Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Using computational theory to constrain statistical models of neural data

Scott W Linderman[1] and Samuel J Gershman[2]

Computational neuroscience is, to first order, dominated by two approaches: the 'bottom-up' approach, which searches for statistical patterns in large-scale neural recordings, and the 'top-down' approach, which begins with a theory of computation and considers plausible neural implementations. While this division is not clear-cut, we argue that these approaches should be much more intimately linked. From a Bayesian perspective, computational theories provide constrained prior distributions on neural data — albeit highly sophisticated ones. By connecting theory to observation via a probabilistic model, we provide the link necessary to test, evaluate, and revise our theories in a data-driven and statistically rigorous fashion. This review highlights examples of this theory-driven pipeline for neural data analysis in recent literature and illustrates it with a worked example based on the temporal difference learning model of dopamine.

Addresses
[1] Department of Statistics, Columbia University, United States
[2] Department of Psychology and Center for Brain Science, Harvard University, United States

Corresponding author: Gershman, Samuel J (gershman@fas.harvard.edu)

Current Opinion in Neurobiology 2017, **46**:14–24

This review comes from a themed issue on **Computational neuroscience**

Edited by **Adrienne Fairhall** and **Christian Machens**

For a complete overview see the Issue and the Editorial

Available online 18th July 2017

http://dx.doi.org/10.1016/j.conb.2017.06.004

Introduction

The statistical toolbox for neuroscience has been steadily growing in sophistication — relaxing restrictive assumptions, increasing expressiveness, and enhancing computational efficiency. These advances have enabled a recent blossoming of 'data-driven' approaches to neuroscience, which aim to provide insight into neural mechanisms without testing specific computational theories. Data-driven approaches are appealing, at least in principle, for several reasons: they do not require the scientist to explicitly specify a set of hypotheses, they are unprejudiced by the scientist's theoretical dispositions, and they avoid the problem that many computational theories are difficult to fit to real data.

In this paper, we argue that such faith in data-driven approaches is misplaced. Far from escaping the explicit specification of hypotheses, any statistical model of neural data inevitably makes assumptions about the structure of the data, and there is no principled distinction between statistical assumptions and scientific hypotheses. (Admittedly, a purely data-driven approach is something of a straw-man, but we pursue this line of argument for pedagogical purposes.) A corollary of this point is that theoretical dispositions are inescapable: it is impossible to specify a statistical model without making assumptions. The question then becomes what assumptions to make. We argue that these assumptions should be derived from computational theories (which provide strong and principled constraints), coupled with flexible statistical parametrizations that compensate for inaccuracy and underspecification of the theories.

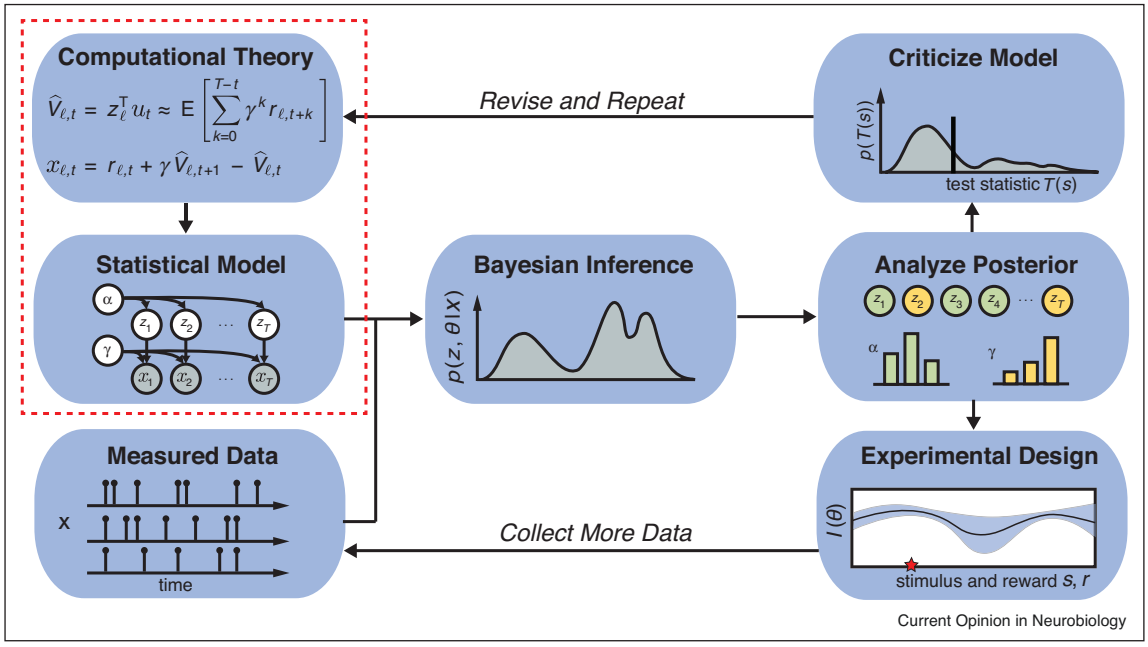
This is not a radically novel perspective; indeed, these ideas date back decades in classical statistics [1–3] and find their roots in the works of Popper [4]. This combination of top-down and bottom-up modeling is becoming more common in neursocience — as recently reviewed by Durstewitz *et al.* [5] — though it is still the exception rather than the rule. Our goal is to espouse this type of approach, to simplify it by breaking it down into well defined blocks, to illustrate it through examples, and to highlight some of the recent work in the Bayesian machine learning and statistics communities that could aid in various steps of the process.

We illustrate this approach with a worked example, using a paradigmatic neurocomputational theory: the temporal difference learning model of dopamine. We show how the computational theory can be augmented with modern statistical tools to produce a powerful data analysis methodology. This approach generates a more complete and flexible specification of the theory. Moreover, we show that this approach offers insights into the mechanisms underlying neural data that are inaccessible to purely data-driven approaches.

A theory-driven pipeline for neural data analysis

Neural data analysis is an iterative process that begins with a data set and an idea of the underlying processes that shaped it. The first step, and arguably the most important one, is to turn that idea into a model. With a

Figure 1



A theory-driven pipeline for neural data analysis based on 'Box's Loop' [3,6]. This review illustrates many examples of translating theory into statistical model (red box). The benefits are many. Given a model, we may leverage a powerful toolbox of statistical techniques for inference, model criticism, and experimental design. Equally important, theory constrains the space of models and provides a critical lens through which to interpret the posterior. We will discuss advances in each stage of this pipeline.

model in hand, we fit it to the data and investigate the learned parameters, searching for patterns that shed new light on the system under study. But the process does not end here; we then interrogate our model, see where it captures the data well and where it fails, and use these criticisms to suggest model enhancements or subsequent experiments. Thus, model criticism leads to a new model and another iteration of the process.

