Psychon Bull Rev (2014) 21:227–250
DOI 10.3758/s13423-013-0530-0

THEORETICAL REVIEW

A generalized, likelihood-free method for posterior estimation

Brandon M. Turner · Per B. Sederberg

Published online: 21 November 2013

Abstract Recent advancements in Bayesian modeling have allowed for likelihood-free posterior estimation. Such estimation techniques are crucial to the understanding of simulation-based models, whose likelihood functions may be difficult or even impossible to derive. However, current approaches are limited by their dependence on sufficient statistics and/or tolerance thresholds. In this article, we provide a new approach that requires no summary statistics, error terms, or thresholds and is generalizable to all models in psychology that can be simulated. We use our algorithm to fit a variety of cognitive models with known likelihood functions to ensure the accuracy of our approach. We then apply our method to two real-world examples to illustrate the types of complex problems our method solves. In the first example, we fit an error-correcting criterion model of signal detection, whose criterion dynamically adjusts after every trial. We then fit two models of choice response time to experimental data: the linear ballistic accumulator model, which has a known likelihood, and the leaky competing accumulator model, whose likelihood is intractable. The estimated posterior distributions of the two models allow for direct parameter interpretation and model comparison by means of conventional Bayesian statistics—a feat that was not previously possible.

Keywords Probability density approximation · Cognitive modeling · Likelihood-free inference · Estimation · Error-correcting criterion model · Leaky competing accumulator model · Linear ballistic accumulator model

Code is freely available for implementing the PDA method in an online appendix, and on each authors' websites.

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0530-0) contains supplementary material, which is available to authorized users.

B. M. Turner (✉)
Department of Psychology, Stanford University, Stanford, CA, USA
e-mail: turner.826@gmail.com

P. B. Sederberg
Department of Psychology, The Ohio State University, Columbus, OH, USA
e-mail: sederberg.1@osu.edu

Introduction

The goal of cognitive modeling is to understand complex behaviors within a system of mathematically specified mechanisms or processes. While cognitive models can vary in complexity from relatively simple models of choice response time (e.g., Brown & Heathcote, 2008) to full cognitive architectures (e.g., Anderson, 2007), cognitive models all have a set of parameters that govern the proposed mechanisms and, ideally, lend themselves to psychologically meaningful interpretations. For example, a parameter might correspond to bias, the tendency to prefer one alternative over another, or discriminability, the degree of perceptual clarity a stimulus provides.

Cognitive models are important because they are designed to embody a set of psychological principles or cognitive theories. In order to properly test the assumptions made by a cognitive theory, a researcher would begin by designing an experiment that directly tests the theory. The next step is to fit the cognitive model to the data. A good model fit to the data would support the underlying cognitive theory, whereas a bad fit would refute the theory. After fitting the model, we obtain an estimate of the model parameters, which carry valuable information about how the model captures the observed behavior.

Given the detailed information that the parameters of a cognitive model provide, from a theoretical perspective, it is essential that we fully understand how the parameters of a model affect the model predictions. Furthermore, it is critical that we understand how the parameters behave and how groups of parameters might interact with one another. From an inferential perspective, it is important not only that our

parameter estimates are accurate, but also that the degree of uncertainty in our estimates can be properly assessed. Without accurate and informative parameter estimates, we risk drawing incorrect conclusions not only about the data, but about the models as well.

Despite the importance of understanding the full range of valid parameter estimates, the most common approach to parameter inference is least squares estimation. In this procedure, the goal is to determine the set of parameter values that minimize some discrepancy measure (e.g., a Euclidean distance metric) between the simulated and observed data. Another approach is Bayesian modeling, which provides a framework within which one can simultaneously understand both the estimates of the model parameters and the uncertainty about them.[1] Implementing the Bayesian approach requires both a prior distribution for the parameters of the model and a likelihood function for the data. One can (virtually) always specify a prior, which usually comes in the form of some convenient distribution (e.g., a Gaussian distribution). By contrast, the likelihood function can be difficult to fully specify. Indeed, the difficulties encountered in deriving the full likelihood function have prevented the application of fully Bayesian analyses for many cognitive models.

Recent advances have made it possible to perform Bayesian analyses without having an explicit likelihood function, and these techniques have generated new insights into simulation-based models (Beaumont, 2010; Beaumont, Cornuet, Marin, & Robert, 2009; Beaumont, Zhang, & Balding, 2002; Marjoram, Molitor, Plagnol, & Tavare, 2003; Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999; Sisson, Fan, & Tanaka, 2007; Turner, Dennis, & Van Zandt, 2013; Turner, & Sederberg, 2012; Turner & Van Zandt, 2012, 2013; Wood, 2010). While likelihood-free techniques have spurred many new avenues for Bayesian analyses, at best, present likelihood-free methods rely on an assumption that can rarely be justified in practice. Specifically, likelihood-free methods require that the set of summary statistics used by the algorithm are *sufficient* for the model parameters of interest. A sufficient statistic is a statistic that when calculated from a set of data, provides just as much information about the unknown parameters of a model as the entire data set. When a set of summary statistics are not sufficient, one cannot guarantee convergence to the correct posterior distribution. Because it is often impossible to guarantee that a summary statistic is sufficient when a likelihood function is unavailable, likelihood-free algorithms that do not have sufficient statistics necessarily introduce error into the posterior distribution, and this error is not directly measurable (Beaumont, 2010).

Here, we present a fully generalizable method for performing likelihood-free Bayesian parameter estimation that does not depend on summary statistics. The method is based on previous efforts for maximum likelihood estimation (Fermanian & Salanié, 2004). Here, we adapt the method for Bayesian estimation, extend it to data consisting of both discrete and continuous measures (e.g., choice response time), and apply it to several cognitive models. We begin with a brief discussion of standard Bayesian methods. We then introduce the probability density approximation (PDA) method and discuss how it can be applied to discrete, continuous, and mixed (i.e., data consisting of both discrete and continuous measures) data types through illustrative examples. We then apply the method to two real-world problems. First, we fit a hierarchical, dynamic, criterion adjustment model of signal detection to experimental data. Despite being dynamic, the model's likelihood function can be evaluated, and true estimates of the posterior can be compared with estimates obtained using the PDA method. Second, we fit two hierarchical models of choice response time—the linear ballistic accumulator (LBA; Brown & Heathcote, 2008) model, and the leaky competing accumulator (LCA; Usher & McClelland, 2001) model—to experimental data. The LBA model also has a tractable likelihood function, which allows us to further assess the accuracy of the PDA method. By contrast, the LCA model is intractable, and so it has never been analyzed using Bayesian techniques. Our Bayesian analyses provide a new understanding of the parameters within all three models.

Bayesian estimation

The goal of Bayesian inference is to obtain the *posterior* distribution of the parameters of interest θ conditioned on the observed data $Y=\{Y_1,Y_2,\ldots,Y_N\}$.[2] To do so, one must specify both a likelihood function for the data, $L(\theta|Y)$, and a prior distribution, $\pi(\theta)$, for the parameters θ. To specify the likelihood function, we require a probability density function relating any observation Y_i to the cognitive model. We will denote this probability density function as Model$(y|\theta)$. If we can assume that the observations Y are independent and identically distributed (i.i.d.), the likelihood function is given by

$$L(\theta|Y) = \prod_{i=1}^{N} \text{Model}(Y_i|\theta). \quad (1)$$

To specify the prior distribution, we need only provide a distribution that reflects our prior knowledge of the distribution of the parameters θ. The selection of the prior is

[1] While it is true that similar information can be obtained through the Hessian or Fischer information matrix, such a procedure requires that the log likelihood function be twice differentiable. When the likelihood function is unknown—the central problem we address in this article—a Hessian matrix cannot be calculated.

