

Maximum Likelihood Estimation

A tutorial on the MLE concept for model fitting is given by Myung (2003) in the context of mathematical models in psychology (see Hays, 1994, for a more general context). The general idea is to find the particular (vector-valued) parameter $\boldsymbol{\theta}$ that corresponds to the maximum of the likelihood function given the observed data. This parameter value is used as a parameter estimate and, therefore, termed *maximum likelihood estimate*.

Fitting models to data based on the likelihood has considerable statistical advantages over using other statistics for fitting (Myung, 2003). First, the likelihood guarantees sufficiency, that is, raw data do not constrain the parameters more than the maximum likelihood criterion. Second, for the likelihood, there is asymptotic consistency, such that for large samples the estimate converges to the correct parameter value if the data were generated from the model. Third, the likelihood has asymptotic maximum efficiency, that is, for large samples, there is no consistent estimate with smaller variance. Finally, the likelihood estimate is not changed by the reparametrization of the model, which is known as *parametrization invariance*.

In numerical simulation models like the SceneWalk model, the maximum of the likelihood can be found using an optimization algorithm that evaluates the likelihood $L_M(\boldsymbol{\theta}\,|\,\text{data})$ varying the model parameters $\boldsymbol{\theta}$. Most optimization algorithms try to change the parameters gradually to improve the likelihood and can thus be trapped in local extrema, where the likelihood is higher than for surrounding parameter values, but not the globally best parameter value. If the global optimum is found, it must not depend on the specific optimization algorithm or starting position. Consequently, it is common practice to run multiple optimizations with different starting positions. If one of the local extrema is clearly better than the others and the optimizations end up in clusters, one can be reasonably sure that one found the global optimum.

Alternatively the field of global optimization designs algorithms to find global minima. Two well-known families of algorithms for global optimization are (1) *simulated annealing*, which—inspired by the cooling of physical materials—first explores broadly and later allows fewer bad objective values settling near the optimum (Kirkpatrick, Gelatt, & Vecchi, 1983; Kirkpatrick, 1984) and (2) the *genetic algorithm*, which simulates a population of parameter values over generations in which points with high objective function values have higher probability to reproduce in the next generation (Golberg, 1989; Holland, 1975; Houck, Joines, & Kay, 1995). Variants of both these algorithms are available for most higher programming languages like MATLAB (2016) or python (Jones et al., 2001). As a promising idea for the future the relatively recent metamodeling approach aims to model our knowledge about the function gained so far and to conclude which points to sample to gain the most information about the optimum (Jones, Schonlau, & Welch, 1998; Villemonteix, Vazquez, & Walter, 2009; Hennig & Schuler, 2012).

For optimization of the parameters of the SceneWalk model, we used the genetic algorithm for global optimization as implemented in MATLAB. We used 200 individuals on the logarithm of the parameters with a range from −10 to 10, corresponding to a range from 0.000 045 to 22 026 for the parameters. Subsequently we further optimized using the Nelder-Mead simplex algorithm as implemented as fminsearch in MATLAB. Using the standard settings for all other options these algorithms found the global maximum reliably, as confirmed by some standard optimization runs from random start positions, the sampling we did for Bayesian inference and the fits we computed for cross-validation as described in the following text.

Bayesian Inference

If the likelihood $L_{\mathrm{M}}(\boldsymbol{\theta}\,|\,\text{data})$ of the data can be computed for a given model *M*, then Bayesian inference (Marin & Robert, 2007; Gelman et al., 2014, for overviews) is a viable method for parameter estimation. The main advantage of Bayesian inference in the current context is that it provides not only the best fitting parameter values, but also a full distribution of possible parameter values. Thus, there is information on which other parameter values could also explain the data and thus how well the parameters of the assumed model are constrained by given data. In Bayesian inference, the goal is the computation of a posterior distribution $\mathrm{P}(\boldsymbol{\theta}\,|\text{data})$ that indicates the most probable parameter values $\boldsymbol{\theta}$ under the assumption of model *M* and the given data. Based on the likelihood $L_M(\boldsymbol{\theta}\;|\text{data})$ and a prior distribution $P(\boldsymbol{\theta})$, which describes our knowledge or beliefs about the parameters prior to data collection, the posterior distribution is computed as

$$P(\boldsymbol{\theta}\,|\,\text{data}) = \frac{L(\boldsymbol{\theta}\,|\,\text{data})P(\boldsymbol{\theta})}{\int_{\Omega} P(\boldsymbol{\theta})L(\boldsymbol{\theta}\,|\,\text{data})d\boldsymbol{\theta}}, \tag{15}$$

where, computationally, the main problem is that quantities of interest are usually integrals over the posterior $P(\boldsymbol{\theta}\,|\,\text{data})$ like the expected value of the posterior, its variances or correlations. To compute these integrals, it is often necessary to use Markov Chain Monte Carlo (MCMC) methods (Brooks, Gelman, Jones, & Meng, 2011; Robert & Casella, 2013). These methods produce—sometimes weighted—samples from the posterior using only local evaluations of the likelihood and prior. These samples can then be used to replace integrals by sample means. This especially avoids the direct calculation of the denominator $P(\text{data}) = \int_{\Omega} P(\boldsymbol{\theta})L(\boldsymbol{\theta}\,|\,\text{data})d\boldsymbol{\theta}$, which in turn can be computed from the samples if one is interested in this value.

The most controversial aspect of Bayesian statistics is the choice of prior. The main reason is that the prior may serve very different functions in different situations.

The first, most literal interpretation of priors is that they shall represent all available believes prior to the experiment. If one manages to formulate all prior believes into the prior distribution, the posterior represents the believes one should have after the experiment to do proper reasoning (Jaynes, 2003, Chapter 1). If we had an estimate of the parameters from some other experiment, or had any other kind of information what the parameters or predictions of the model should be, the prior offers a possibility to include this knowledge. In the absence of prior information, the general recommendation is to use relatively broad uninformative priors to avoid biasing the conclusions too much. If a bias is unavoidable, then the recommendation is modified to use a prior which favors the opposite of the suspected conclusion to achieve a conservative analysis showing how well the data should convince a sceptic (Gelman et al., 2014, Chapter 2.8, Jaynes, 2003, Chapter 11 and 12).

The notion of an uninformative prior can be formalized mathematically, which leads to Jeffreys' (1946) priors. Other mathematically preferable kinds of priors are conjugate priors, for which the posterior has the same form as the prior (Gelman et al., 2014, Chapter 2.4), such that posteriors can be parametrized and analytically analyzed. Neither Jeffreys' priors nor conjugate priors are particularly relevant for the complex models we study here, as they are rarely known or even computable for highly complex models.

A second more objective interpretation is that the priors shall represent the actual distribution of parameters as close as possible. In this interpretation, which is popular in machine learning, the prior becomes part of the model to be evaluated. The better the prior represents the distribution of parameters needed to fit data, the better it is. Obviously, such evaluations require multiple instances for which a parameter is fitted. Once one starts to adjust the prior to fit some data, this approach becomes essentially equivalent to hierarchical models which we discuss in the following text.

Prior assumptions on parameters also represent a helpful tool to include information obtained from other experiments and other knowledge (e.g., physiological constraints) or to regularize the model, which is a general expression for preferring some parameter values of the model over others, if both parameter values explain the data equally well. The term regularization is used usually in Frequentist contexts and justified as a means to stabilize model fitting when the parameters are not sufficiently constrained by the data.

For regularization purposes, one typically differentiates whether parameter values are considered only less likely or impossible. Only the former is usually called *regularization*, whereas the latter is usually called *constrained estimation*. This distinction is mainly necessary because once there are areas of parameter space which are impossible the algorithms for optimization or sampling need to be changed. For the effect of the priors on the model, this is a more gradual distinction. Although it is usually discouraged to entirely exclude parameter values a priori, that is, to set their prior probability to 0, very small prior probabilities will have the same effect on the model predictions and parameter fits.

The different aims for priors partially work against each other. Regularizing or including prior knowledge helps mostly when the parameters cannot be constrained well by the data at hand, that is, when the prior excludes parameters that can also fit the data convincingly. When doing this one can obviously not interpret the posterior as information how well these parameters are constrained by the data. Thus different aims might require different priors for the same model and data.

As we do not require regularization and have little to no prior information about the parameters of the model we investigate, we chose an extremely broad prior not to influence our parameter estimates. We assume a log-normal distribution with a standard deviation of 30 units (log-space) around 0 (in log-space).

Results on Model Parameter Estimation

For the SceneWalk model, we used the same dataset as in the original article (Engbert et al., 2015). In the experimental data, gaze positions were recorded via eye tracking from 35 human observers in a memorization task. Experimental stimuli consisted of 15 natural images and 15 texture images, where the latter are photographs of relatively homogeneous textures like grass or a stone wall.

The numerical optimization of the model parameters required less computation time than the original fitting method, as the likelihood objective is not stochastic, although we fitted four more parameters (the pooling exponents λ and γ, the weighting of the inhibitory map c_F and the weight of the uniform map in the mixture ζ).