Statisticians have formalized and automated many pieces of this pipeline: models are joint distributions over data, latent variables and parameters; 'fitting' is performed by posterior inference; criticism is carried out with statistical tests; and optimal experimental design suggests what experiment to run next. This cyclic process of probabilistic modeling, inference, and statistical criticism is known as 'Box's loop' [1–3,6], and later sections of this review will discuss many recent advances in each stage of the pipeline (Figure 1).

Still, the art of carving a tractable class of models from the infinite space of possibilities remains the province of the practitioner. It is here that computational theory can play a vital role, since theories suggest what structure and patterns may exist in the data. In doing so, theories constrain the class of models and make it easier to search, and provide a lens through which to interpret model parameters. These benefits are reciprocated: once a theory has been translated into a probabilistic model, a vast statistical toolbox can be harnessed to test and refine it in light of data.

Theory-driven statistical models are the norm in many fields, most notably in physics, where strong quantitative predictions can be derived from first principles. For example, the discovery of the Higgs boson relied on statistical tests based on predictions of the standard model [7]. Perhaps it is unsurprising, then, that some of the best examples of theory-driven statistical analyses in neuroscience arise from detailed, biophysical models of single cells. For example, Huys and Paninski [8] use the Hodgkin–Huxley model to derive a probabilistic model for noisy membrane potential recordings. The conductances of various ion channels are free parameters of their model, and the time-varying channel activations are their latent states. Given the membrane potential, their goal is to infer the conductances, integrating over possible activation states. The highly nonlinear nature of the Hodgkin–Huxley dynamics and the potentially large number of different channel types present a formidable challenge, but biophysical constraints limit the space of feasible parameters. In recent work, these methods have been extended to data in which only spike trains are observed [9], which present an even greater challenge.

Many models in neuroscience are phenomenological rather than mechanistic in nature. One step up from biophysical models are firing rate models like autoregressive Poisson models, a form of generalized linear model (GLM) [10–12]. Recent work has extended these classical models to make them more flexible [13], more biophysically inspired [14], and more interpretable [15•]. While the GLM omits many mechanistic details, in fully-observed networks its weights can be roughly interpreted as synaptic strengths [16,17]. However, the weights of the standard GLM are static, even though synaptic plasticity may be at work in many neural recordings. While the space of all possible dynamic GLM's is intractably large, theories of synaptic plasticity place strong constraints on how synaptic weights evolve over time in response to preceding activity. A number of authors have leveraged these constraints to develop theory-driven GLM's with time-varying weights and have shown how alternative models of synaptic plasticity can be compared on the basis of their fit to spike train data [18–20].

This approach extends to computational theories as well, and is exemplified in the work of Latimer *et al.* [21••]. The authors reconsider the long-standing theory of evidence accumulation in lateral intraparietal (LIP) cortex [22], and ask whether patterns that emerge in trial-averaged data are borne out in individual trials. Specifically, do the firing rates of neurons in LIP slowly ramp as evidence is accumulated, or do they exhibit a discrete jump in firing rate? Theory suggests the former, whereas the latter would indicate that LIP may not be the site of integration (with the caveat that integration might still be implemented in LIP at the population level without all neurons behaving like integrators). Critically, both theories would yield the appearance of a ramp in trial-averaged firing rate. Latimer *et al.* [21••] formulate both theories as probabilistic models for single trial data, fit these models with Bayesian inference, compare them on the basis of the marginal likelihood of the data, and find that a large fraction of neurons are better explained by the discrete jump model. This provides statistical evidence with which to assess and reevaluate canonical theory. Indeed, this work has prompted further assessments of their modeling assumptions and the validity of their conclusions [23] — a prime example of Box's loop in action post-publication.

Integrative approaches to computational theory and statistical analysis have also been pursued in higher-level cognition. Detre and colleagues [24] used Bayesian inference to identify a nonmonotonic relationship between memory activation (as measured by functional MRI) and subsequent memory, as predicted by a competition-dependent theory of episodic memory [25]. The same analytical approach was used to identify other nonmonotonic effects of retrieval strength on memory [26,27].

The aforementioned examples stand in contrast to many dimensionality reduction methods like PCA, tSNE [28], and others [29], and differ as well from general-purpose state space models [30–32] and recurrent neural network models [e.g. 33••] for neural data. Such methods start with very weak assumptions — linear embeddings or low-dimensional dynamics — and, in this sense, allow the data to speak freely. Thus, they are invaluable exploratory tools. However, in the absence of a theory, the inferred low-dimensional states and projections require careful interpretation. In many cases, theories correspond to special cases of these general-purpose models, and thus help address issues of interpretability.

The landscape of neural data-analysis is not as strictly divided into top-down and bottom-up approaches as the preceding discussion may suggest. Indeed, many models fall somewhere in the middle, incorporating aspects of theory while allowing flexibility in aspects that are less certain. Wiltschko *et al.* [34] strike such a balance in their model for depth videos of freely behaving mice. Starting with the classic ethological theory that behavior is composed of a sequence of discrete, reusable units, or 'syllables,' the authors propose an autoregressive hidden Markov model to discover these syllables from raw data. Since the number of syllables is not known *a priori*, the authors use a Bayesian nonparametric prior distribution [35] to determine the number of states in a data-driven manner.

These works exhibit a diverse array of 'theory-driven' neural data analyses, but the best way to understand this pipeline is through an example.

A worked example

There is no single recipe for translating computational theories into probabilistic models of data, but the conversion necessarily involves answering a few basic questions. Which theoretical variables and parameters are observed and which are latent? How are they encoded by the neural system under study? How do these variables evolve over time? What are the sources of noise in the system and in the measurements? The answers to these questions inform statistical models of data that in turn define distributions of likely patterns of neural activity. We will illustrate this translation with a simple worked example.[1]

Temporal difference (TD) learning [37] is a classical algorithm by which agents, over the course of many trials, learn to use sensory cues to predict the discounted sum of future rewards. Assume that there are L trials, each lasting T time steps. On trial ℓ, the agent receives a sequence of stimuli, which are stored and encoded as vectors, $u_{\ell,t}$, and a corresponding sequence of rewards, $r_{\ell,a}, \ldots, r_{\ell,T}$, most of

[1] Code to run this example and reproduce Figures 2 and 3 is available at https://github.com/slinderman/tdlds.