[2] A detailed introduction to Bayesian methodology is beyond the scope of this article. We recommend that the interested reader consult one of the many texts on the topic (e.g., Christensen, Johnson, Branscum, & Hanson, 2011; Gelman, Carlin, Stern, & Rubin, 2004; Kruschke, 2011; Lee, & Wagenmakers, 2012)).

subjective and will vary from one researcher to the next. Because any number of distributions can be used to reflect one's prior knowledge, it is virtually always possible to specify a prior.

Once the likelihood function and prior distribution are specified for the model, the posterior distribution is given by

$$\pi(\theta|Y) = \frac{L(\theta|Y)\pi(\theta)}{\int L(\theta|Y)\pi(\theta)d\theta} \tag{2}$$

$$\propto L(\theta|Y)\pi(\theta). \tag{3}$$

Equation 2 shows that the posterior distribution is a function that provides the probability of the parameters θ, conditioned on the data that were observed. Having such a distribution allows us to simultaneously assess the "best" estimate for the parameter as well as quantify the uncertainty in that estimate in a way that does not depend on hypothetical data, as in null hypothesis testing (see Wagenmakers, 2007).

While the Bayesian framework is appealing and powerful in theory, estimating the posterior distribution depends on our ability to specify the likelihood function in Eq. 1. Because the mechanisms used by cognitive models are often complex, the likelihood function can be difficult to evaluate or even impossible to specify mathematically. There are already a number of models with intractable likelihood functions in psychology, and we suspect that this number will grow with the increased demand for neurological plausibility, temporal dependency (e.g., free recall; Howard & Kahana, 2002; Polyn, Norman, & Kahana, 2009; Raaijmakers & Shiffrin, 1981; Sederberg, Howard, & Kahana, 2008), and unified explanations of multiple measures of behavioral data (e.g., choice, response time, and confidence; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009).

Consider, for example, the LCA model (Usher & McClelland, 2001). The LCA model was proposed as a neurologically plausible model for choice response time in a k-alternative task. The model possesses mechanisms that extend other diffusion-type models (e.g., Ratcliff, 1978) by including leakage and competition by means of lateral inhibition. Because the evidence accumulation process used by the LCA model was designed to mimic actual neuronal activation patterns, one critical assumption is that the degree of evidence for any choice alternative can never be negative. To accommodate this assumption, if any accumulator in the model becomes negative, the degree of evidence for that accumulator is reset to zero. The LCA model also assumes a competition among response alternatives that depends on the current state of each of the accumulators. Together, these features of the model sufficiently complicate the equations describing the joint distributions of choice and response time such that the likelihood function for the LCA model has not been derived. As a result, all model evaluations to this point have been performed using either a model simplification or least squares estimation (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz, Usher, Zhang, & McClelland, 2007; Gao, Tortell, & McClelland, 2011; Tsetsos, Usher, & McClelland, 2011; Usher & McClelland, 2001; van Ravenzwaaij, van der Maas, & Wagenmakers, 2012).

When one cannot evaluate Eq. 1, standard Bayesian estimation is simply not possible. As a result, and given the increased popularity of the Bayesian approach, there has been a recent surge of interest in developing new techniques to approximate the standard Bayesian solution. The general form of a likelihood-free algorithm is to first propose a value for the parameter of interest. Second, one simulates the model of interest many times under the proposed parameter value. Third, one calculates a set of summary statistics $S(\cdot)$ for both the simulated and observed data (i.e., the data to be fit by the model). Finally, the two sets of statistics are compared, and the proposal parameter is assigned a "score" based on this comparison (for tutorials, see Csilléry, Blum, Gaggiotti & François, 2010; Turner & Van Zandt, 2012). If specified properly, the scores associated with the proposals should approximate the likelihood function. However, for likelihood-free algorithms to work, the summary statistics must be sufficient for the parameters of interest.

The problem of sufficiency

Condensing the data down to a set of statistics $S(\cdot)$ provides a computationally convenient way to assess the similarity of the simulated data X to the observed data Y. However, some summary statistics carry more information about the unknown model parameters θ than do others. The ideal situation is when the summary statistics $S(\cdot)$ are sufficient for the parameters. When using sufficient statistics, no information about the unknown parameters θ is lost in the compression of Y to $S(Y)$, and so the relationship

$$\pi(\theta|Y) = \pi(\theta|S(Y)),$$

is satisfied. If the likelihood function is known, one can use the Fisher–Neyman factorization theorem (Rice, 2007) to prove that a statistic is sufficient for θ or that a set of summary statistics are jointly sufficient for θ.

Requiring that $S(\cdot)$ be sufficient is problematic because one cannot guarantee sufficiency when the likelihood function is unknown or is intractable because the Fisher–Neyman factorization theorem cannot be applied. The most common resolution is to select a large set of summary statistics and hope that the set of statistics are jointly sufficient for the parameters of interest. While adding more summary statistics will tend to provide more information about θ, this is not necessarily true in general. When a set of summary statistics are not sufficient for the parameters, the influence of the

information conveyed by the observed data will be weaker, resulting in posterior distributions that are inaccurate, particularly with respect to the degree of variability in the estimated posteriors (Beaumont, 2010).

The probability density approximation method

Our method differs from other likelihood-free algorithms in two substantial ways. First, the PDA method makes no assumption that a set of summary statistics be jointly sufficient for the parameters of interest. Second, the PDA method is a nonparametric approach, and so it does not require any restrictive assumptions about the distribution of the summary statistics $S(\cdot)$, as required by other approaches discussed below. The equations involved in describing the PDA method differ depending on whether the data are discrete or continuous. Thus, we first present the algorithm in a general form and then provide more specific details for each of the different types of data.

We again assume that the observed data $Y = \{Y_1, Y_2, \ldots, Y_N\}$ arise from a model so that $Y \sim \text{Model}(\theta)$.[3] We begin by generating a proposal θ^*. The method for generating θ^* can be one of many options (e.g., importance sampling). We then use θ^* to simulate a set of data $X = \{X_1, X_2, \ldots, X_J\}$ from the assumed model, so that $X \sim \text{Model}(\theta^*)$. Next, we estimate the form of the random distribution of X, which we call the "simulated probability density function" (SPDF) and denote $f(x|X)$. Using the SPDF, we evaluate the density of the observed data Y under a given θ by the equation

$$\text{Model}(Y_i|\theta) = f(Y_i|X). \tag{4}$$

Thus, after evaluation of Eq. 4, we obtain a density under the assumed model for every point in the data set Y. Because the data are always sufficient to themselves, our density estimation procedure allows us to guarantee sufficiency because the summary statistics are computed for each individual observation Y_i.

In correspondence with Eq. 1, an approximation of the likelihood function is

$$\mathcal{L}(\theta|Y) = \prod_{i=1}^{N} \text{Model}(Y_i|\theta). \tag{5}$$

Because the likelihood function in Eq. 5 is still an approximation, we denote it $\mathcal{L}(\theta|Y)$ to draw a distinction from Eq. 1. Thus, for a given proposal θ^*, the "pseudo-likelihood" is determined by plugging θ^* in for θ in Eq. 5. Finally, the posterior density, up to a constant of proportionality, would be approximated by the equation

$$\pi(\theta|Y) \propto \mathcal{L}(\theta|Y)\pi(\theta). \tag{6}$$

As was mentioned above, the construction of the SPDF differs for different types of data. Specifically, for discrete data, we construct an empirical probability mass function, whereas for continuous data we construct a continuous probability density function of the simulated data X. We now discuss each of these specific applications in more detail.