The results of the maximum likelihood estimation (MLE) are listed in Table 1. As they agree with values from Bayesian estimation we shall discuss their meaning after explaining the origin of the Bayesian estimates. To perform Bayesian inference about the parameters of the SceneWalk model, we sampled the posterior distribution with a Metropolis Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). A hand-tuned multivariate Gaussian proposal distribution was chosen to have a covariance matrix roughly proportional to the covariance of the sampled distribution and to reach an acceptance rate of roughly 25% as recommended as optimal for Gaussians by Gelman, Roberts, and Gilks (1996). We restricted us to reproduce the diagonal of the covariance matrix, that is, to the variances of the individual parameters, and three particularly strong covariances, between σ_A and σ_F, C_F and λ and C_F and ζ respectively. Using this scheme, we sampled three chains with 50,000 samples each starting with a small displacement from the Maximum A Posteriori (MAP) estimate. We then discarded the first 1,000 samples as burn in, which covered the initial transient back toward the MAP in all parameters.

First we checked that our sampling algorithm converged using the $\hat{R}$ statistic (Gelman & Rubin, 1992; Brooks & Gelman, 1998), which quantifies how large the variance between chains is compared with the variance within the chains, that is, whether the chains sampled different regions. The $\hat{R}$ statistic is always greater than one and, when the chains under analysis converged to the same stationary distribution, the $\hat{R}$ statistic should be close to one. For our chains we obtained values in the range from 1.00 to 1.06 for different parameters and a value of 1.06, when $\hat{R}$ was computed as a multivariate statistic. We thus concluded that our chains converged to their common stationary distribution, which we also confirmed by investigating visually and by comparison of the distributions obtained from the three independent chains.

Next we checked that our chains mixed sufficiently well, that is, we tested that the samples were sufficiently uncorrelated with each other and, therefore, that the samples provide an adequate representation of the posterior distribution. The mixing property was analyzed via the effective sample size, which is an estimate of the number of independent samples one would need to get an equally good representation of the posterior. This estimate is computed from the autocorrelation of the chain for each individual parameter. As a result, we obtained an estimate of the effective sample size for each parameter, although the true efficiency of the sampling algorithm is a single quality of the method. For our chains, the effective sample sizes turned out to range from 624 to 22,806 for the different parameters. This indicates that our sampling algorithm provides at least the information of a few hundred samples, which we considered as sufficient for our purposes.

However, our findings on the effective sample size also indicate that the Metropolis Hastings algorithm could probably be improved in efficiency as its sampling efficiency (effective sample size divided by the number of drawn samples) was less than 1%. When the algorithm is well tuned to the problem, a sampling efficiency of several percent can be reached (Gelman et al., 1996).

The sampled posterior distributions are displayed in Figure 3. The distributions clearly indicate the most likely values of the parameters. All parameters except for the decay of the excitatory map ω_A and the exponent γ were well constrained by the data. Their posterior marginals concentrate on a range of $\leq \pm 10\%$ around the best fitting values and are much narrower than the prior (± 10 log-units).

From an analysis of the marginal posterior distributions displayed in Figure 3, we can extract point estimates and credible intervals, which characterize a single optimal model parameter and a range that contains the true parameter value with a given probability. For our model we extracted the mean estimate and a 95% credible interval for each parameter listed in Table 1 to compare them with the parameter estimates obtained in the original article (Engbert et al., 2015). For the well constrained parameters the MLE and mean estimates agree closely as expected. These esti-

Table 1
Table of the Parameter Values Obtained From Different Point Estimates

Parameter name	Original estimate	MLE	Posterior mean estimate		95% Credible interval	
ω_A	6.607	2.4×10^{30}	1.1×10^{45}	$\pm 8 \times 10^{44}$	417.6	4.373×10^{30}
ω_F	.00903	1.9298	1.973	±.001601	1.876	2.071
σ_A	4.88	5.9082	5.903	±.000640	5.838	5.967
σ_F	3.9436	4.5531	4.558	±.002282	4.445	4.671
γ	.3[a]	44.780	3.3×10^{12}	$\pm 4.5 \times 10^{11}$	43.83	3.249×10^{13}
λ	1[a]	.8115	.8130	±.000422	.7896	.8354
c_F	1[a]	.3637	.3605	±.000321	.3658	.3767
ζ	—	.0722	.0712	±.000046	.0662	.0764

Note. Displayed are the maximum likelihood estimate (MLE), the posterior mean estimate (±estimated sampling error), and a confidence interval from the Bayesian estimation we present, compared with the values from the original study by Engbert, Trukenbrod, Barthelmé, and Wichmann (2015).
[a] Values were fixed without fitting in the original article.

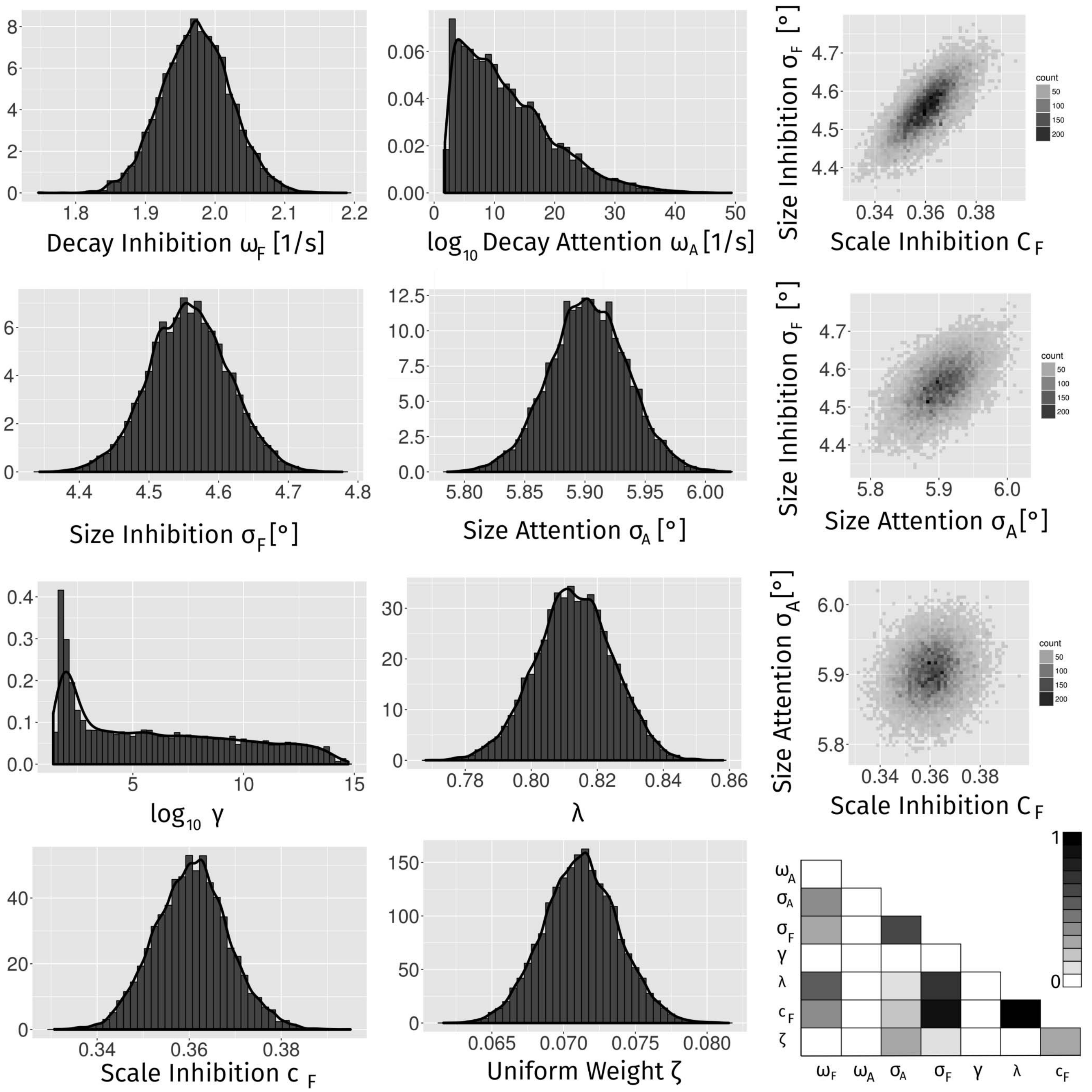

Figure 3. Sampling results for the posterior distribution for the example model's parameters. In the left two columns we show histograms and density estimates for all eight parameters. Except for γ and ω_A, all parameters seem to be well constrained by the data. In the right column we show two dimensional histograms of two parameters against each other illustrating their dependencies. The first indicates the strong correlation between the spatial scale and scaling factor of the inhibition. The second shows the medium strength dependency between the sizes of inhibition and attention pathway. The third plot illustrates the near independence of the spatial scale of the attention map and the scaling factor highlighting the nontransitivity of correlations. In the lower right corner, we present a summary plot about the correlations between parameters. The darkness of each rectangle in this plot indicates the absolute correlation between two parameters, which each could be shown as a two-dimensional histogram as we did for the previous three examples.

mates can only differ when the posterior is relatively broad. Consequently, our interpretation is the same for both parameter estimates.