which may be zero. In a classical conditioning experiment, the stimulus may be a light at time t followed by a reward some number of time steps in the future, and $u_{\ell,t}$ may encode, for example, the number of time steps since the bell was heard. The agent then uses this encoding to compute a *value function* for the given trial and time step,

$$V_{\ell,t} = z_\ell^T u_{\ell,t}. \tag{1}$$

In reinforcement learning, the value is the total amount of future reward to be expected after receiving input $u_{\ell,t}$. However, according to the theory, the reward is discounted by how long one must wait before receiving it. For example, a reward k time steps into the future is down-weighted by a factor of γ^k, where $\gamma \in [0, 1]$ is the *discount factor*. The agent's goal is to adjust the weights[2] of its value function, z_ℓ, such that the value function approximates this discounted sum of expected future rewards,

$$V_{\ell,t} \approx V_{\ell,t} = \mathbb{E}\left[\sum_{k=0}^{T-t} \gamma^k r_{\ell,t+k}\right]. \tag{2}$$

If the environment is a Markov decision process, the target value function can be written recursively as $V_{\ell,t} = \mathbb{E}[r_{\ell,t} + \gamma V_{\ell,t+1}]$. When the value function equals the cumulative discounted reward, the *reward prediction error*,

$$x_{\ell,t} = r_{\ell,t} + \gamma V_{\ell,t+1} - V_{\ell,t}, \tag{3}$$

will equal zero. Intuitively, the reward prediction error provides an instantaneous estimate of how well the value function predicts the received reward. Thus, to improve its value function, the agent should adjust its weights to reduce this error. Indeed, this is accomplished by the simple learning rule,

$$z_{\ell+1} = z_\ell + \alpha_\ell \sum_{t=1}^{T} x_{\ell,t} u_{\ell,t}, \tag{4}$$

which can be seen as a form of stochastic gradient descent on the (squared) reward prediction error with learning rates $\alpha_1, \ldots, \alpha_L$. In the following experiments, we will consider two learning schedules: a power-law schedule, $\alpha_\ell = (\ell+1)^{-\tau}$, and a constant schedule, $\alpha_\ell \equiv \tau$. In both cases, assume $\tau \in [0, 1]$.

Schultz *et al.* [38] found that the firing rates of dopaminergic neurons in the ventral tegmental area (VTA) mimic the reward prediction errors essential to the TD-learning algorithm. Moreover, it is hypothesized that cortex represents the stimulus, striatum represents the value function estimate, and VTA activity modulates plasticity of synapses from cortex to striatum [39]. Still, many important questions remain, like how learning schedules, which affect this plasticity, vary from trial to trial in real neural circuits. As a didactic exercise, we will we use the TD learning theory to construct a probabilistic model for neural data, and use that model to compare between different learning schedules in a statistically rigorous manner.

Suppose that we have access to simultaneous noisy recordings of a VTA neuron and an upstream population of N cortical neurons. As has been hypothesized, we will assume the VTA neuron encodes reward prediction error, $x_{\ell,t}$, and the cortical neurons carry the stimulus encoding, $u_{\ell,t}$. Moreover, assume we know the reward signal, $r_{\ell,t}$. These assumptions may not be warranted in practice, and they must be tested, as we discuss below. According to the TD learning theory, the cortical and VTA signals are related via a value function, which is determined by an unobserved and dynamic set of weights at each trial. In other words, the theory implies that the reward prediction errors follow a latent state space model whose hidden states are the weights, z_ℓ, and whose parameters vary from trial to trial according to the cortical inputs, rewards, and prediction errors. If we assume Gaussian noise in the weight updates and observations, the theory implies that the VTA activity follows a Gaussian linear dynamical system (LDS) with non-stationary parameters.

To see this equivalence, we rewrite the TD learning updates in standard state space notation:

$$z_{\ell+1} \sim \mathcal{N}(A_\ell z_\ell + b_\ell, \epsilon I), \tag{5}$$

$$x_\ell \sim \mathcal{N}(C_\ell z_\ell + d_\ell, \sigma I). \tag{6}$$

Here, the latent states are the weights, $z_\ell \in \mathcal{R}^N$, and their dynamics are determined by $A_\ell = I$ and $b_\ell = \alpha_\ell \sum_{t=1}^{T} x_{\ell,t} u_{\ell,t}$. That is, the weights follow a random walk biased by the learning rate, error signal, and inputs. The emissions are vectors of observed VTA activity, $x_\ell = [x_{\ell,1}, \ldots, x_{\ell,T-1}]$, and they are determined by the matrix $C_\ell = [c_{\ell,1}^\top; \ldots; c_{\ell,T-1}^\top]$, where $c_{\ell,t} = \gamma u_{\ell,t+1} - u_{\ell,t}$, and by the bias vector $d_\ell = [d_{\ell,1}, \ldots, d_{\ell,T-1}]$, where $d_{\ell,t} = r_{\ell,t+1}$. Note that both the dynamics and emission parameters are non-stationary; that is, they vary from trial to trial. The noise in the weight updates is governed by ϵ, and the noise in the observations is governed by σ. Referring back to Eqns 1–4, we see that the exact TD learning model is recovered in the noise-free limit. The free parameters are $\theta = (\tau, \gamma, \epsilon, \sigma)$ — the learning rate parameters, discount factor, and noise variances.

We call this constrained model a temporal difference LDS (TD-LDS). Importantly, by translating the TD learning theory into a constrained Gaussian LDS, we have reduced it to an essentially solved model with very mature estimation and interpretation procedures [40]. In the next section we will show how to infer the states and parameters of the TD-LDS from data.

What assumptions did we make in deriving the TD-LDS? First, we assumed Gaussian noise in both the observed

[2] We denote the weights by z instead of something more traditional, like w, since this will highlight the connection to state space models.

reward prediction errors and the weight dynamics. If we observed spike counts instead, the resulting model would be more akin to a Poisson linear dynamical system (PLDS) [30,31]. If we had assumed a nonlinear model for the value function, that is, $V_{\ell,t} = f(z_\ell, u_{\ell,t})$, then both the dynamics and observation models would be nonlinear in z_ℓ, which would necessitate more sophisticated inference procedures. We will only consider the linear Gaussian case in this didactic example.

Bayesian inference

Bayesian inference algorithms take as input the observed data, x, and a probabilistic model, $p(x, z, \theta)$, and output the posterior distribution over the latent variables and parameters of the model, $p(z, \theta \mid x)$. By Bayes' rule, this posterior distribution is given by,

$$p(z, \theta|x) = \frac{p(x|z, \theta)p(z|\theta)p(\theta)}{p(x)} = \frac{p(x|z, \theta)p(z|\theta)p(\theta)}{\int p(x|z, \theta)p(z|\theta)p(\theta)\, \mathrm{d}z\, \mathrm{d}\theta}. \quad (7)$$

With this posterior distribution in hand, we can answer a host of scientific questions. We can estimate the posterior mean and mode (the maximum *a posteriori* estimate), and we can provide Bayesian credible intervals by computing the quantiles of the posterior distribution. Moreover, we can predict what future data would look like with the *posterior predictive* distribution,

$$p(x^*|x) = \int p(x^*|z^*, \theta)p(z^*|\theta)p(\theta, z|x)\, \mathrm{d}z^*\, \mathrm{d}z\, \mathrm{d}\theta. \quad (8)$$

which integrates over the space of parameters and latent variables, weighting them by their posterior probability given the data seen thus far. As we will show below, these functions of the posterior distribution provide principled means of comparing and checking models.