Discrete data

For discrete data such as typical confidence responses (e.g., a Likert scale) or the number of hits and false alarms (and by extension, hit and false alarm rates), the SPDF $f(x|X)$ is constructed by means of a probability mass function. We first define a sample space $S = \{s_1, s_2, \ldots, s_n\}$ as the set of all possible outcomes in our experiment.

For example, in a recognition memory experiment, after a study phase, subjects are asked to classify test items as either "new" (i.e., not on the previously studied list) or "old" (i.e., on the previous list). During the test phase, an experimenter can present a distractor, an item that was not on the study list, or a target, an item that was on the previously studied list. For each item type presented at test, there are two possible responses, and so there are only four possible stimulus–response outcomes. By convention, memory researchers focus on the number of hits and false alarms, which occur when a response of "old" is elicited for a target item or a distractor item, respectively. Hit rates are determined by dividing the number of hits by the number of target items presented (i.e., the number of *possible* hits). Similarly, false alarm rates are determined by dividing the number of false alarms by the number of distractors presented. If we let the total number of targets in an experiment be denoted T and the number of total distractors be denoted D, then the sample space for hit rates is $S = \{0, 1/T, 2/T, \ldots, (T-1)/T, 1\}$, and the sample space for false alarm rates is $S = \{0, 1/D, 2/D, \ldots, (D-1)/D, 1\}$.

Due to randomness in the process of model simulation, the simulated data X are random variables with their own sample space. If we define the set of possible outcomes as $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, we can define a probability function $P(\cdot)$ such that

$$P(X = x_i) = P\left(s_j \in S : X\left(s_j\right) = x_i\right), \tag{7}$$

which restricts the sample space of the simulated data $\mathcal{X}$ to lie in the sample space of the experiment S. Equation 7 is a way of mathematically expressing that the probability of X equaling a given value of x_i is simply the number of times that the simulated data equal the given value x_i, divided by the

[3] We note that the notation Model(θ) describes the distribution of a random variable, whereas the notation $Y \sim \text{Model}(y|\theta)$ denotes the probability density at the location y, conditional on the parameters θ, as in Eq. 1.

total number of model simulations. Because Eq. 7 defines a probability for all possible outcomes, we use it to construct our SPDF so that

$$f(x|X) = P(X = x), \tag{8}$$

which is then used in Eqs. 4 and 5 to evaluate the pseudo-likelihood of the proposal θ^*.

The signal detection theory model

As a concrete example, consider the equal-variance model of signal detection theory (SDT; see Egan, 1958; Green & Swets, 1966; Macmillan & Creelman, 2005). For this model, the parameters of interest are the discriminability parameter d' and the bias parameter β, such that the hit rates h and false alarm rates f are

$$\begin{aligned} h &= \Phi\left(d'/2-\beta\right), \text{ and} \\ f &= \Phi\left(-d'/2-\beta\right), \end{aligned} \tag{9}$$

where $\Phi(x)$ is the cumulative distribution function of the normal distribution evaluated at the location x (Lee, 2008; Lee & Wagenmakers, 2012; Rouder & Lu, 2005).

For this model, the true PDF is well-known, which permits a direct evaluation of Eq. 1. To compare the quality of the SPDF with the true PDF, we performed an illustrative simulation study. We set the model parameters to $\theta = \{d' = 1, \beta = 0.1\}$ and used them to generate data consisting of an "old"/"new" judgment for 100 test items, 50 of which were targets and 50 were distractors. To construct the SPDF, we simulated the model J = 1,000 times so that $X = \{X_1, X_2, \ldots, X_{1000}\}$, where $X_j = \left[\widehat{h}_j, \widehat{f}_j\right]$ and $\widehat{h}_j$ and $\widehat{f}_j$ are the empirical hit and false alarm rates for the jth simulation. We then evaluated Eq. 8 to obtain an estimated probability of the data for each location in the data space.

Figure 1 shows the PDFs (densities) and the SPDFs (black vertical lines) for hit (left panel) and false alarm (right panel) rates. The true PDFs were obtained by evaluating

$$\begin{aligned} P(H = x|\theta) &= \text{Bin}(50x|50, h), \text{ and} \\ P(F = x|\theta) &= \text{Bin}(50x|50, f), \end{aligned}$$

where H and F are the possible hit and false alarm rates (i.e., the x-axis in Fig. 1), respectively, h and f are calculated by Eq. 9, and $\text{Bin}(x|a, b)$ is the binomial density evaluated at the location x with the number of trials equal to a and the probability b of a single-trial success. The close match between the true PDFs and the SPDFs demonstrates that simulating the model 1,000 times provides a reasonably stable approximation to the true PDF.

If we were interested in fitting the equal-variance SDT model to observed data Y, we would then evaluate Eq. 8 by plugging in each Y_i for x. The densities obtained, $f(Y_i|X) = \text{Model}(Y_i|\theta)$ can then be multiplied to form the pseudo-likelihood function in Eq. 5.

Continuous data

When the data Y have continuous measurements, we cannot rely on a probability mass function to characterize the random distribution of Y. Instead, we must use our simulated data X to form an approximation of the PDF via a density function. While there are many ways of specifying the density function, we propose a nonparametric procedure by constructing a kernel density estimate (see Silverman, 1986). A kernel density estimate provides a way to estimate the true probability density function by using the full simulated data set.

To use the PDA method, we proceed in the same way as in the discrete case by first generating a proposal θ^* and using the proposal to generate a sequence of observations from the model, so that $X \sim \text{Model}(\theta^*)$. We then construct a kernel density estimate of the simulated data so that

$$f(x|X) = \frac{1}{hJ}\sum_{j=1}^{J} K\left(\frac{x-X_j}{h}\right), \tag{10}$$

where $\int f(x|X)dx = 1$. The function $K(\cdot)$ is the kernel, and h is a smoothing parameter known as the bandwidth.[4] The kernel is usually chosen to be unimodal and symmetric about zero to place a decreasing weight on observations X_j further from the point where the density is being estimated. While the kernel can take many forms, in this article we will consider only the Epanechnikov kernel, given by

$$K(x) = \begin{cases} \frac{3}{4}\left(1-x^2\right) & \text{if } x \in [-1, 1] \\ 0 & \text{if } x \notin [-1, 1] \end{cases}. \tag{11}$$

The accuracy of kernel density function is measured by the mean integrated squared error (MISE), a measure of divergence between a true and an estimated density function. The Epanechnikov kernel was derived on the basis of minimizing the asymptotic MISE, and so it is optimal in a statistical sense (Epanechnikov, 1969; Silverman, 1986). To select a bandwidth, we use Silverman's rule of thumb, so that

$$h = 0.9min\left(SD(X), \frac{IQR(X)}{1.34}\right)n^{-1/5}, \tag{12}$$

where $SD(\cdot)$ denotes the standard deviation and $IQR(\cdot)$ denotes the interquartile range. While these choices work well

[4] Although our kernel density estimate depends on the bandwidth h, we drop this conditioning from the notation, for simplicity.