Qualitatively, we reproduce the patterns observed in the original article: The activation on the excitatory attention map is larger and faster than the inhibitory fixation map ($\omega_A > \omega_F$, $\sigma_A > \sigma_F$). Quantitatively, the parameters differ substantially from the ones in the original study. In particular, compared with the original study, (a) the Gaussian input around the current fixation is larger by roughly a degree for both maps, (b) the inhibitory fixation map is 2.5 log-units faster, the attention map could be arbitrarily fast and (c) the pooling exponents (γ and λ) converged to very different values than those chosen by hand.

The fact that the two parameters γ and ω_A are not well constrained can be explained as follows. The parameter ω_A determines the rise-rate of the attention map. Once this rate is fast enough, changes of the parameter value will not influence predictions any more. Similarly, high values of gamma produce all very similar nonlinearities in the inhibition map and thus do not change any predictions. As we discussed earlier, one could have used a prior to restrict these parameters to ranges over which they change predictions to avoid the result of parameters which are unconstrained over such wide ranges. This would however hide the fact that they are not well constrained from the posterior sampling result.

From the posterior distribution, we can also extract two-dimensional marginal distributions as histograms or density estimates. These marginal distributions illustrate posterior couplings between pairs of parameters. Such couplings indicate that obtaining information of one of the two parameters would constrain both of them better. For example, we show two-dimensional histograms for 3 pairs of parameters (see Figure 3):

- For σ_F and C_f we find a relatively strong coupling which indicates that models with stronger inhibition require it to be spread wider to explain the data equally well.
- For σ_A and σ_F we find a weaker, but still visible coupling, which indicates that the inhibition and attention window need to covary in size to explain the data.
- Finally, σ_A and C_F turned out to be approximately independent. Fixing one of these parameters would not constrain the other parameter.

This last point additionally illustrates that posterior correlations are not necessarily transitive.

In summary, the posterior marginal distributions can be reduced to the correlation coefficient, which captures the strength of the linear dependence between the parameters. These correlation coefficients are also plotted in Figure 3 for each combination of two parameters. The samples from the posterior also contain all higher-order dependencies between parameters, although they are more difficult to visualize or summarize.

Inter-Participant Differences and Hierarchical Models

For many cognitive tasks, participants differ in meaningful ways, which we might want to include into our models. For eye movements, one important participant-specific parameter is the average length of saccades (Castelhano & Henderson, 2008). For our participants who generated the longest saccades, we observed average saccade lengths twice as large as the saccade lengths for participants with the shortest saccades (see Figure 4).

One popular method for integrating differences between participants into models are hierarchical models. In hierarchical models the differences between participants are explained by assuming different parameter values for each participant which follow an additional model for the distribution of parameters in the population.[3] The main advantage of using a model for the distribution of parameters in the population is to stabilize the estimates for participants, whose parameters are not well constrained by the data alone.

We implemented a hierarchical model which allows the sizes of the attention span and of the inhibited area to differ between participants to explain the observed differences in saccade length. To simplify the analysis, we fixed all other parameters of the model to their MAP estimates over all participants and images from the model fitting explained earlier.

As our model for the parameter distribution in the population, we introduced a two dimensional Gaussian, which we parametrized using means and variances for the two parameters and the correlation between parameters as a fifth parameter. As we now aim to estimate these five parameters together with the individual-participants parameters, we defined a prior on each parameter individually and assumed the priors to be mutually independent. For each of the means and their correlation we chose a uniform distribution, whereas for the variances we selected an inverse Gamma distribution with parameters 0.25 and 1, which yields a very broad distribution over the positive real axis with a peak at 1.

It is possible to fit the hierarchical model using the same procedures we applied to the orginal model. We skip optimization and frequentist analysis here though. Instead we directly sample the posterior using Gibbs sampling (Casella & George, 1992) with parameter groups for each participant and one group for the hyperparameters, sampling each marginal distribution using the Metropolis Hastings algorithm. Specifically, we first cycled through each participant performing one Metropolis Hastings sampling step for the corresponding two individual parameters. Next, we performed one Metropolis Hastings step for the parameters of the Gaussian distribution, which was assumed for the parameter distribution in the population. All proposal distributions were Gaussians with diagonal covariance matrix, adjusted by hand to approximately achieve 25% acceptance rate, and variances roughly proportional to the posterior variances of the parameters (Gelman et al., 1996). We used the same proposal distribution for each participant. Gibbs sampling is especially efficient for hierarchical models because sampling the parameters of each participant requires only the likelihood for the data of that participant. Thus a whole sweep is computationally only as costly as single likelihood evaluation for updating all parameters. We sampled three chains of 10,000 sweeps through the parameters each starting at the maximum a posteriori estimates over all data. As burn in we removed the first 1,000 samples of each chain, which seemed sufficient after visual inspection of the chains. This yielded an effective sample size between 347 and 4,472 for the different parameters and the chains seemed to have converged according to visual inspection of

[3] The hierarchical model framework can also be used to model effects of other properties of the task like item and image effects.

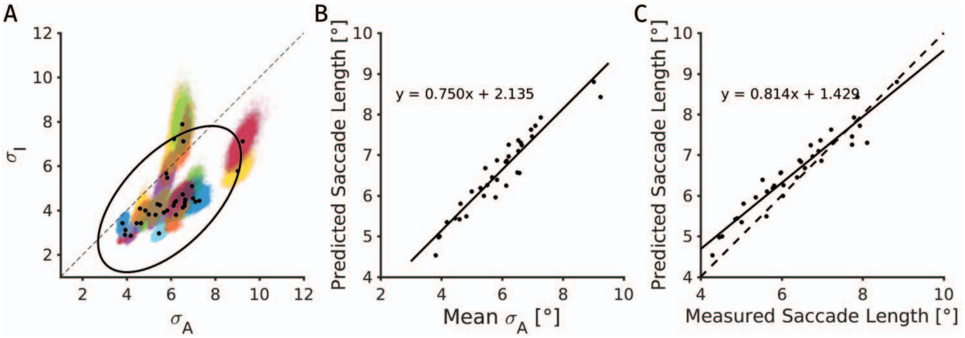

Figure 4. Results for the Hierarchical model. Panel A: Fits for the two parameters σ_A and σ_F for the different observers. Each observer is represented by a black marker marking their posterior mean and a colored point cloud representing the posterior samples. Additionally the dashed line marks the $\sigma_A = \sigma_F$ diagonal and a large black ellipse marks the 95% contour line of the Gaussian population model estimated as the posterior mean over hyper-parameters. Panel B: Predicted saccade length for each participant against their posterior mean estimate for σ_A with a linear least squares regression line. Panel C: Predicted mean saccade length from the posterior mean estimate against the measured mean saccade length for each participant. The dashed and continuous line marks the equality diagonal and a linear least squares regression line. See the online article for the color version of this figure.

the chains and the $\hat{R}$ statistic which had an upper CI bound of 1.06 or less in all cases.

The results of the hierarchical model analysis are shown in Figure 4. First in Panel A, we observe that different participants are fitted by considerably different sizes for both σ_A and σ_F and that the estimates for the two parameters are highly correlated, that is, participants who have a larger fitted attention span also have a larger fitted inhibition area. Second in Panel B we show that the mean saccade length predicted by the model depends strongly on σ_A and consequently on σ_F, as they are highly correlated. Finally, we compare the measured mean saccade length to the mean saccade length predicted by the fitted model by simulating as much data as measured for each participant with their posterior mean parameters. The two observables are strongly related, indicating that varying the two spans in the SceneWalk model could account for the difference in saccade length between participants. Additionally, we can observe that the predicted mean saccade length grows with a slope slightly smaller than 1 with the measured saccade length, indicating a slight regression to the mean, as expected and intended for a hierarchical model.

Looking at the individual participant estimates more closely, we can observe that most participants (30 of 35) fall into a large cluster, with slightly smaller σ_F than σ_A. However, three participants have larger fitted inhibition spans and two participants have extraordinarily large attention and inhibition spans.

Model Comparison in the Likelihood Approach

The likelihood concept can be used as a general approach to evaluate how well a given model fits experimental data. Thus, it is possible to compare different models. For likelihood-based comparisons between models one usually assumes fitted parameters. Thus one uses the maximum likelihood, that is, the best likelihood value a model can reach on the data, when the model's parameters are optimally adjusted. In the following, we denote the maximum likelihood as $L(M) = \max_\theta L_M(\boldsymbol{\theta} \mid \text{data})$.