Unfortunately, the normalizing constant on the right-hand side of Bayes' rule, $p(x)$, also known as the *marginal likelihood*, requires an integral over all possible parameters. This integral is intractable for all but the simplest models, so in practice we must resort to approximate techniques like Markov chain Monte Carlo (MCMC) [41] or variational inference [42,43]. MCMC algorithms approximate the posterior distribution with a collection of samples collected by a Markov chain that randomly walks over the space of parameters. With a carefully tuned random walk, the stationary distribution of the Markov chain is equal to the desired posterior distribution so that, once the chain has converged, parameters are visited according to their posterior probability. In contrast, variational inference algorithms specify a family of 'simpler' distributions and search for the member of this family that best approximates the desired posterior. Thus, they convert an integration problem of computing the denominator of Bayes' rule into an optimization problem of searching over the variational family. Of course, both approaches present challenges — how to tell if a Markov chain has converged? How to select and search over a variational family and diagnose errors in the obtained approximation? — making Bayesian inference both an art and a science.

Fortunately for the practitioner, as probabilistic programming packages grow in sophistication, the nuances of approximate inference play a lesser role. Probabilistic programming languages like Anglican [44], Stan [45], Venture [46], and Edward [47] remove the burden of deriving and implementing an inference algorithm, and simply require the practitioner to specify their probabilistic model and supply their data. Under the hood, these packages automatically derive suitable MCMC or variational inference algorithms. In practice, some care must be taken to ensure these systems provide accurate inferences, and these tools still cannot compete with well-tuned, model-specific inference algorithms. However, they can dramatically accelerate the scientific process by enabling rapid iteration over models. Once a model has been selected, time may be invested in deriving bespoke inference algorithms for peak performance.

We have taken an intermediate approach to inference in our working example. After reducing TD learning theory to a canonical state space model, we leverage off-the-shelf inference algorithms for the latent states and develop model-specific updates only for the parameters. Specifically, given the discount factor and the learning schedule, the posterior distribution over latent states is found with a standard message passing algorithm [43]. Given a distribution over latent states, we estimate the most likely learning schedule parameters and discount factor with hand-derived updates. We alternate these two steps — updating the latent states and re-estimating the parameters — in our variational inference algorithm.

Figure 2 illustrates some of the results of our Bayesian inference algorithm. Panel (e) shows the posterior mean of the states, which in this model correspond to the weights of the value function. From the posterior distribution over weights, we derive the distribution over the value function, which is linear in the weights (c.f. 1). Panel (f) shows the true and inferred value function at early (blue), middle (red), and late (yellow) trials, along with the uncertainty under the posterior. Likewise, panel (g) shows the inferred learning rate under two different models: a model with constant rates and a model with rates that decay according to a power law (the true model in this case). Posterior visualizations like these play a critical role in the scientific process, providing views of the low-dimensional structure of complex data. However, these visualizations are only useful to the extent that the model captures meaningful structure. Panel (h) exemplifies this point: a standard LDS with the same latent

dimension as the TD-LDS provides a very good fit to the data, but its latent states look like pure noise. Without a theoretical structure with which to interpret this low-dimensional projection, the latent states are meaningless.

Bayesian inference is only one method of estimation, and it stands in contrast to other approaches like maximum likelihood and the method of moments. These could have been substituted in the center panel of Figure 1, but the model criticism and experimental design methods discussed below assume access to the posterior distribution. Avoiding statistical dogmatism, our view is simply that the posterior distribution of parameters and latent variables is often the object of interest, and this is the central object of study in the Bayesian approach. However, this requires a choice of prior distribution, which must be checked, just like the rest of the model, and it requires a challenging approximate computation, whose accuracy must also be assessed. The next section addresses the former; a number of previous works have addressed the latter [e.g. 48–51]. Finally, we note that posterior predictive checks discussed in the following section are essentially frequentist tests of Bayesian estimators, a pragmatic blend of approaches.

Model criticism and comparison

Bayesian inference is not the end of the scientific process, but rather an intermediate step in the iterative loop of hypothesizing, fitting, criticizing, and revising a model. Still, posterior inference provides a rigorous and quantifiable method of guiding model criticism and revision. Intuitively, if the model is a good match for the data, then samples from the fit model should 'look like' the observed data. *Posterior predictive checks* (PPC's) [3,52–54], which are essentially Bayesian goodness-of-fit tests, formalize this intuition in a statistically rigorous manner. Our presentation here parallels that of Blei [6].

PPCs compare the observed data to datasets sampled from the posterior predictive distribution 8 of the model. If the sampled data differs from the observed along important dimensions, the model fails the PPC. These 'important dimensions' are determined by the practitioner's choice of a test statistic, $T(x)$: a function that identifies a particular aspect of the data, x. For example, in our TD learning simulations, a salient characteristic is the propagation of error signal from the onset of reward to the presentation of the cue. Thus, a simple statistic is amplitude of the error signal in particular trials and time bins. The PPC is defined as the probability that the test statistic of sampled data exceeds that of observed data, $\text{PPC} = \Pr(T(x^*) > T(x) \mid x)$.

The choice of test statistic is left to the practitioner. Clearly, probabilistic modeling under computational constraints necessitates trade-offs and assumptions; no model is perfect. PPCs are a diagnostic tool for assessing whether the model recapitulates salient features of the data, as determined by the practitioner. In this sense, PPCs provide a targeted means of criticizing models, shining spotlights on the most important parts. Moreover, there is no limit to the number of PPCs that may be applied, and the marginal cost of estimating multiple PPCs is negligible since they can all be estimated using the same sampled data.