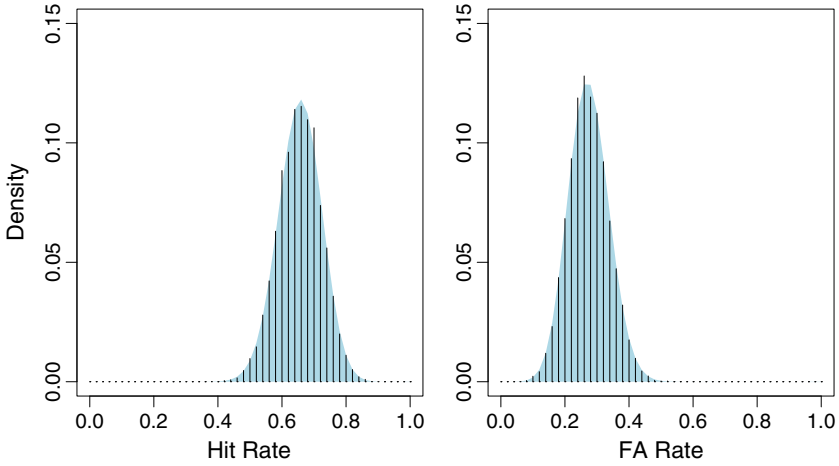

Fig. 1 Simulated probability density functions (black lines) along with the true probability density functions (shaded densities) for the equal-variance signal detection theory model. Hit rates are shown in the left panel, and false alarm rates are shown in the right panel

for all of the examples we have performed and are the standard methods available in the scientific libraries in R and Python, we speculate that our method of kernel density estimation could be further improved upon, especially in the case of small samples (see, e.g., Chapeau-Blondeau & Rousseau, 2009; Kontkanen & Myllymäki, 2007).

Once we have constructed the SPDF $f(x|X)$ via kernel density estimation, we can calculate the pseudo-likelihood function by evaluating Eq. 5, and the posterior density is determined by Eq. 6. Equations 4 and 5 together show that each data point is used in the evaluation of the likelihood, so while our kernel density estimation procedure uses statistics, there is no compression of the observed data into summary statistics. Hence, because the data are always sufficient to themselves, our method guarantees sufficiency.

The Wald model

To illustrate our approach for continuous data, we now demonstrate how the PDA method can be used to estimate the parameters of the Wald model. The Wald distribution (also known as the inverse Gaussian distribution) is a statistical model of response times that is both analytically simple and theoretically motivated, because it describes the behavior of the first-passage time of a diffusion process with a single boundary (Chhikara & Folks, 1989; Wald, 1947). Ratcliff (1978) extended the single boundary diffusion process to a two-boundary process, but in so doing, rendered the likelihood function analytically intractable. Specifically, the likelihood function for a single-choice response time pair involves the sum of an oscillating but convergent infinite sum (see Feller, 1968; Lee, Fuss, & Navarro, 2006; Navarro & Fuss, 2009). However, for simple response time, where subjects provide a single response as soon as a signal is perceived (e.g., Heathcote, 2004; Luce, 1986; Schwarz, 2001), only one boundary is required for sequential-sampling-based psychological models. Thus, for a single-choice response time task, the Wald model can be used to provide a simple interpretation of the underlying processes theorized to be at work (see Rouder, Yue, Speckman, Pratte, & Province, 2010, for a sophisticated application).

The three-parameter Wald distribution has the density

$$f(y|\alpha,\nu,\tau)=\frac{\alpha}{\sqrt{2\pi(y-\tau)^3}}exp\left(-\frac{[\alpha-\nu(y-\tau)]^2}{2(y-\tau)}\right), \qquad (13)$$

where the parameter ν is related to the speed of cognitive processing, α is related to the amount of accumulated perceptual evidence that an observer requires before eliciting a response, and τ represents a nondecision component composed of processes such as perceptual encoding and motor control.

To demonstrate the PDA method, we simulated the single-boundary diffusion process with $\theta=\{\alpha=2,\nu=2.2,\tau=0.1\}$ $J=10{,}000$ times. The simulated data X were then used to construct the SPDF $f(x|X)$ by means of Eq. 10. We used the Epanechnikov kernel, as in Eq. 11, and determined the bandwidth parameter h by evaluating Eq. 12. The black line in Fig. 2 shows the SPDF $f(x|X)$. For comparison, we also calculated the true density by evaluating Eq. 13 along the interval [0, 5], which is represented in Fig. 2 as the shaded density. A visual comparison of these two distributions shows that the SPDF closely resembles the true PDF, and a two-sample Kolmogorov–Smirnov test failed to reject the null hypothesis that the two distributions were from the same generating mechanism ($D=0.072$, $p=.138$).

For continuous measures, the computational complexity of the PDA method may seem high. To lighten our computational burden, we can exploit the common assumption made in mathematical modeling of i.i.d. data. We can benefit from the i.i.d. assumption here because, once the SPDF has been constructed for a proposal θ^*, the cost associated with evaluating the likelihood function in Eq. 5 is

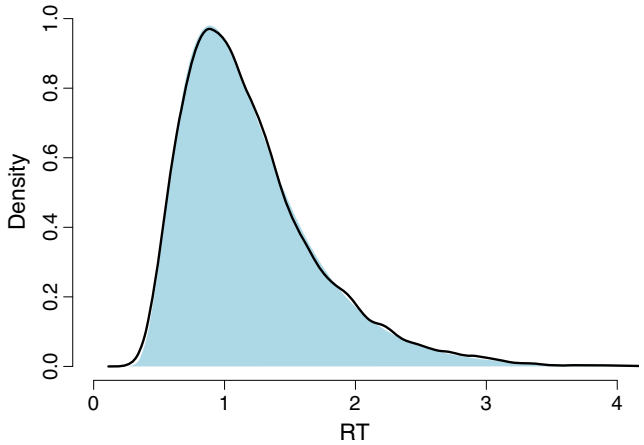

Fig. 2 Simulated probability density function (black line) along with the true probability density function (shaded density) for the Wald distribution

negligible with increases in the dimensionality of Y. While this is a convenient result and as we will show later in this article, the i.i.d. assumption is not required for the PDA method to work. For example, the PDA method could be used to fit models that do not assume an i.i.d. process (e.g., Craigmile, Peruggia, & Van Zandt, 2010; Dorfman & Biderman, 1971; Howard & Kahana, 2002; Kac, 1962, 1969; Peruggia, Van Zandt, & Chen, 2002; Sederberg et al., 2008; Turner, Van Zandt, & Brown, 2011; Vickers & Lee, 1998, 2000) by constructing a density estimate for each data point in the sequence Y (e.g., outcomes of trials within an experiment) and then evaluating the density of each observed data point under the corresponding SPDF. For these dynamic models, it is less clear how one would specify summary statistics, as required by other likelihood-free techniques.

Mixed data

Although we have outlined the PDA method for both discrete and continuous measurements, we have not made the extension to data consisting of both types explicit. Recently, the demand for models that can explain multiple sources of data simultaneously has increased (e.g., Cox & Shiffrin, 2012; Nosofsky, Little, Donkin, & Fific, 2011; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009). One reason is because data with multiple sources provide additional tests on the legitimacy of the assumptions made by the models. However, accounting for more sources of data generally requires more complicated mechanisms or more sophisticated assumptions, which can potentially render the likelihood function for these more complicated models analytically intractable.