For the comparisons that we will carry out in the following text, it is important that the log-likelihood is always a relative measure because it depends on the grid for the observation of fixation positions, the size of the dataset and other dataset specific aspects. Therefore, only the log-likelihood-ratios between models can be compared between different data sets, models, or viewing conditions. Given a null model M_0, which defines a reference point, one can compute a likelihood ratio Λ to compare a model M_1 to the model M_0, i.e.,

$$\Lambda(M_1) = \frac{L(M_1)}{L(M_0)}. \tag{16}$$

The likelihood ratio Λ informs about how many times more likely the data are generated by model M_1 than by model M_0. For theoretical considerations and for most computations the log-likelihood ratio λ is a better choice,

$$\begin{aligned}\lambda(M_1) = \log(\Lambda(M_1)) = \log\frac{L(M_1)}{L(M_0)} \qquad (17)\\ = \log(L(M_1)) - \log(L(M_0)).\end{aligned}$$

The log-likelihood ratio is additive and can be interpreted in a straightforward way, for example, if M_2 is one bit better than M_1, which is one bit better than M_0, then M_2 is two bits better than M_0 and the data are 4 times more likely under model M_2 than under model M_0.

Also, the log-likelihood ratio can be interpreted in information theoretic terms as the *information gain* about the data generated by the new model compared with the information explained by the original model. Thus the log-likelihood ratio measures how much communication could be saved when specifying a sequence of fixations using a code based on the model. As information theory is well developed (Ash, 1990, for an introduction), it provides a strong theoretical background for log-likelihood ratios in model comparisons.

In principle likelihood ratios measure the relative quality of the model fits. However, models tend to fit aspects of the data which are purely random, a phenomenon known as *overfitting* (e.g., Dietterich, 1995). Overfitting is the main reason why *model selection*—to which Zucchini (2000) gives an introduction for psychologists—should not be done by directly comparing the likelihoods based on the data used for fitting the models (Myung, 2000). Ultimately the goal of model comparison approaches is to compare the expected likelihood on new data, not on the data used for fitting. Proper model selection and comparison methods are especially critical for comparing models which differ in their flexibility. More flexible models always explain more details of the dataset they are fit to, and thus produce larger likelihood values for the dataset they are fit to. However, more flexible models should only be preferred if the additionally explained details generalize to new data.

There are two popular quantities model comparison techniques try to estimate and use for comparing models. The first one is the *out-of-sample-prediction error* (Gelman, Hwang, & Vehtari, 2013), that is, one tries to estimate the likelihood of the parameters fitted on the given data on a new dataset. The second one is the *evidence* for a model which is the denominator of the Bayesian formula—$\int_\Omega P(\boldsymbol{\theta})L(\boldsymbol{\theta}\,|\,\text{data})d\boldsymbol{\theta}$—that is, the total probability to observe the data according to the model with the given prior $P(\boldsymbol{\theta})$. For a new dataset this means the evidence estimates the models performance using only the prior information about the parameter value. Consequently the evidence critically depends on the prior and can be arbitrarily bad if the prior assigns large probability to parameters with low likelihood. The ratio of evidences for two models is called the Bayes factor.

The first approach for model selection are metrics which add a correction or penalty term for more flexible models. These metrics are generally called information criteria and are usually formulated in terms of the *deviance* ($-2\lambda(M)$)—a general measure of prediction error—which is directly computed from the likelihood and contains exactly the same information, but reverses the sign. Thus smaller information criteria correspond to better models.

Classical examples for this procedure are the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). The AIC was formally introduced as a first model selection criterion, defined as $AIC(M) = -2\lambda(M) + 2\dim(M)$.[4] It represents a simple large sample bias correction obtained from Fischer information theory estimating out-of-sample-prediction error. The BIC (Schwarz, 1978) was introduced as an approximation to the evidence in favor of a model in the case of an exponential family model. Thus it effectively aims to estimate the generalization quality to new data which requires new fitted parameters. For n independent observations it is defined as: $\text{BIC}(M) = -2\lambda(M) + \log(n)\dim(M)$.[5] This obviously does not contain the prior and is a coarse approximation to the evidence. From very small data sets on this penalty will be larger for the BIC than for the AIC, for example, the BIC will prefer parsimonious models more strongly than the AIC corresponding to the harder generalization task estimated by BIC.

The classical information criteria—AIC and BIC—both result in very small corrections of the raw likelihood. Our dataset contained 13908 and 13306 fixations for natural images and texture images respectively. Thus for our model with 8 free parameters the AIC and BIC penalties would maximally be $0.0008\frac{\text{bit}}{\text{fix}}$ and $0.0041\frac{\text{bit}}{\text{fix}}$ respectively, whereas the differences between models are much larger. In contrast, our cross-validation results in the following text suggest that the actual difference between fitted data and new data is much larger. Thus AIC and BIC seem to provide bad estimators in our case of complex dynamical models.

Very similar Bayesian evaluations exist (Gelfand & Dey, 1994), which estimate generalization of the posterior predictive distribution instead of generalizations based on a point estimate for the parameters. Nonetheless, the aim stays to predict how likely new data will be according to the model.

Fortunately direct formulas to approximate model performance in fully Bayesian terms from sampling results exist (Gelman et al., 2013). Thus a Bayesian Model comparison is possible, once a representative sampling is available for the posterior on the parameters of each model. Examples for this approach aimed at generalization to new data from the same parameters are the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) which approximates the posterior as the mean estimate and the Widely Applicable Information Criterion (WAIC, Watanabe, 2010), which directly uses the sampling estimate for the posterior predictive. Both these criteria also use the posterior samples to their advantage to produce a more accurate estimate for the out of sample prediction quality. Similarly, there is also a Bayesian alternative to the BIC, the Widely Applicable Bayesian Information Criterion (WBIC, Watanabe, 2013).

Calculation of the Bayesian information criteria requires an estimate for the posterior distribution on the model parameters, that is, a sampling of the posterior. As we compare 10 models and only have a sampling for one of these models, we do not perform these analyses here. However, such analyses should be considered especially when one studies other models like hierarchical models for example for which cross-validation is not straightforward. And of course, once the posterior predictive is used for prediction, this should be the measure to be compared in the cross-validation.

One should note that the penalties of all information criteria per data point (i.e., fixation or scanpath) converge to zero for growing dataset size. Thus larger data sets will raise a preference for more detailed models if there is any advantage for prediction. This makes sense as the criteria penalize complexity only when this complexities cannot be calibrated well enough to improve predictions with the given data (Burnham & Anderson, 2004).

A different more data driven approach to estimate the quality of out of sample predictions is *cross-validation*, which is frequently used in machine learning, but has been introduced to the psychological literature as well (Browne, 2000). For cross-validation the dataset is split into n subsets. Then the model is fitted to $n - 1$ of the subsets—the *training set*—and evaluated on the one subset not used for fitting—the *test set*. This is repeated for each of the subsets being the test set and the results are averaged. This procedure applies to Bayesian and Frequentist evaluation equally, but is more frequently used with point estimates and Frequentist evaluation.

For dynamical models for eye movements in scene viewing, two separate factors induce variability for which overfitting could occur: human observers (participants) and stimuli. To avoid problems of overfitting for these two factors, we split our

[4] Dim(M) representing the dimensionality of the model, that is, the number of parameters, n the number of independent observations.

[5] The original criterion was half the value described here. However, the version reported here seems to be the more commonly used one today.

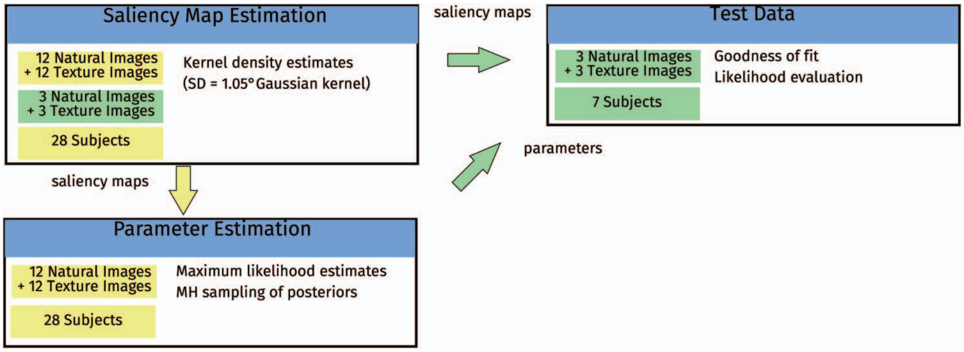

Figure 5. To guarantee that the model is fit to a different dataset than the one used for evaluation many possible separations exist. Here we display the separation of our dataset into training and test data used for each fold of cross-validation. Data from 28 human observers on 2 × 12 images (yellow/lighter) were used for parameter fitting, whereas the data from 7 different observers on 2 × 3 test images were used for model tests (green/darker). See the online article for the color version of this figure.

data across both factors and perform fivefold cross-validation using splits into training and test set as illustrated in Figure 5: For each fold we used the data obtained from 28 participants on 12 natural images and 12 texture images for *training*. For evaluation we run the model on data obtained from 7 other participants on three other natural images and three other texture images. To compute the empirical saliency maps, we used the 28 training participants on both training and test images. There are also data for the training participants on the test images and the test participants on the training images, both of which are not used here to completely isolate training and test sets from each other.