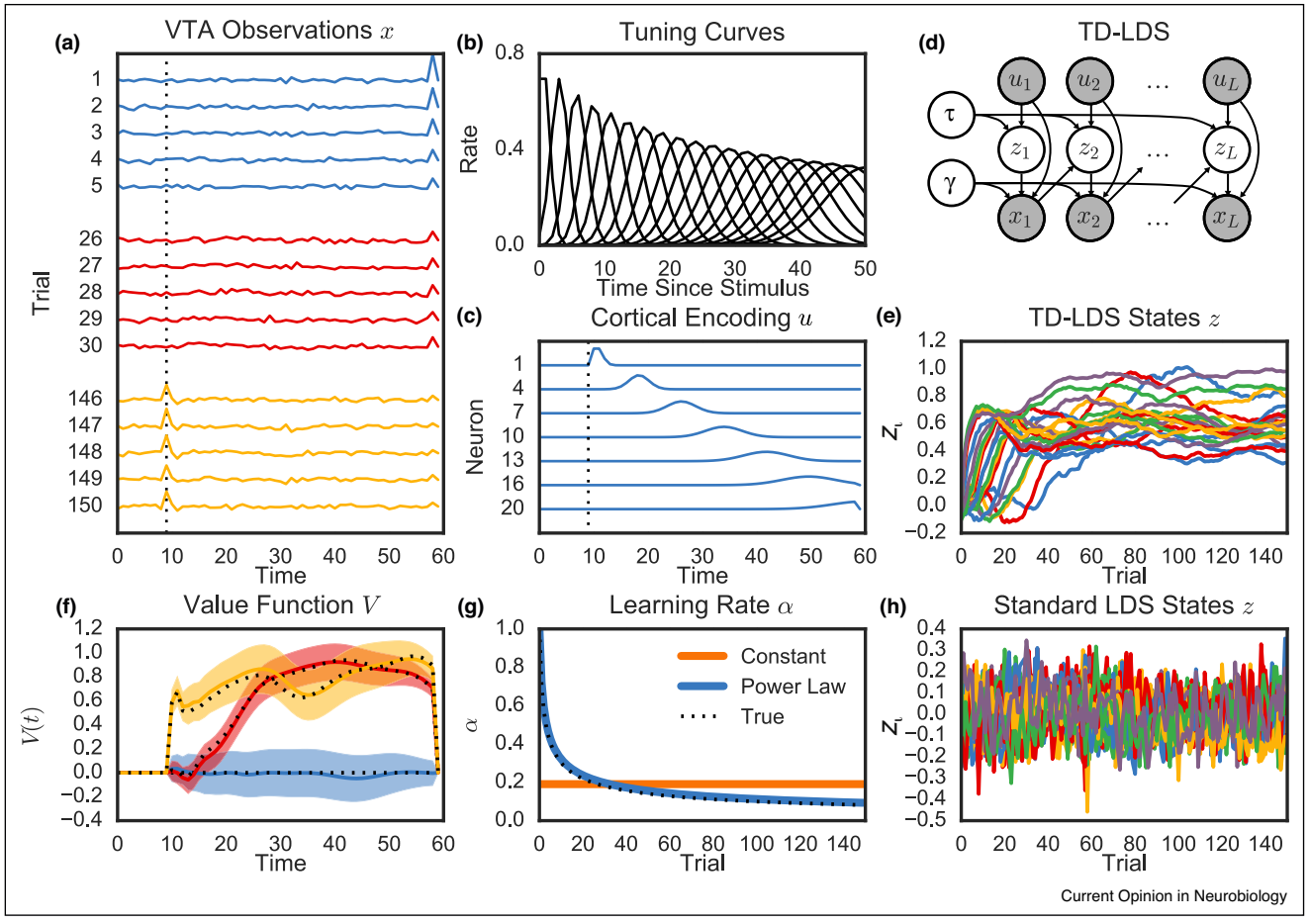
Figure 3 illustrates a very simple posterior predictive check for the TD learning model. Panels (a–c) show the observed data (black) and the quantiles of the posterior predictive distribution for the tenth trial, estimated with 1000 samples from the posterior predictive distributions. In this case, the true model uses a power law learning rate, and indeed this is the only model that consistently captures the data. The constant model overestimates the response to the reward (time 60) and the standard LDS incorrectly predicts a response at cue onset. We quantify this with PPC's for the simplest statistics, $T_{t,t'}(x) = x_{t,t'}$. Panels (d–f) show the PPCs for each trial and time bin. This reveals the delayed responses of the constant model in early trials, and the tendency of the standard LDS to predict a response at cue onset regardless of trial. Under the true model, these PPCs are uniformly distributed on [0, 1]. Panels (g–f) show that only the power law achieves this.

While PPCs, in absolute terms, how well the model fits the data, in some cases we seek a relative comparison of two models instead. For example, we often cascade models of increasing complexity — factor analysis is a special case of an LDS, which in turn is a special case of a switching LDS — and we need means of justifying this increased capacity. The most straightforward approach is to measure predictive likelihood on held-out data. A better model should assign higher posterior predictive probability, $p(x^* \mid x)$, to the held-out data. We see that the predictive probability 8 is an expectation with respect to the posterior. Since this is typically intractable, we estimate the predictive probability with samples from the approximate posterior.

This is by no means the only method of comparing models. In 'fully Bayesian' analyses, it is common to compare models on the basis of their marginal likelihood, $p(x)$ [55,56]. Recall that this is the denominator in Bayes' rule 7, and it is generally intractable. Variational methods provide a lower bound on this quantity, and Monte Carlo estimates like annealed importance sampling [57] can yield unbiased estimates of it. In general, however, marginal likelihood estimation is an active area of research [58–60].

Model criticism suggests not only new theories to test, but also new experiments to run. Specifically, we should choose an experiment that is most likely to reduce the uncertainty of the posterior. Equivalently, we should