For ease of exposition, we consider the common case of data consisting of one discrete measurement (e.g., choice) and one continuous measurement (e.g., response time). For the discrete measurements, suppose there are C options, and for the continuous measurements, there is an infinite number of possible values. We adopt a new notation for our simulated data, $X = (X^{(1)},X^{(2)},\ldots,X^{(C)})$, where $X^{(c)}$ is the set of continuous measurements for the cth discrete alternative. We then introduce a vector containing the set of the number of observations for each alternative, so that $n = \{n^{(1)},n^{(2)},\ldots, n^{(C)}\}$ and $J = \sum_{c=1}^{C} n^{(c)}$ (i.e., J denotes the *total* number of model simulations). We will similarly denote a set of bandwidth parameters $h = \{h^{(1)},h^{(2)},\ldots,h^{(C)}\}$, so that

$$h^{(c)} = 0.9min\left(SD\left(X^{(c)}\right), \frac{IQR\left(X^{(c)}\right)}{1.34}\right)\left(n^{(c)}\right)^{-1/5}. \quad (14)$$

For each discrete alternative, we construct a kernel density estimate for the simulated data (i.e., an SPDF), given by

$$f^{*}_{n^{(c)}}\left(x|X^{(c)}\right) = \frac{1}{h^{(c)}n^{(c)}}\sum_{j=1}^{n^{(c)}} K\left(\frac{x-X_j^{(c)}}{h^{(c)}}\right). \quad (15)$$

As a result of the construction of each individual $f^{*}_{n^{(c)}}\left(x|X^{(c)}\right)$, they integrate to one so that $\int f^{*}_{n^{(c)}}\left(x|X^{(c)}\right)dx = 1$. That is, each individual $f^{*}_{n^{(c)}}\left(x|X^{(c)}\right)$ is a proper probability density function, but they do not take into account the other $C-1$ alternatives. To create an SPDF that takes into account both the discrete and continuous random variables, we require that

$$\int \sum_{c=1}^{C} f_{n^{(c)}}\left(x|X^{(c)}\right)dx = 1,$$

for some function $f(\cdot)$, which is called the *defective* distribution because it does not integrate to one for a single alternative c (i.e., in nontrivial cases). To satisfy this constraint, we scale each density by the corresponding frequency of the alternative, so that

$$f_{n^{(c)}}\left(x|X^{(c)}\right) = \frac{n^{(c)}}{J} f^{*}_{n^{(c)}}\left(x|X^{(c)}\right).$$

Thus, for a proper SPDF function, Eq. 15 becomes

$$f_{n^{(c)}}\left(x|X^{(c)}\right) = \frac{1}{h^{(c)}J}\sum_{j=1}^{n^{(c)}} K\left(\frac{x-X_j^{(c)}}{h^{(c)}}\right). \quad (16)$$

For the observed data, we denote continuous data as $Y = \{Y_1,Y_2,\ldots,Y_N\}$ and discrete data as $Z = \{Z_1,Z_2,\ldots,Z_N\}$. Thus, for the ith pair (Y_i, Z_i), the density under the assumed model, conditional on the parameter θ, is given by

$$\text{Model}(Y_i, Z_i|\theta) = f_{n^{(Z_i)}}\left(Y_i|X^{(Z_i)}\right),$$

and the pseudo-likelihood function is given by

$$\mathcal{L}(\theta|Y,Z) = \prod_{i=1}^{N} \text{Model}(Y_i, Z_i|\theta).$$

Finally, the pseudo-likelihood function can be combined with the prior distribution for θ to form the posterior distribution

$$\pi(\theta|Y,Z) \propto \pi(\theta)\mathcal{L}(\theta|Y,Z).$$

Given the difficulty of the equations above, the reader may wonder how challenging it is to implement our method. In fact, the method is surprisingly easy to program because many statistical software packages, such as R, Python, and MATLAB, already possess density functions that can be modified to use the Epanechnikov kernel and Silverman's rule of thumb for bandwidth selection. Thus, in practice, implementing the method involves (1) calling the density function for each of the C alternatives and (2) scaling (i.e., multiplying) the resulting density values obtained by the number of times the corresponding alternative was chosen in the simulation. These scaled densities serve as Eq. 16.[5]

Simulation study

In this section, we will use the PDA method to fit the LBA model to choice response time data (i.e., data of mixed type) simulated from the LBA model. In a standard choice response time paradigm, following the presentation of a stimulus, subjects are instructed to make a decision between two or more alternatives. For example, in a random dot motion task, subjects are presented with a stimulus that contains many dots that randomly move in one direction or the other (e.g., left or right) within a fixed region of space. However, only a certain percentage of these dots move coherently toward the correct alternative, while the remaining dots move completely randomly. Subjects are asked to make a decision about what direction the majority of the dots are moving in, and the amount of time that passes from the onset of the stimulus to the decision that is elicited serves as the response time.

Suppose we wished to use likelihood-free techniques to fit a simulation-based model to data from a choice response time experiment. For reasons previously discussed, we cannot know what summary statistics will be sufficient for which parameters in the model, and so we will need to choose a set of summary statistics that we feel adequately characterize the relationship between the model parameters and the choice response time data. One option is to use the quantiles of both the correct and error response time distributions, along with the probability of a correct response. The use of quantiles to summarize response time distributions has a long history in mathematical psychology (e.g., Heathcote, Brown, & Mewhort, 2002; Luce, 1986; Van Zandt, 2000). Although it has been acknowledged that inference based on the quantiles is not equivalent to likelihood-based inference (Heathcote & Brown, 2004; Heathcote et al., 2002; Speckman & Rouder, 2004), this acknowledgment has not discouraged the use of quantiles when performing frequentist estimation. We speculate that there are two reasons for using the quantiles over the likelihood function. First, quantiles are dramatically more efficient because there are fewer densities to calculate when using quantiles. Second, quantiles are more robust to outliers when recovering point estimates of the parameter values (see Heathcote, Brown, & Cousineau, 2004; Speckman & Rouder, 2004). That is, outlying observations have less of an effect on parameter estimates when quantiles are used.

We argue that in the Bayesian context the use of quantiles is inappropriate because the quantiles do not provide a suitable approximation of the likelihood function. To examine this conjecture, we will use the computationally tractable LBA model to compare the true posterior estimates (obtained via standard Bayesian estimation techniques) with estimates obtained using the synthetic likelihood algorithm with quantiles as the summary statistics. If the estimates using the likelihood-free method and quantiles are different from the true posterior distributions (estimated using likelihood-informed methods), we would have evidence to support our claim that the chosen quantiles are not sufficient statistics for the LBA model parameters. We can also compare the true posterior estimates with the estimates obtained using our PDA method, which requires no summary statistics. If the estimates obtained using the PDA method are similar to the true posteriors, the results would speak to the utility and generality of our approach as a method for likelihood-free estimation.

The linear ballistic accumulator model

The LBA model provides a simple explanation of choice and response time (Brown & Heathcote, 2008). It eliminates many complexities assumed by previous models, such as competition between alternatives (e.g., Brown & Heathcote, 2005; Usher & McClelland, 2001), passive decay of evidence ("leakage"; e.g., Usher & McClelland, 2001), and even within-trial variability (e.g., Ratcliff, 1978; Stone, 1960). The model's simplicity allows closed-form expressions for the first passage time distributions for each accumulator. With these equations, one can specify the likelihood function for the model parameters, which has been instrumental in the LBA model's success (e.g., Donkin, Averell, Brown, &

[5] We have made code available online that constructs an SPDF estimate for all three types of data discussed above.

Heathcote, 2009; Donkin, Brown, & Heathcote, 2011; Donkin, Heathcote, & Brown, 2009; Forstmann et al., 2010; Forstmann et al., 2008; Forstmann et al., 2011).

The LBA model assumes that following the presentation of a stimulus, evidence for each response alternative is gathered ballistically until one of the alternatives reaches a threshold amount of evidence b. The rate of evidence accumulation d_c for the cth alternative is randomly sampled for each trial from a normal distribution with mean $v^{(c)}$ and standard deviation s. The model further assumes that each response alternative may start with an initial amount of evidence k_c, a value that is sampled randomly for each trial from a continuous uniform distribution on the interval [0, A]. Finally, the LBA model assumes that the observed response time distribution is the result of the decision process described above and a nondecision process (e.g., time required to initially process the stimulus and to physically elicit a response) captured by the parameter τ.