For each fold we fitted the model to the training data using the genetic algorithm of MATLAB with settings as for the original fitting process on all data described earlier. However we noted that there was exactly one more local maximum to be found at small ($\sigma_F \approx .5°$), fast ($\omega_F \geq 10$) inhibitions, to which the genetic algorithm converged for some folds. To find the global maximum in every case nonetheless, we started a subsequent fminsearch optimization from each of these two maxima for each fold and took the better one as the global maximum. In all folds and all models, the global maximum had similar sized attention window and inhibition and generally similar parameter values to the fit of the subtractive model to all data described earlier. The other local maximum was usually around 1,000 worse on the log-likelihood scale for the training data. Thus the decision was always clear cut. Nonetheless this additional local maximum can be understood. Effectively it implements an inhibition for saccade targets very near to the current fixation. Saccades to these targets would not be detected as such by the data preprocessing such that such short saccades indeed do not occur in the dataset and cannot occur in a dataset. Thus this model adaptation indeed would be predictive, but not informative about any underlying processes of eye movement behavior.

Results on Model Comparison

To perform our comparison we split the data as explained earlier, fitted the model to each of the five training sets and computed the log-likelihood of each model on each test dataset. Then we divided the resulting likelihood value by the number of fixations to normalize the results regarding the size of the dataset. Thus we measure all differences in bits per fixation [bit/fix]. According to this null model, the uniform distribution over the whole image distributes a probability of 2^{-14} for every fixation to each grid point because we calculated all maps on a 128 × 128 grid. This results in a log-likelihood of −14 bit/fix. We ran separate evaluations for texture images and object-based natural scenes presented in the experiments; the log-likelihoods are plotted in Figure 6. Overall, we find a gain for the empirical saliency model over center-bias prediction and a considerable gain in likelihood for the SceneWalk model.

The information gain for the saliency model differs strongly between natural textures and natural scenes, which was expected as the gaze patterns over texture images were more uniform than the corresponding data for natural scenes. This difference carries over to our dynamical model, as this uses the empirical saliency as an input predicting where human observers want to look. However, the increase in likelihood due to the dynamical principles is comparably large for texture images and for scenes. This result lends support to the view that the same dynamical principles of scanpath generation are underlying texture images and natural scenes.

We also evaluated the model with the parameters values fitted by Engbert et al., (2015). This yields a likelihood value of −12.96 bit/fix for natural images and −13.10 bit/fix for texture images for the training data (not shown in the figure). This indicates that the model explained the data better than empirical saliency even with the parameters not optimized for the likelihood. However, with the new parameter values the model generates higher likelihood values per fixation on the test sets it was not trained on (natural scenes: −12.38 bit/fix, textures: −12.68 bit/fix).

To compare different model specifications against each other, we generated two new model variants—one without inhibition and one with divisive inhibition—described in detail earlier. Additionally we questioned whether the introduction of the exponents λ and γ were necessary. To test this, we generated model variants with one or both of the exponents fixed yielding four variants of the

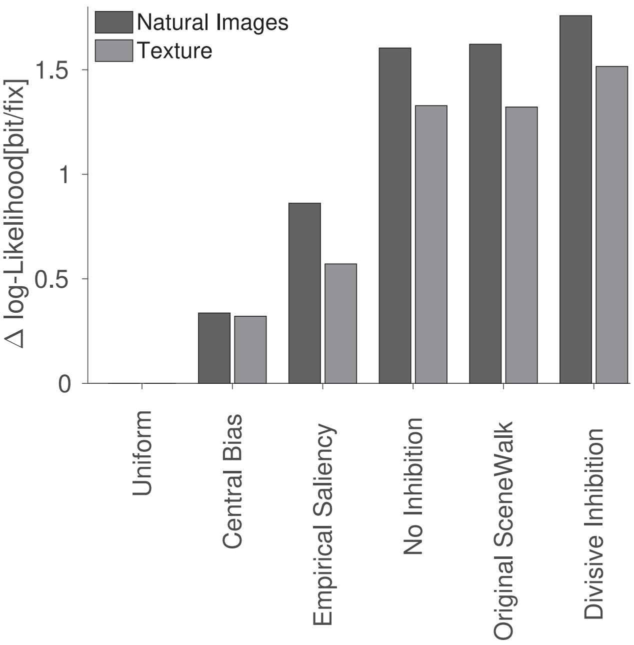

Figure 6. Bar plots for the models' log-likelihood differences to the uniform distribution null model. We split here by the two experimental conditions, which differed in the images presented. For the texture models the density map is much less informative than for the natural images. The central bias/central fixation bias model is a kernel density estimate from the fixations on all other images. The empirical saliency is the kernel density estimate from the fixations of other observers on the same image. Finally, no inhibition, original SceneWalk, and divisive inhibition refer to the three variants of the SceneWalk model, which we investigate in detail.

subtractive original SceneWalk model, four for the divisive model and two for the model without inhibition.

First, as a check on the results it is informative to look at the performance of the models on the training data, we display in Figure 7A, although these values should not be used for model comparison. Evaluated on the training data a model which contains another model as a special case must be at least as good as the contained model on each of the training sets. This sanity check was how we first noticed that some of the optimizations had ended in a different, wrong local maximum. Also comparing the training set and test set results provides some insight how substantial the flexibility problem is for the specific model.

The test set results of these more detailed comparisons are displayed in Figure 7B. We find that overall the divisive inhibition model provides the best performance followed by the original SceneWalk model and finally the model without inhibition. Within each model type the exponent γ seems to improve the model fit, whereas the fits with free λ yield equally good performance or even worse performance than fixing $\lambda = 1$ (using the attention map without nonlinear distortion). The model to choose from our pool is thus the divisive inhibition model with a large, fitted γ and λ fixed to 1.

Note that all the models with inhibition have a qualitatively similar behavior and typically computed statistics on scanpaths cannot discriminate these models, as we discuss in the following text. Thus the likelihood based comparisons allow us to differentiate models we could not differentiate otherwise. A restriction of these model comparisons is, however, that they do not come with a measure of uncertainty like standard errors, credible or confidence intervals or adequate statistical tests.[6] Thus we cannot provide a hard statistical measure how sure we are about the order of the models although the differences can be interpreted in size.

Goodness-of-Fit for Specific Measures and Spatial Statistics

Although we used the likelihood as a general measure of model fit to experimental data, the likelihood remains a relative (i.e., depending on a null model) and global measure (i.e., no specific statistical properties are addressed). Thus, there are at least two reasons to check other statistics additional to performing a likelihood-based approach to parameter estimation or model comparison. First, to analyze the absolute performance of the model, and, second, to understand which aspects of the data are modeled adequately and which other aspects are modeled poorly.

The first reason, judging the absolute quality of models, is to check that they are good enough to be interesting, which is subsumed under goodness-of-fit analysis in statistics (Pitt, Myung, & Zhang, 2002; Wichmann & Hill, 2001). In statistics, the importance of goodness-of-fit analyses is emphasized because the theory of parameter estimation for models is built on the assumption that there is a correct solution, that is, model parameter values exist that actually generated the data. So, if a model cannot explain the data well for any parameter value, the best estimate for the parameter might be meaningless, even when the best parameter value is defined by generating the highest likelihood for a given model. For the same reason, Bayesian inference methods may fail if there are no good models in the set assumed a priori.

To get an idea about the absolute quality of the model's predictions for data, the easiest way is to simulate data by the model and to compute statistics for these data in exactly the same way as it is done for the interpretation and statistical analysis of experimental data. A comparison of the resulting statistics gives a good indication of the quality of the model's fitness.

On the basis of the likelihood it is also possible to test how (un-)likely the measured data are compared with the expected likelihood of data from the model. This expected likelihood can be computed by simulating larger amounts of data from the model and computing its likelihood. For a perfect fit, the measured data should have a similar likelihood as data sets simulated from the model, which represents a test whether the model's output variability matches the variability of the observed data.

We performed such an analysis by simulating as much data as we had collected and computed the likelihood of this data. We compare histograms over the log-likelihood per fixation for simulated and experimental data in Figure 8. First, in Figure 8A, we ran the analysis on a model without the mixture with a uniform distribution, that is, choosing $\zeta = 0$. According to this model some of the observed fixations were extremely unlikely, that is, the model predictions were to specific, which motivated us to include the mixture with a uniform distribution. In Figure 8B, we show a

[6] Some classical χ^2 tests of model fit exist. As they are based on the same approximations as the AIC and BIC, we doubt that they produce correct conclusions here.