Figure 2

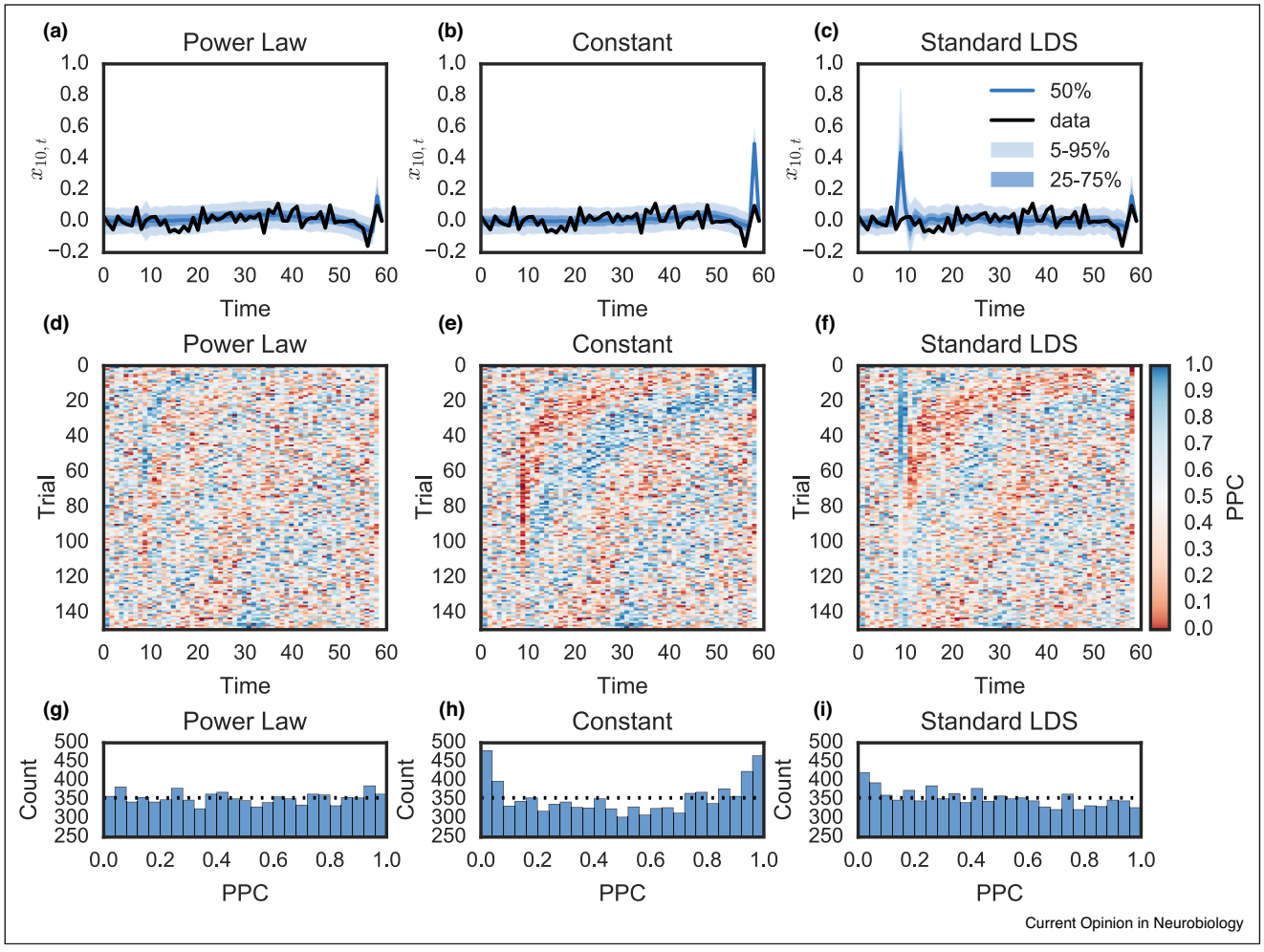


An illustrative example of using the theory of TD learning to constrain a probabilistic state space model for neural data. **(a)** Simulated example of a dopamine neuron encoding reward prediction error in VTA. Over many trials, the response shifts from the delivery of reward (at $t = 60$) to the onset of stimulus (at $t = 10$, dashed line). **(b)** Hypothetical cortical neurons encode time since stimulus onset with a set of temporal tuning curves, as has been suggested [36]. **(c)** Thus, on each trial, the cortical neurons exhibit a cascade of activity. **(d)** We use TD learning theory to constrain a state space model for the activity of cortex and VTA, whose graphical model is shown here (rewards omitted). The latent states are the weights relating cortical activity to an unobserved value function. **(e)** The posterior mean of the latent states of the TD learning state space model. Though not particularly insightful on their own, when combined with cortical activity, the weights determine the posterior distribution of the value function **(f)**. Colors correspond to trials 1, 30, and 150, as in (a). Dotted black line: ground truth. **(g)** We also learn the learning rate, α_t, under two different models: a constant model and a power-law decay model. **(h)** In contrast to the TD-LDS, fitting a standard LDS to the VTA activity yields accurate predictions, but its latent states are uninformative and do not correspond to weights of a value function.

perform the experiment that yields the maximal information gain in expectation. This intuition is the basis of Bayesian optimal experimental design [55,61–63] and is also the guiding principle underlying Bayesian optimization [64]. In our working example, these methods could suggest the combination of stimulus and reward patterns that would be most informative of the underlying learning rate. These methods have been proposed for sampling the voltage on dendritic trees in high-noise settings [65], as well as for designing training regimes for animals [66•].

Just as probabilistic programming languages and automated inference algorithms are relieving the burden of Bayesian inference, recent work has attempted to automate model criticism and model comparison. Automatic two-sample tests [67,68] search for test statistics that best discriminate between the observed data and a model's samples. In this sense, these approaches are similar to generative adversarial networks [69], which simultaneously train competing generator and discriminator networks. Likewise, automatic model composition methods [70,71] iteratively construct models, adding increasingly sophisticated structure to capture nuances of the data and comparing on the basis of marginal likelihood. While these advances have still not taken the human 'out of the loop,' recent work suggests that these approaches do indeed mimic the process by which humans learn the complex structure of data [72].

Figure 3



Model criticism using posterior predictive checks (PPCs). **(a–c)** PPC of the data on trial 10 for three models: the TD-LDS with a power-law learning schedule (i.e. the true model that generated the data); the TD-LDS with a constant learning rate; and a standard LDS. Blue line: posterior predictive median; blue shading: posterior predictive quantiles; black line: observed data. The constant learning rate fails the PPC because it generates a much larger prediction error at time $t = 59$. The standard LDS fails because it always predicts large signals at $t = 10$, regardless of trial. **(d–f)** A summary view of the PPC for all trials and time points. Color denotes the PPC value estimated from 1000 generated trajectories. Blue: model predictions larger than data; red: data larger than model predictions. Values close to zero or one indicate model mismatch. **(g–i)** A histogram of values in (d-f), respectively. The true model should yield uniformly distributed PPCs (dotted line), as indeed the power law does. The other models generated data that systematically differs from the true data.

Finally, in our worked example, we skipped one of the hardest steps: how does one arrive at the theory of temporal difference learning in the first place, not to mention these hypotheses of where and how various signals are encoded? We relied on these assumptions to place critical constraints on the space of models, and when they were taken into account, we obtained a very different view of the data than with the standard LDS. However, in the regime with few constraints and only vague ideas of how the systems under study work, standard models are invaluable tools for exploratory analysis [73]. That is, in the early stages of the pipeline, when computational theory is lacking, relatively unconstrained models are invaluable tools for generating hypotheses than can then be and refined with this pipeline.

Conclusions

The idea of combining statistical models with computational theories is not new [c.f. 5], but researchers are only beginning to appreciate the range of possibilities that have opened up with advances in probabilistic modeling. Richly expressive probabilistic programming languages, efficient inference algorithms, and flexible Bayesian nonparametric priors allow complex models to be specified and fit to data much more easily than in the past. Model criticism and comparison techniques can be used to guide

the refinement of modeling assumptions, as in Box's loop. We have shown how this statistical toolbox can be seamlessly integrated with computational theory, using a worked example from reinforcement learning. The key lesson from this modeling exercise is that data-driven and theory-driven approaches to neuroscience need not be mutually exclusive; indeed, the most powerful insights can be gained by using computational theories as constraints on data-driven statistical models.