To perform our simulation study, we first generated 500 responses from the LBA model by setting the threshold b = 1.0, the upper bound of the start point A = 0.75, the drift rate for correct responses $v^{(C)}$ = 2.5, the drift rate for incorrect responses $v^{(I)}$ = 1.5, and the nondecision time τ = 0.2.[6] We conventionally set s = 1 to satisfy mathematical scaling properties of the model. The remaining parameter values were selected to be representative of subjects from experimental data (see, e.g., Turner, Sederberg, Brown, & Steyvers, 2013).

To fit the model in a Bayesian framework, we first specified uninformative uniform priors for each of the parameters, given by

$$b, A, v^{(C)}, v^{(I)}, \tau \sim CU(0, 10),$$

where $CU(a, b)$ denotes the continuous uniform distribution with lower bound a and upper bound b. We specified completely uninformative priors for the simulation study so that any accuracy observed in the posterior estimates would be entirely due to a method's ability to accurately estimate the likelihood function, and not due to the influence of the prior. In our real-world examples below, we will specify informative priors.

Estimating the posterior

To estimate the posterior distribution of $\theta = \{b, A, v^{(I)}, v^{(C)}, \tau\}$, we use three different approaches. The first method is the standard approach that makes use of the likelihood function (see Donkin, Averell, et al., 2009; Donkin, Heathcote, & Brown, 2009; Turner, Sederberg, et al., 2013). The second method is the PDA method for mixed data types as described above. The final method is the synthetic likelihood algorithm (Wood, 2010), which requires the specification of a set of summary statistics $S(\cdot)$, and despite selecting plausible statistics, we will show the dangers of reducing the observed data to a set of summary statistics that are not necessarily sufficient.

As is shown in Turner, Sederberg, et al. (2013), the parameters of the LBA model are generally highly correlated, which can make conventional sampling algorithms such as Markov chain Monte Carlo (MCMC; Robert & Casella, 2004) inefficient to use. As such, for each of the methods below, we used a genetic algorithm called differential evolution (DE) with MCMC (DE-MCMC; ter Braak, 2006; Turner, Sederberg, et al., 2013). DE-MCMC is a population Monte Carlo algorithm that generates proposals on every trial on the basis of the information learned in the current estimate of the posterior. The communication between the "chains" in the algorithm allows DE-MCMC to generate proposals to match the shape of the posterior, regardless of how correlated the parameters may be.

Likelihood-informed method The first method uses the likelihood function, which was derived from the defective probability density functions provided in Brown and Heathcote (2008). The estimates obtained from this method will be used to evaluate the accuracy of the remaining methods. For brevity, we do not report our methods for fitting the model here, but interested readers can consult Turner, Sederberg, et al. (2013) for an application of DE-MCMC or Donkin, Heathcote, and Brown (2009) and Donkin, Averell, et al. (2009) for applications of the program WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000).

Synthetic likelihood Wood (2010) proposed the synthetic likelihood algorithm as a method for likelihood-free parameter estimation that, unlike previous likelihood-free algorithms, does not require the use of error terms that can sometimes produce inaccurate posteriors. To implement the synthetic likelihood algorithm, we first generate a proposal value θ^* and simulate J new data sets of the same size and design as the observed data so that $X = \{X_1, X_2, \ldots, X_J\}$, where $X_j = \{X_{j,1}, X_{j,2}, \ldots\}$. For the jth simulated data set, we then compute a vector of summary statistics $S^{(j)}(X_j) = \{S_1^{(j)}(X_j), S_2^{(j)}(X_j), \ldots, S_M^{(j)}(X_j)\}$. The summary statistics across the J simulated data sets are then used to compute the mean vector $\widehat{\mu}_\theta$, so that

$$\widehat{\mu}_\theta = \frac{1}{J}\sum_{j=1}^{J} S^{(j)}(X_j),$$

and the covariance matrix $\widehat{\Sigma}_\theta = QQ^T/(J-1)$, where

$$Q = \left[S^{(1)}(X_1) - \widehat{\mu}_\theta, S^{(2)}(X_2) - \widehat{\mu}_\theta, \ldots, S^{(J)}(X_J) - \widehat{\mu}_\theta\right].$$

The same summary statistics are computed for the observed data, which we denote *S(Y)* (i.e., without the super

[6] The drift rates are arbitrarily labeled as "correct" and "incorrect" here, but we could have just as well labeled them "left" and "right" responses.

script index), and we assume that they have the parametric form

$$S(Y) \sim \mathcal{N}(\mu_\theta, \Sigma_\theta),$$

where $\mathcal{N}(a, b)$ denotes the normal distribution with mean a and variance covariance matrix b. Finally, using the normality assumption and the central limit theorem, the log synthetic likelihood function is given by

$$\mathcal{SL}(\theta|Y) = -\frac{1}{2}\left(S(Y) - \widehat{\mu}_\theta\right)^T \widehat{\Sigma}_\theta^{-1} \left(S(Y) - \widehat{\mu}_\theta\right) - \frac{1}{2} log\left|\widehat{\Sigma}_\theta\right|. \quad (17)$$

To implement the synthetic likelihood algorithm for response time data, we must first choose summary statistics $S(\cdot)$ that will adequately characterize the relationship between the model parameters and the observed data. To fit the LBA data, we chose $S(\cdot)$ to be the quantiles {0.1,0.3,0.5,0.7,0.9} for both the correct and incorrect response time distributions, along with the proportion of responses in each choice. Thus, for a given response time data set Y and Z, we summarized the data by computing the vector $S(Y, Z)$ comprising 11 statistics: 5 quantiles for each of the two choices and 1 proportion of total responses to alternative 1, without loss of generality for the two-choice task.

While the synthetic likelihood approach has certain advantages over other likelihood-free algorithms, the disadvantage of using this approach is that it is more computationally costly. For each proposal, we generated 100 data sets—each of size $N = 500$—which were then used to evaluate the synthetic likelihood shown in Eq. 17. Thus, we performed a total of 50,000 model simulations per proposal to evaluate the synthetic likelihood.[7] The major difference between the algorithms is that the synthetic likelihood algorithm requires sufficient summary statistics, whereas the PDA method does not.

Probability density approximation The final method we used to estimate the posterior distribution was the PDA method for data of mixed type. For each proposal, we simulated the model 10,000 times to form a stable SPDF, given by Eq. 16. The bandwidth parameters h were calculated for each proposal by means of Eq. 14. To increase the accuracy of the Epanechnikov kernel density approximation, we applied a log transformation to the simulated response times, which helped produce more normally-distributed data. As described above, we scaled the approximate density functions for each choice by the corresponding proportion of total responses out of the J simulations to determine the defective distribution for each choice. Once the SPDF was constructed, we approximated the likelihood function by evaluating

$$\mathcal{L}(\theta|Y, Z) = \prod_{i=1}^{N} \text{Model}(Y_i, Z_i|\theta) = \prod_{i=1}^{N} f_{n(z_i)}\left(Y_i|X^{(Z_i)}\right).$$

Results

For each of the four different likelihood evaluation methods, we implemented a DE-MCMC sampler, with 24 chains for 4,900 sampling iterations following 100 burn-in iterations. We set the within-group migration probability to .05, and for each DE proposal, we randomly sampled the scaling factor $\gamma \sim CU[0.5, 1]$. Additional implementation details of the sampler can be found in Turner, Sederberg, et al. (2013).