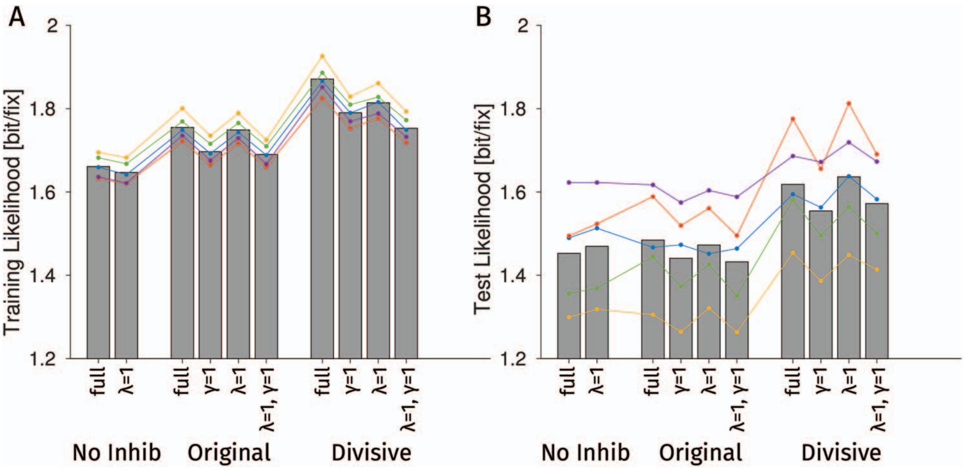

Figure 7. Bar plot comparing log-likelihood differences to the uniform distribution null model, exploring the effects of the exponents. Each bar is the average test set performance of the five folds of our cross-validation procedure. The colored lines plot the results for the five folds. Panel A: The likelihoods on the training data sets, which should not be used to judge the models, but are informative, whether the model fitting worked properly. Panel B: The likelihoods on the test data sets, which can be used to compare models. See the online article for the color version of this figure.

histogram of the log-likelihoods for the full model, again for the measured data and simulated data from the model. For the full model, the mean log-likelihood of the simulated data is -12.11 bit/fix, $\Delta = 1.89$ bit/fix (raw value, difference Δ to a uniform distribution), which is roughly equal to the likelihood for the training data of -12.08 bit/fix, $\Delta = 1.92$ bit/fix, but larger than for the test data for which the model reaches only -12.67 bit/fix, $\Delta = 1.33$ bit/fix. The small difference between training data and model-generated data suggests that the model did not overfit the data dramatically, that is, we would expect the model to be roughly as good as it is for the data, if the data were generated by the model. The difference between training and test data suggests that the model does not generalize to the test dataset perfectly, which is mainly caused by an increased number of highly unlikely fixations (Figure 8B). It seems plausible that these are fixations in regions where none of the observers in the training set fixated (regions of low empirical saliency). This indicates that a higher number of observers for estimating the empirical saliency map would be beneficial to our approach.

The second motivation for additional model analyses is to decide which aspects of the data are modeled well, and which are not described adequately. It is important to further improve models and to choose appropriate models for different situations and modeling goals. Generally, measures used for this analysis should be interpretable for the modeler and other researchers. Some more detailed information can also be extracted from the likelihood calculations as this calculation is split over the different observations. Thus for each individual observation a separate likelihood can be computed and one can check which measured scanpaths or individual fixations are especially likely or unlikely according to the model providing some additional, more specific information.

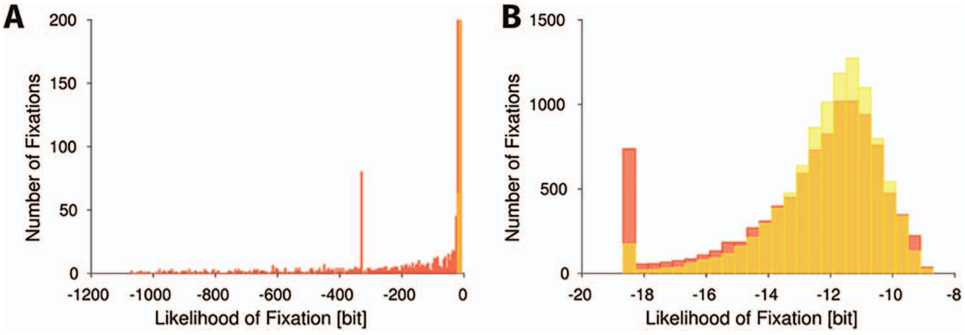

Figure 8. Histograms of the likelihood of individual fixations on the test dataset (red) and on data generated from the model (yellow). Panel A: Using a model without mixing with a uniform distribution (setting $\zeta = 0$ in Eq. (13) and using an earlier method to allow fixations at points with $u^* = 0$). The considerable number of extremely unlikely fixations led us to include the mixture with a uniform distribution in Eq. (13). Panel B: Using the full model with the mixture, extremely unlikely fixation positions no longer occur. See the online article for the color version of this figure.

For the SceneWalk model, we started with an analysis of standard statistics from eye-movement experiments. As a first step, we compared the overall fixation density of model and data. To quantify the comparison, we computed the Kullback Leibler divergence (KL-divergence) of the fixations predicted by the model against the fixations made in our experiment. This standard measure is computed as

$$KL = \int_I p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{18}$$

where the integral is computed over the full image I.

The fixation density generated by the model does not fit the empirical saliency perfectly, but perturbs it slightly through its dynamics. However, the predicted distributions diverge less from the true density (average KL-divergence = 0.1997) than any saliency models, which minimally reach 0.54 and 0.37 for the two data sets in the MIT saliency benchmark (Bylinskii et al., 2016). The good performance of the SceneWalk model is not surprising here because we used the empirical fixation density as an input to our model.

Next, we looked at the distribution of the saccade lengths, a first aspect of the model dynamics. The results of this analysis are given in Figure 9. The saccade lengths in the model and data are very similar and the variance over images is small in both model and data, whereas the variance over participants is substantial as we discussed earlier. Also the competitor models without inhibition and with divisive inhibition fit the distribution of saccade lengths well such that the saccade length distribution does not clearly differentiate these models from each other. However, simply drawing fixations independently from the empirical saliency map yields an entirely different, wrong distribution.

Recently, methods from the theory of spatial point processes were introduced into the analysis of fixation patterns in scene viewing (Barthelmé et al., 2013; Engbert et al., 2015). Most of the standard statistical measures are first-order statistics, for example, the two-dimensional density of fixations. For the SceneWalk model, we computed the pair correlation function (Engbert et al., 2015) as an example for a second-order spatial statistic. The pair correlation function describes how frequently two fixations with a certain distance occur in one scanpath normalized against the frequency expected for a random selection from the fixation density. Values higher than one indicate that fixation patterns are more aggregated than could be expected from the first-order spatial inhomogeneity of the process. As the pair correlation function includes later returns to earlier fixated positions, this function measures a different property than the saccade length distribution. In experimental data, the pair correlation function usually indicates a clustering at small distances below 3 and 4 degrees (Engbert et al., 2015). Comparing the pair correlation functions estimated from experimental data and model predictions in Figure 10, it is obvious that all models fit the pair correlation function much better than a simple random process that draws fixations from the empirical density map. However, this measure seems not to differentiate between the different types of inhibition either.

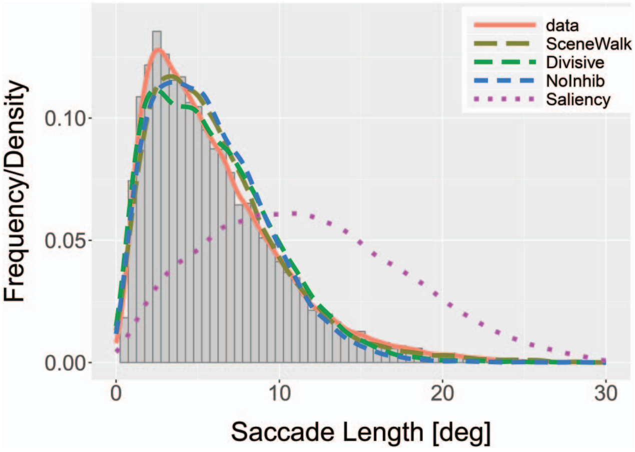

Figure 9. Comparison of model and data based on saccade lengths. The plots present the saccade length distribution over all images for experimental data and model simulations. See the online article for the color version of this figure.