Conversely, flexible statistical models can enrich computational theories. Historically, computational tractability has biased the kinds of models we fit towards simplicity (conjugacy, convex optimization problems, unimodal posteriors, low-dimensional parametrizations). With faster computers, larger datasets and new algorithms, machine learning has increasingly pushed the envelope towards much more complex models [33••,74,75], altering the usual tradeoff between neuroscientific realism and computational tractability. We are now in a position to start experimentally testing a vast range of computational theories.

Although we have emphasized probabilistic models in this paper, the same ideas apply to deterministic models, where apparent randomness is due to ignorance of latent variables and measurement noise. For example, although spike generation is often modeled as a random process, neurophysiological experiments suggest that spike generation may be highly reliable when a neuron is stimulated with white noise inputs [76]. Thus, neurons *seem* random until we condition on the relevant latent variables. The Bayesian framework does not require an ontological commitment to randomness; uncertainty can be purely epistemic. The practical motivation for building probabilistic models of deterministic processes is that it allows us to parse the different sources of uncertainty. Once we know that spike generation can be highly reliable, we should push our uncertainty into other parts of the model (synaptic inputs, ion channels, etc.). Constraining uncertainty in this way can be a driving force for the discovery of new latent variables that explain away residual randomness.

Conflict of interest statement

Nothing declared.

Acknowledgements

SWL is supported by a Simons Collaboration on the Global Brain postdoctoral fellowship (SCGB-418011). SJG is supported by the National Institutes of Health (CRCNS R01MH109177). We thank Liam Paninski for his helpful feedback on this work.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Box GEP, Hunter WG: **A useful method for model-building**. *Technometrics* 1962, **4**:301-318.

2. Box GEP: **Science and statistics**. *J Am Stat Assoc* 1976, **71**:791-799.

3. Box GEP: **Sampling and Bayes' inference in scientific modelling and robustness**. *J R Stat Soc Ser A (Gen)* 1980:383-430.

4. Popper KR: *The logic of scientific discovery*. 1959.

5. Durstewitz D, Koppe G, Toutounji H: **Computational models as statistical tools**. *Curr Opin Behav Sci* 2016, **11**:93-99.

6. Blei DM: **Build, compute, critique, repeat: data analysis with latent variable models**. *Annu Rev Stat Appl* 2014, **1**:203-232.

7. The ATLAS Collaboration: **A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider**. *Science* 2012, **338**:1576-1582.

8. Huys QJM, Paninski L: **Smoothing of, and parameter estimation from, noisy biophysical recordings**. *PLoS Comput Biol* 2009, **5**.

9. Meng L, Kramer MA, Middleton SJ, Whittington MA, Eden UT: **A unified approach to linking experimental, statistical and computational analysis of spike train data**. *PLOS ONE* 2014, **9**:e85269.

10. Paninski L: **Maximum likelihood estimation of cascade point-process neural encoding models**. *Network* 2004, **15**:243-262.

11. Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN: **A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects**. *J Neurophysiol* 2005, **93**:1074-1089.

12. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP: **Spatio-temporal correlations and visual signalling in a complete neuronal population**. *Nature* 2008, **454**:995-999.

13. Park IM, Archer EW, Priebe N, Pillow JW: **Spectral methods for neural characterization using generalized quadratic models**. *Advances in Neural Information Processing Systems*. 2013:2454-2462.

14. Latimer KW, Chichilnisky EJ, Rieke F, Pillow JW: **Inferring synaptic conductances from spike trains with a biophysically inspired point process model**. *Advances in Neural Information Processing Systems*. 2014:954-962.

15. Linderman SW, Adams RP, Pillow JW: **Bayesian latent structure discovery from multi-neuron recordings**. *Advances in Neural Information Processing Systems*. 2016:2002-2010.
A probabilistic method for recovering circuit organization from multi-neuron spike trains.

16. Gerhard F, Kispersky T, Gutierrez GJ, Marder E, Kramer M, Eden U: **Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone**. *PLoS Comput Biol* 2013, **9**:e1003138.

17. Soudry D, Keshri S, Stinson P, Oh M-h, Iyengar G, Paninski L: **Efficient "shotgun" inference of neural connectivity from highly sub-sampled activity data**. *PLoS Comput Biol* 2015, **11**:e1004464.

18. Linderman SW, Stock CH, Adams RP: **A framework for studying synaptic plasticity with neural spike train data**. *Advances in Neural Information Processing Systems*. 2014:2330-2338.

19. Stevenson I, Koerding K: **Inferring spike-timing-dependent plasticity from spike train data**. *Advances in Neural Information Processing Systems*. 2011:2582-2590.

20. Robinson BS, Berger TW, Song D: **Identification of stable spike-timing-dependent plasticity from spiking activity with generalized multilinear modeling**. *Neural Comput* 2016.

21. Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW: **Single-trial spike trains in parietal cortex reveal discrete steps during decision-making**. *Science* 2015, **349**:184-187.
A comprehensive analysis that calls into question the traditional view of gradual evidence accumulation in parietal cortex.

22. Gold JI, Shadlen MN: **The neural basis of decision making**. *Annu Rev Neurosci* 2007, **30**:535-574.

23. Zylberberg A, Shadlen MN: **Cause for pause before leaping to conclusions about stepping**. *bioRxiv* 2016:085886.

24. Detre GJ, Natarajan A, Gershman SJ, Norman KA: **Moderate levels of activation lead to forgetting in the think/no-think paradigm**. *Neuropsychologia* 2013, **51**:2371-2388.

25. Norman KA, Newman E, Detre G, Polyn S: **How inhibitory oscillations can train neural networks and punish competitors**. *Neural Comput* 2006, **18**:1577-1610.

26. Kim G, Lewis-Peacock JA, Norman KA, Turk-Browne NB: **Pruning of memories by context-based prediction error**. *Proc Natl Acad Sci U S A* 2014, **111**:8997-9002.

27. Lewis-Peacock JA, Norman KA: **Competition between items in working memory leads to forgetting**. *Nat Commun* 2014, **5**.

28. van der Maaten L, Hinton G: **Visualizing data using t-SNE**. *J Mach Learn Res* 2008, **9**:2579-2605.

29. Cunningham JP, Yu BM: **Dimensionality reduction for large-scale neural recordings**. *Nat Neurosci* 2014, **17**:1500-1509.

30. Paninski L, Ahmadian Y, Ferreira DG, Koyama S, Rad KR, Vidne M, Vogelstein J, Wu W: **A new look at state-space models for neural data**. *J Comput Neurosci* 2010, **29**:107-126.

31. Macke JH, Buesing L, Cunningham JP, Yu Byron M, Shenoy KV, Sahani M: **Empirical models of spiking in neural populations**. *Advances in Neural Information Processing Systems*. 2011:1350-1358.

32. Linderman SW, Miller AC, Adams RP, Blei DM, Paninski L, Johnson MJ: *Recurrent switching linear dynamical systems*. 2016 arXiv:1610.08466.