Figure 3 shows the estimated posterior distributions obtained by the PDA method (top row) and the synthetic likelihood method (bottom row). The columns of Fig. 3 correspond to the threshold parameter b, start point upper bound parameter A, drift rate for correct responses $v^{(C)}$, the drift rate for incorrect responses $v^{(I)}$, and the nondecision time parameter τ. In each panel, the true estimated posterior distribution (i.e., the likelihood-informed estimate) is shown as the density function, and the true parameter value used to generate the data is shown as the vertical dashed line.

The figure illustrates two important results. First, the estimates obtained using the PDA method are similar to the true estimates. Because the PDA method is a general technique that uses the entire data set, we can be sure that the accuracy of the estimates will depend only on the quality of the kernel density estimate. The second important result is that the estimates obtained using the synthetic likelihood algorithms are inaccurate. The most likely explanation for this inaccuracy is that the summary statistics we used are simply not sufficient for the parameters of the LBA model. Another possible explanation is that the joint multivariate normality assumption is not satisfied. However, for this example, we suspect that the distributional assumption is appropriate, because the statistics we calculated were far enough from their respective boundaries (i.e., a boundary at zero) and the sample size was large enough ($J = 100$, resulting in 50,000 model simulations per proposal) to satisfy the joint multivariate normality assumption. Furthermore, Q-Q plots of the marginal distribution of the statistics appeared to satisfy the normality assumption.

Although the combination of quantiles and response proportion has been used extensively in psychology, in our example we have found clear evidence that the use of quantiles does not result in accurate estimates of the posterior distribution. One of the reasons quantiles are used so frequently in more traditional fitting techniques (e.g., maximum likelihood estimation) is because they are resistant

[7] As we will explain below, when using the synthetic likelihood method, we generated 40,000 samples per proposal more than when using the PDA method.

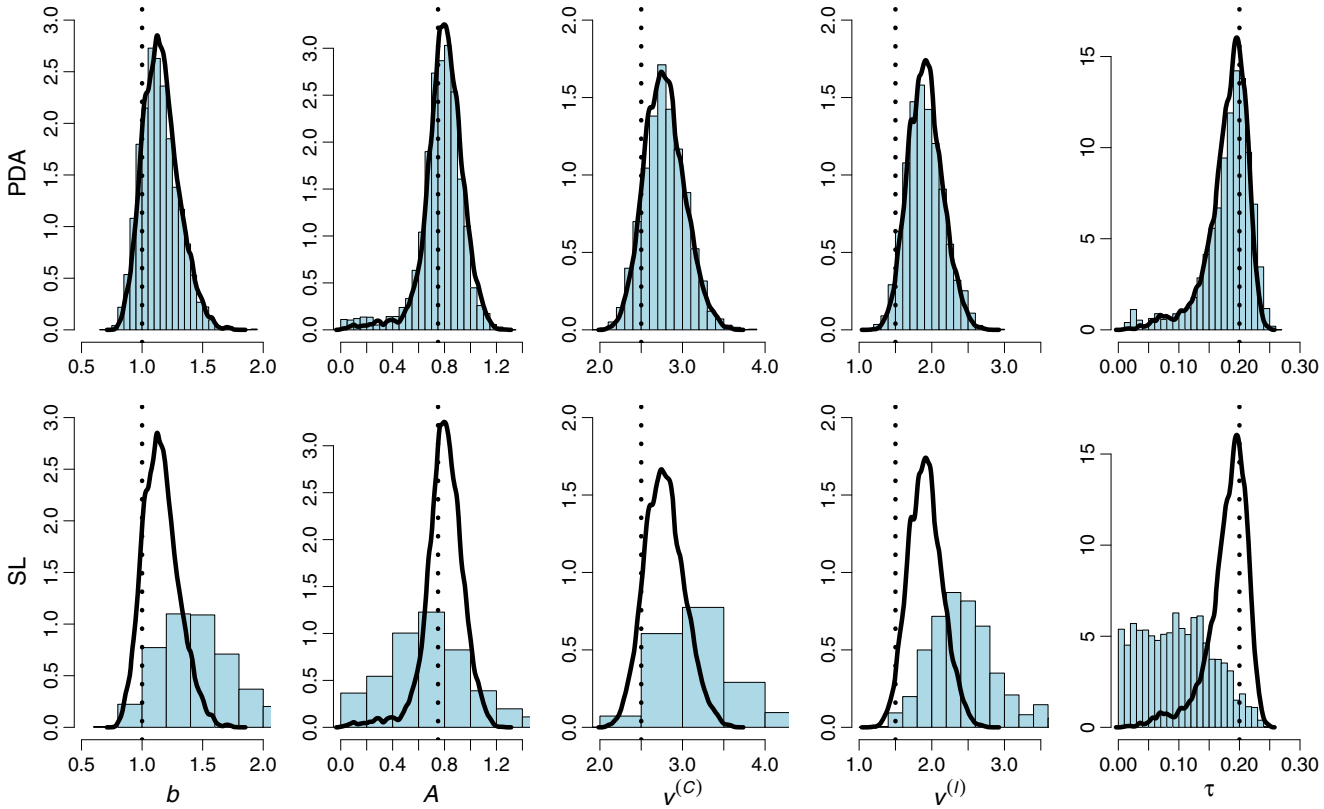

Fig. 3 Estimated marginal posterior distributions obtained using the probability density approximation (PDA) method (top row) and the synthetic likelihood algorithm (SL; bottom row). In each panel, the true estimate of the posterior distribution (i.e., the likelihood-informed estimate) is shown as the black density, and the true parameter value used to simulate the data is shown as the vertical dashed line

to outliers; however, in this case, using quantiles masks errors in the fits between the simulated and actual response time distributions and allows proposals to be accepted that should be rejected.

Real-world example 1: error-correcting criterion model

The classical SDT model (see above) posits the existence of two representations of sensory affect and a criterion that is fixed for the duration of the experiment (Green & Swets, 1966; Macmillan & Creelman, 2005). These assumptions, while convenient for SDT's use as a measurement tool, are not well-suited for SDT's use as a process model of perceptual decision making (e.g., Balakrishnan, 1998a, 1998b, 1999). Specifically, the classic SDT model does not provide any mechanistic account of how stimulus representations are established, nor does it explain how the criterion is placed or adjusted as a function of accuracy, payoffs, or fluctuations in the stimulus stream. Over the years, many alternatives have been proposed to extend SDT-as-model to account for basic experimental manipulations, stimulus properties, feedback, probability matching, and adjustments following responses and the accuracy of those responses (Atkinson & Kinchla, 1965; Benjamin, Diaz, & Wee, 2009; Dorfman & Biderman, 1971; Dorfman, Saslow, & Simpson, 1975; Erev, 1998; Kac, 1962, 1969; Kubovy & Healy, 1977; Lee & Dry, 2006; Mueller & Weidemann, 2008; Treisman & Williams, 1984; Turner et al., 2011). While these models take wildly different approaches to the same problem, they all propose a dynamic adjustment procedure where the stimulus representations take a new form on every trial. The dynamic process assumed by most of these models makes model evaluation and parameter inference difficult.