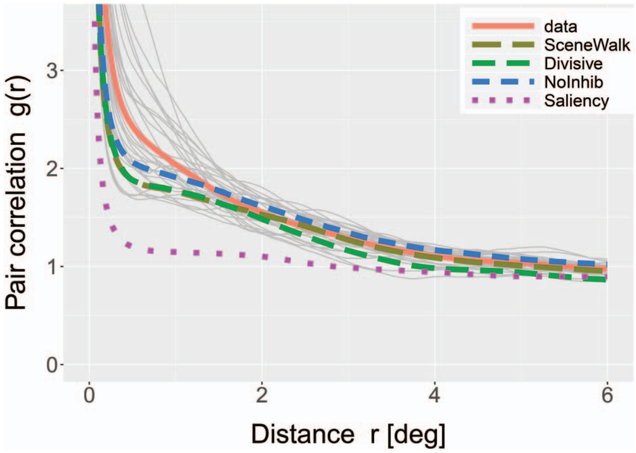

Figure 10. Comparison of models and data based on the pair correlation function (PCF). The mean PCF for each of the models is plotted in color. For the data the mean is shown in color as well and the pair correlation functions for individual images are plotted in gray. Higher values than one indicate clustering or aggregation, that is, fixations at distance r are more abundant than expected on average from independently drawn fixations from the fixation density. Values smaller than one indicate repulsion, that is, fixations at distance r are rarer than expected for independently drawn fixations. See the online article for the color version of this figure.

Discussion

The key motivation for the current study was to apply the likelihood approach to the evaluation of dynamical cognitive models and, in particular, for model parameter estimation and model comparison. Dynamical cognitive models are formulated by evolution equations (temporally discrete or continuous) and evaluated against time-ordered data (time series). As a specific example, we investigated the problem of dynamical scanpath models, where the dynamical model determines the probability $\pi(x, t)$ to select a saccade target position x at time t. In the SceneWalk model (Engbert et al., 2015), this probability is computed from activation fields at any point in time. Thus, we can compute the corresponding probability for a fixation and force the model to generate the gaze shift to the new fixation position. This procedure of direct computation of the likelihood will work for the broad class of dynamical models that generate continuous-time activations for the prediction of discrete behavioral events (Erlhagen & Schöner, 2002).

For the interpretation, we normalized the likelihood with respect to the number of fixations in a given dataset to obtain a measure that is independent of the size (length) of the fixation sequence. Furthermore, we suggested to compare the likelihood to the likelihood obtained from a uniform distribution to get a measure which is independent of grid and image sizes. For simpler, nondynamical models this comparison to chance performance is a standard procedure. Additional nondynamical models were used to generate likelihoods to compare to the dynamical model. Such nondynamical density models (e.g., the central fixation bias, Tatler, 2007) represented a convenient statistical baseline for our computations. Finally, we investigated two variants of the SceneWalk model to show that the likelihood can be applied as a powerful tool to distinguish different dynamical models with highly specific assumptions.

The likelihood as a global measure of model performance can be used as a tool for the estimation of model parameters. Fitting models based on the maximum likelihood concept has a long tradition in statistics and some clear advantages over other parameter fitting procedures, including mathematical proofs for the convergence and sufficiency of the parameter estimate. A practical advantage is that the likelihood is a scalar value, which does not rely on simulating complex discriminating statistics. Additionally, model fitting based on the likelihood is the starting point for Bayesian inference about parameter values, which provides new insights to other parameters that could explain the data and, thus, statistical comparisons on whether the parameters differ between data sets or conditions.

For the SceneWalk model (Engbert et al., 2015), we computed parameter values using MLE and sampled the posterior for Bayesian parameter estimation. This parameter estimation technique allowed us to fit all the parameters of the model, which was impossible in the original publication. The parameters found by optimizing the likelihood reproduce all the statistics the original publication reported, whereas the parameters from the original publication perform significantly worse in terms of likelihoods. Additionally, we computed a full posterior probability over the parameters that informs about which parameters are constrained by the data well and which parameters are not constrained by the data.

Furthermore, the likelihood-based evaluation helped us to improve the original model. Using a hierarchical model, we found that the known differences between participants in their average saccade length (Castelhano & Henderson, 2008) could be fit well, by allowing the size of the attention window and the size of the inhibition to vary between participants. Furthermore, likelihood based comparisons between models allowed us to show that the dynamics and the inhibition both improve model predictions. And additionally, we could differentiate different variants how the excitatory and inhibitory maps are combined. For experimentally motivated statistics, these specific model variants made very similar predictions. Among the models analyzed here, a divisive inhibition model with a fixed numerator exponent λ seems to fit the data best—and even better than the original SceneWalk model.

With the SceneWalk model, we focus on fixation locations and take fixation durations as given (or a random process with given mean and variance). This is, however, not necessarily a restriction of the likelihood approach. Models which compute probabilities for fixation durations (Nuthmann, Smith, Engbert, & Henderson, 2010; Trukenbrod & Engbert, 2014) or for both the durations and locations of fixations (Tatler, Brockmole, & Carpenter, 2017) could be fit and evaluated using the same techniques we present here for locations only. There are recent studies on fixation durations for scene viewing (e.g., Laubrock, Cajar, & Engbert, 2013). Furthermore, the prediction of fixation durations is a main aim for models of eye movements during reading (Reichle, Rayner, & Pollatsek, 2003; Engbert, Nuthmann, Richter, & Kliegl, 2005).

In this article, we used relatively simple gradient free optimization algorithms and the Metropolis Hastings algorithm for their conceptual simplicity, which eased the presentation. However, there might be more efficient algorithms for solving the optimization and sampling problems in the SceneWalk model and certainly different algorithms will be best or easiest to implement for different models. Also, the optimizations and samplings for complex models may take hours, days or even months of computation time. Thus efficiency is important as it may make the difference whether an analysis is feasible with given computational resources or not. Consequently, it can be worthwhile to invest some time to try different optimization algorithms including global optimization algorithms, when local minima are a problem. Similarly, there is broad literature on how to (adaptively) tune MCMC-algorithms (e.g., Andrieu & Thoms, 2008; Gelman et al., 1996; Haario, Laine, Mira, & Saksman, 2006; Haario, Saksman, & Tamminen, 2001; Roberts & Rosenthal, 2009) and efficient sampling algorithms (Brooks et al., 2011; Robert & Casella, 2009, 2013).

An especially large step in efficiency for both optimization and sampling can be made if a gradient of the likelihood can be calculated with reasonable efficiency. For optimization highly efficient gradient based algorithms, that is, quasi-Newton methods like the BFGS algorithm are available. The original gradient based sampling algorithm is the Hamiltonian Monte Carlo (HMC) method introduced by Duane, Kennedy, Pendleton, and Roweth (1987) (see Neal, 2011, for an introduction). By now there are many variants of HMC available, including adaptive methods like the no-u-turn sampler (Hoffman & Gelman, 2014), which works behind Stan (Carpenter et al., 2017), one of the most recent general purpose samplers. These samplers contain automatic differentiation tools, which remove the necessity to code a gradient computation by hand. Also independent tools to compute derivatives automatically are able to differentiate virtually any computable function (Abadi et al., 2015; Theano Development Team, 2016), which allows computation of a derivative for many models.

As a next step the likelihood evaluation permits comparisons between different models. To avoid overfitting, we carried out such comparisons using cross-validation. Here, the SceneWalk model (Engbert et al., 2015) was compared with a statistical model of the central fixation bias and to a model that sampled fixation positions from the empirical saliency map. We found that the SceneWalk model outperforms the empirical saliency model by $0.75\frac{\text{bit}}{\text{fix}}$, which highlights the importance of incorporating influences of previous fixations into predictions for upcoming saccade targets. Consequently, a saliency model alone is not a good model for scanpaths, no matter how closely it matches the fixation density.

As the likelihood is a relative measure, it is necessary to check whether the fitted model is reasonably good in terms of absolute measures. For the SceneWalk model, we demonstrated the adequacy by comparing different summary statistics computed on model predictions to the corresponding statistics obtained from experimental data. We found that the model reproduced the fixation density, saccade length distribution, and the pair correlation function with parameters computed via MLE.

For scanpath models in eye-movement research, the likelihood approach to parameter estimation and model comparison is most interesting as there is no general consensus on a metric for comparing models so far (Le Meur & Baccino, 2013; Pitt et al., 2002). Instead, many statistics on specific aspects of scanpaths were proposed, which allow judgments whether a given model shows some specific effects or not. However, a global account of how adequately the model fits the experimental data is currently lacking. We demonstrated that such global measures could be provided by the likelihood approach.

In the likelihood approach, any scanpath observed in humans must have a probability larger than zero under the model, as the likelihood vanishes otherwise, indicating only that the model cannot explain the data. A second constraint on the model is that the likelihood can be computed. As we showed earlier, it is sufficient to be able to numerically generate the probability for the next fixation given the previous ones. This is not a strong constraint as most eye movement models on natural scenes even explicitly represent a probability map for the next fixation (Le Meur & Liu, 2015; Zelinsky, 2008; Zelinsky, Adeli, Peng, & Samaras, 2013).

We believe that model evaluations based on the likelihood are promising for many other psychological models. Indeed, for some models the evaluation is already routinely done using likelihoods, for example for receiver operating curves (Ogilvie & Creelman, 1968), diffusion models (Ratcliff & Tuerlinckx, 2002), psychometric functions (Wichmann & Hill, 2001), and recently for saliency models and fixations on static images (Barthelmé et al., 2013; Kümmerer et al., 2015).