33. Sussillo D, Jozefowicz R, Abbott LF, Pandarinath C: **LFADS — latent factor analysis via dynamical systems**. *Advances in Neural Information Processing Systems*. 2016.
A dimensionality reduction technique for multi-neuron recordings that uses constraints from dynamical systems theory.

34. Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, Abraira VE, Adams RP, Robert Datta S: **Mapping sub-second structure in mouse behavior**. *Neuron* 2015, **88**:1121-1135.

35. Gershman SJ, Blei DM: **A tutorial on Bayesian nonparametric models**. *J Math Psychol* 2012, **56**:1-12.

36. Gershman S, Moustafa A, Ludvig E: **Time representation in reinforcement learning models of the basal ganglia**. *Front Comput Neurosci* 2014, **7**:194.

37. Sutton RS, Barto AG: **Toward a modern theory of adaptive networks: expectation and prediction**. *Psychol Rev* 1981, **88**:135.

38. Schultz W, Dayan P, Read Montague P: **A neural substrate of prediction and reward**. *Science* 1997, **275**:1593-1599.

39. Niv Y: **Reinforcement learning in the brain**. *J Math Psychol* 2009, **53**:139-154.

40. Durbin J, Koopman SJ: **Time Series Analysis by State Space Methods**. Oxford University Press; 2012.

41. Robert C, Casella G: *Monte Carlo Statistical Methods*. Springer Science & Business Media; 2013.

42. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK: **An introduction to variational methods for graphical models**. *Mach Learn* 1999, **37**:183-233.

43. Wainwright MJ, Jordan MI: **Graphical models, exponential families, and variational inference**. *Found Trends Mach Learn* 2008, **1**:1-305.

44. Wood F, van de Meent J-W, Mansinghka V: **A new approach to probabilistic programming inference**. *AISTATS*. 2014:1024-1032.

45. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A: **Stan: a probabilistic programming language**. *J Stat Softw* 2016, **20**.

46. Mansinghka V, Selsam D, Perov Y: *Venture: a higher-order probabilistic programming platform with programmable inference*. 2014 arXiv:1404.0099.

47. Tran D, Kucukelbir A, Dieng AB, Rudolph M, Liang D, Blei DM: *Edward: A library for probabilistic modeling, inference, and criticism*. 2016 arXiv:1610.09787.

48. Gelman A, Rubin DB: **Inference from iterative simulation using multiple sequences**. *Statistical Science*. 1992:457-472.

49. Cowles MK, Carlin BP: **Markov chain Monte Carlo convergence diagnostics: a comparative review**. *J Am Stat Assoc* 1996, **91**:883-904.

50. Cook SR, Gelman A, Rubin DB: **Validation of software for Bayesian models using posterior quantiles**. *J Comput Graph Stat* 2006, **15**:675-692.

51. Cusumano-Towner MF, Mansinghka VK: *AIDE: an algorithm for measuring the accuracy of probabilistic inference algorithms*. 2017 arXiv:1705.07224.

52. Rubin DB: **Bayesianly justifiable and relevant frequency calculations for the applied statistician**. *Ann Stat* 1984, **12**:1151-1172.

53. Meng X-L: **Posterior predictive *p*-values**. *Ann Stat* 1994:1142-1160.

54. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: *Bayesian Data Analysis*. edn 3. CRC Press; 2013.

55. MacKay DJC: **Bayesian interpolation**. *Neural Comput* 1992, **4**:415-447.

56. Kass RE, Raftery AE: **Bayes factors**. *J Am Stat Assoc* 1995, **90**:773-795.

57. Neal RM: **Annealed importance sampling**. *Stat Comput* 2001, **11**:125-139.

58. Grosse RB, Maddison CJ, Salakhutdinov RR: **Annealing between distributions by averaging moments**. *Advances in Neural Information Processing Systems*. 2013:2769-2777.

59. Grosse RB, Ghahramani Z, Adams RP: *Sandwiching the marginal likelihood using bidirectional Monte Carlo*. 2015 arXiv:1511.02543.

60. Carlson D, Stinson P, Pakman A, Paninski L: **Partition functions from Rao-Blackwellized tempered sampling**. In *Proceedings of The 33rd International Conference on Machine Learning*. 2016:2896-2905.

61. Lindley DV: **On a measure of the information provided by an experiment**. *Ann Math Stat* 1956:986-1005.

62. Paninski L: **Asymptotic theory of information-theoretic experimental design**. *Neural Comput* 2005, **17**:1480-1507.

63. Lewi J, Butera R, Paninski L: **Sequential optimal design of neurophysiology experiments**. *Neural Comput* 2009, **21**:619-687.

64. Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N: **Taking the human out of the loop: a review of Bayesian optimization**. *Proc IEEE* 2016, **104**:148-175.

65. Huggins JH, Paninski L: **Optimal experimental design for sampling voltage on dendritic trees in the low-SNR regime**. *J Comput Neurosci* 2012, **32**:347-366.

66. Bak JH, Choi J, Witten I, Pillow JW: **Adaptive optimal training of animal behavior**. *Advances in Neural Information Processing Systems*. 2016:1939-1947.
An online optimal experimental design algorithm for inferring the rules generating animal behavior.

67. Lloyd JR, Ghahramani Z: **Statistical model criticism using kernel two sample tests**. *Advances in Neural Information Processing Systems*. 2015:829-837.

68. Lopez-Paz D, Oquab M: *Revisiting classifier two-sample tests*. 2016 arXiv:1610.06545.

69. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y: **Generative adversarial nets**.

Advances in Neural Information Processing Systems. 2014:2672-2680.

70. Grosse R, Salakhutdinov RR, Freeman WT, Tenenbaum JB: **Exploiting compositionality to explore a large space of model structures**. *Uncertainty and Artificial Intelligence (UAI)*. 2012.

71. Duvenaud DK, Lloyd JR, Grosse RB, Tenenbaum JB, Ghahramani Z: **Structure discovery in nonparametric regression through compositional kernel search**. In *Proceedings of the International Conference on Machine Learning (ICML)*. 2013:1166-1174.

72. Schulz E, Tenenbaum J, Duvenaud DK, Speekenbrink M, Gershman SJ: **Probing the compositionality of intuitive functions**. *Advances in Neural Information Processing Systems 29*. 2016:3729-3737.

73. Tukey JW: *Exploratory Data Analysis*. 1977.

74. Jordan MI *et al.*: **On statistics, computation and scalability**. *Bernoulli* 2013, **19**:1378-1390.

75. Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J *et al.*: **Hybrid computing using a neural network with dynamic external memory**. *Nature* 2016, **538**:471-476.

76. Mainen ZF, Sejnowski TJ: **Reliability of spike timing in neocortical neurons**. *Science* 1995, **268**:1503-1506.