One favorable aspect of the SceneWalk model is that it is deterministic—there is only a single way for the model to produce time-dependent activation maps for a given sequence of fixations. If there were multiple possible internal states compatible with the observed data, then the computation of the likelihood would require an integration over all possible internal states. Such integration could render evaluations of the likelihood function less effective or even impossible for other models. For such complex models with many possible internal states and large data sets efficient computational techniques for combined state and parameter estimation have been developed in particular in the field of data assimilation (Law, Stuart, & Zygalakis, 2015; Reich & Cotter, 2015). Furthermore, processing time-ordered data sets leads naturally to the consideration of sequential Monte Carlo methods (Doucet, de Freitas, & Gordon, 2001; Chopin, Jacob, & Papaspiliopoulos, 2013), to bring computational demands into a manageable range.

For some model classes computation of the likelihood might be too time consuming or the likelihood function too complex for further handling. However, even for such models, mathematically well founded approximations to the likelihood methods were proposed: *Pseudolikelihood* methods compute an approximation to the likelihood (Wood, 2010, e.g.). Alternatively, *pseudomarginal Monte Carlo methods* (Beaumont, 2003; Andrieu & Roberts, 2009) can be used which, while involving approximations, can be shown to provide consistent estimates. Here one could also consider replacing the likelihood by an appropriate scoring function (Gneiting, Balabdaoui, & Raftery, 2007) which provides an alternative metric to rank models in an objective manner. Moreover, approximate Bayesian computation allows an approximation to full Bayesian inference without a likelihood (Barthelmé & Chopin, 2011, 2014; Turner & Van Zandt, 2012; Wilkinson, 2013). These methods preserve some of the benefits of the likelihood approach to parameter estimation and model analysis and can even be used to do model selection. For dynamical models this is discussed for example by Toni, Welch, Strelkowa, Ipsen, and Stumpf (2009).

Conclusion

We proposed and studied a likelihood approach for the evaluation of a dynamical cognitive model for the control of saccadic eye movements. The likelihood can be used for parameter estimation and model comparisons as it makes the full range of statistics available, from MLE through Bayesian estimation and hierarchical models to proper model comparisons. Compared with nondynamical models, the dynamical model generated a significant increase in predictive power by introducing sequential dependencies. Our approach is a promising tool for the evaluation of dynamical models that predict sequences of discrete behavior (e.g., fixation position, movement onsets) in general and for human scanpaths in particular.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from www.tensorflow.org

Akaike, H. (1974). A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers Transactions on Automatic Control, 19,* 716–723.

Andrieu, C., & Roberts, G. (2009). The pseudo-marginal approach for efficient Monte-Carlo computations. *The Annals of Statistics, 37,* 697–725.

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing, 18,* 343–373.

Ash, R. B. (1990). *Information theory*. New York, NY: Dover Publications.

Barthelmé, S., & Chopin, N. (2011). ABC-EP: Expectation propagation for likelihood-free Bayesian computation. In *Proceedings of the 28th International conference on machine learning* (pp. 289–296). New York, NY: ACM.

Barthelmé, S., & Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association, 109,* 315–333.

Barthelmé, S., Trukenbrod, H., Engbert, R., & Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision, 13*(12):1, 1–34.

Beaumont, M. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics, 164,* 1139–1160.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: Ideas and concepts*. San Francisco, CA: Holden-Day.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence, 35,* 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2014). What/where to look next? Modeling top-down visual attention in complex interactive environments. *Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics: Systems, 44,* 523–538.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7,* 434–455.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: CRC Press.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44,* 108–132.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art*. Chicago, IL: University Chicago Press.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2016). *MIT saliency benchmark*. Retrieved from http://saliency.mit.edu/

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13,* 51–62.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician, 46,* 167–174.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 62,* 1–14.

Chopin, N., Jacob, P., & Papaspiliopoulos, O. (2013). SMS^2: An efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society B, 75,* 397–426.

Conover, W. J., & Conover, W. (1980). *Practical nonparametric statistics*. New York, NY: Wiley.

Cox, D. R. (2006). *Principles of statistical inference*. New York, NY: Cambridge University Press.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR), 27,* 326–327.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York, NY: Springer-Verlag.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195,* 216–222.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). Swift: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777–813.

Engbert, R., Trukenbrod, H. A., Barthelmé, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision, 15*(1):14, 1–17.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review, 109,* 545–572.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Methodological, 56*, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (2nd ed.). New York, NY: Taylor & Francis.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing, 24,* 997–1016.

Gelman, A., Roberts, G., & Gilks, W. (1996). Efficient Metropolis jumping hules. *Bayesian Statistics, 5,* 599–608.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–511.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology, 56,* 1–12.

Gneiting, T., Balabdaoui, F., & Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B, 69,* 243–268.

Golberg, D. E. (1989). Genetic algorithms and Machine Learning. *Machine Learning, 3*(2–3), 95–99.

Haario, H., Laine, M., Mira, A., & Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing, 16,* 339–354.

Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli, 7,* 223.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51,* 347–356.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Hays, W. L. (1994). *Statistics*. Independence, KY: Wadsworth Publishing.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504.

Hennig, P., & Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research, 13,* 1809–1837.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15,* 1593–1623.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, MI: University of Michigan Press.

Houck, C. R., Joines, J., & Kay, M. G. (1995). *A genetic algorithm for function optimization: A Matlab implementation, Technical Report NCSU-IE-TR-(95-09)*. North Carolina State University, Raleigh, NC. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.4413&rep=rep1&type=pdf

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2,* 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis & Machine Intelligence, 11,* 1254–1259.

Jackson, E. A. (1992). *Perspectives of nonlinear dynamics*. New York, NY: Cambridge University Press.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218). New York, NY: ACM.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. New York, NY: Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 186,* 453–461.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python [Computer software]. Retrieved from http://www.scipy.org/

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization, 13,* 455–492.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision* (pp. 2106–2113). New York, NY: IEEE.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision, 9*(5):7, 1–15.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics, 34*(5–6), 975–986.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated Annealing. *Science, 220,* 671–680.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4,* 138–147.

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, 10,* 346–352.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences, USA, 112,* 16054–16059.

Laubrock, J., Cajar, A., & Engbert, R. (2013). Control of fixation duration during scene viewing by interaction of foveal and peripheral processing. *Journal of Vision, 13*(12):11, 1–20.

Law, K., Stuart, A., & Zygalakis, K. (2015). *Data assimilation*. New York, NY: Springer-Verlag.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 45,* 251–266.

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research, 116,* 152–164.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10,* 707–710.

Marin, J.-M., & Robert, C. (2007). *Bayesian core: A practical approach to computational Bayesian statistics*. New York, NY: Springer Science & Business Media.

MATLAB. (2016). *Optimization toolbox*. Natick, MA: The MathWorks.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21,* 1087–1092.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90–100.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Boca Raton, FL: CRC Press.

Nicholls, J. G., Martin, A. R., Fuchs, P. A., Brown, D. A., Diamond, M. E., & Weisblat, D. A. (2012). *From neuron to brain* (5 ed.). Sunderland, MA: Sinauer Associates.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review, 117,* 382–405.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481.

Reich, S., & Cotter, C. (2015). *Probabilistic forecasting and Bayesian data assimilation*. New York, NY: Cambridge University Press.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences, 26,* 445–476.

Robert, C., & Casella, G. (2009). *Introducing Monte Carlo Methods with R*. New York, NY: Springer.

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. New York, NY: Springer Science & Business Media.

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics, 18,* 349–367.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Smith, T. J., & Henderson, J. M. (2009). Facilitation of return during scene viewing. *Visual Cognition, 17*(6–7), 1083–1108.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64,* 583–639.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review, 124,* 267–300.

Theano Development Team. (2016). *Theano: A Python framework for fast computation of mathematical expressions*. Retrieved from http://arxiv.org/abs/1605.02688

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface, 6,* 187–202.

Trukenbrod, H. A., & Engbert, R. (2014). ICAT: A computational model for the adaptive control of fixation durations. *Psychonomic Bulletin & Review, 21,* 907–934.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56,* 69–85.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21,* 615–628.

Villemonteix, J., Vazquez, E., & Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization, 44,* 509–534.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language, 65,* 109–127.

von Helmholtz, H. (1924). *Treatise on physiological optics (P. C. Southall, Trans.)*. Ithaca, NY: The Optical Society of America.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11,* 3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research, 14,* 867–897.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313.

Wilkinson, R. D. (2013). Approximate bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology, 12,* 129–141.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature, 466,* 1102–1104.

Yarbus, A. L. (1967). *Eye movements during perception of complex objects*. New York, NY: Springer.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115,* 787–835.

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 368,* 20130058.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44,* 41–61.

Received June 8, 2016
Revision received January 23, 2017
Accepted February 26, 2017 ■