In this section, we focus on a simple mathematical model for criterion adjustment in a perceptual decision task. The model is a generalization of the error-correcting criterion (ECC) model proposed by Kac (1962, 1969), and is analytically tractable (see Dorfman & Biderman, 1971, for derivations). Fitting the ECC model to data using both a likelihood-informed and likelihood-free method enables us to further verify our PDA method on a model that does not assume that responses are a realization of an i.i.d. process, as in classical SDT. We will fit a hierarchical version of the model to experimental data that uses a within-subjects signal frequency manipulation. Thus, to capture the effects in the empirical data successfully, a model must be able to dynamically adapt its representations over the course of an experiment—a feat that standard SDT is unable to perform.

The experiment

The data from Experiment 2 of Turner et al. (2011) consisted of a simple signal detection task where subjects were told to diagnose patients from a community that had been overrun by an infectious disease. The stimuli consisted of a number ranging from 1 to 100, which represented the result of a blood test from a randomly selected patient. Patients were either sick or well, and subjects were asked to decide which patients should be treated and which should not. Subjects were told that a sick patient who was left untreated would die and a well patient who received treatment would also die. Subjects were told that well patients would have average blood assay values of 40 and sick patients would have average blood assay values of 60. Following each decision, feedback was provided about the mortality of the patient as a result of the diagnosis.

Subjects were assigned to one of three stimulus conditions, which differed by the type of distribution the stimuli were sampled from. For our purposes, we will focus only on the Gaussian condition, where stimuli were generated by sampling from one of two Gaussian distributions (i.e., the noise or signal distributions), which had means of 40 and 60, respectively, and a common standard deviation of 6.67. Sixteen subjects participated in this condition. Subjects completed five blocks of 100 trials each. Between each block, an unannounced change in the frequency of sick patients occurred. For all subjects, the frequency of sick patients in the first block was .5. In the second block, the frequency shifted to .8. In the third block it shifted back to .5, then to .2 in the fourth block. Finally, the frequency returned to .5 in the fifth and last block.

The model

We let T_i denote the true value of the stimulus class on trial i so that $T_i = 1$ indicates a trial on which a noise stimulus was presented and $T_i = 2$ indicates a trial on which a signal stimulus was presented. We let R_i denote the response on trial i, where $R_i = 1$ and $R_i = 2$ indicate "noise" and "signal" responses, respectively. On each trial, the presented stimulus value S_i is drawn from one of two distributions $f(S|T)$, depending on the value of T_i. For our data, the stimulus distributions are Gaussian with means of $\mu = \{40, 60\}$ and standard deviations of $\sigma = \{6.67, 6.67\}$. Thus, on each trial, $S_i \sim \mathcal{N}\left(\mu_{T_i}, \sigma_{T_i}\right)$.

The model assumes that an observer begins on trial 1 with an initial criterion of $\theta^{(1)}$, and we will denote subsequent locations of this criterion on trial i as $\theta^{(i)}$. As in classical SDT (Green & Swets, 1966), a "noise" response (i.e., $R_1 = 1$) is elicited if $S_i \leq \theta^{(i)}$, whereas a "signal" response (i.e., $R_1 = 2$) is elicited if $S_i > \theta^{(i)}$. As in the experiment, feedback is given following the response, and observers update the location of their criterion so that

$$\theta^{(i+1)} = \theta^{(i)} + \delta_i \Delta_{T_i,R_i},$$

where Δ_{T_i,R_i} is an element of an updating matrix and δ_i is a transformation of the stimulus class variable:

$$\delta_i = \begin{cases} -1 & \text{if } T_i = 2 \\ 1 & \text{if } T_i = 1 \end{cases}.$$

The individual elements of the updating matrix determine how the model behaves as a function of its response and its accuracy goals (Dorfman & Biderman, 1971). For example, Kac (1962) assumed that $\Delta_{1,1} = \Delta_{2,2} = 0$ and that $\Delta_{1,1} = \Delta_{2,1} = \Delta^*$, where Δ^* was free to vary. Under these specifications, Kac's model assumes that no criterion updating occurs when a response is correct, but when a response is incorrect, updating does occur, and by the same amount for both types of errors. The model we will consider here assumes that $\Delta_{1,1} = \Delta_{2,2} = \Delta^{(C)}$ and $\Delta_{1,2} = \Delta_{2,1} = \Delta^{(I)}$, where $\Delta^{(C)}$ and $\Delta^{(I)}$ denote the degree of criterion change following correct and incorrect responses, respectively. Thus, the model we investigate here consists of three parameters per subject: $\Delta^{(C)}$, $\Delta^{(I)}$, and $\theta^{(1)}$.

To extend the model to a hierarchical design, we must make an assumption about how each of the three parameters is distributed across subjects. We assume that each of the model parameters for the jth subject is normally distributed, so that

$$\begin{aligned} \Delta_j^{(C)} &\sim \mathcal{N}\left(\Delta_\mu^{(C)}, \Delta_\sigma^{(C)}\right) \\ \Delta_j^{(I)} &\sim \mathcal{N}\left(\Delta_\mu^{(I)}, \Delta_\sigma^{(I)}\right), \text{ and} \\ \theta_j^{(1)} &\sim \mathcal{N}\left(\theta_\mu, \theta_\sigma\right). \end{aligned}$$

The parameters governing the distribution of the subject-specific parameters (e.g., $\Delta_\mu^{(C)}$ above) are called *hyper parameters*, and they capture patterns in the data at the group level. To complete our hierarchical Bayesian model, we must specify a prior distribution for each of these hyper parameters. Although we had some information about the criterion adjustment parameters from Dorfman and Biderman (1971), we remained cautious when specifying the priors for the hyper parameters, using only mildly informative priors:

$$\begin{aligned} \Delta_\mu^{(C)}, \Delta_\mu^{(I)} &\sim \mathcal{N}(0, 10), \text{ and} \\ \Delta_\sigma^{(C)}, \Delta_\sigma^{(I)} &\sim \Gamma^{-1}(4, 15), \end{aligned}$$

where $\Gamma^{-1}(a,b)$ denotes the inverse gamma distribution with shape parameter a and scale parameter b. Our choice for the parameters of the inverse gamma reflects our uncertainty about the degree of individual differences in our data. A shape of 4 and scale of 15 produce a distribution with a mean of 5, a standard deviation of 3.45, and a 95 % confidence interval of (1.71, 13.86). Because we suspected a high degree of

variability in the criterion adjustment parameters, we believed that this limited specification was appropriate for our data.

For the initial criterion hyper parameters θ_μ and θ_σ, we centered the prior distribution for the mean θ_μ on 50, with a small (i.e., relative to the stimulus distributions) degree of variability. We made this decision on the basis of the experimental instructions, which told the subjects the means of the two types of stimuli. We used the same prior as above for θ_σ because, while we suspected far less variability in this parameter, the scales were different enough to warrant a slightly ambiguous prior. Thus, we specified the priors as

$$\theta_\mu \sim \mathcal{N}(50, 10), \text{ and}$$
$$\theta_\sigma \sim \Gamma^{-1}(4, 15).$$

The distributions assumed in the prior specifications above were primarily made out of computational convenience. Specifically, the priors for the hyper parameters, along with the assumptions about the distribution of subject-specific parameters, facilitate Gibbs sampling because the conditional distributions for each of the hyper parameters can be derived (see Gelman et al., 2004).

Results

For this model, the likelihood function is easy to derive (see Dorfman & Biderman, 1971), and so we can compare the estimated posterior distribution using the PDA method with the estimated posterior distribution obtained using the (true) likelihood function. To implement the PDA method, we simulated the model 1,000 times per proposal using the same stimuli that were presented to the subjects. Thus, our simulated data consisted of a (500 × 1,000) matrix of signal and noise responses for each subject. The data for a given subject consists of 500 signal or noise responses, and so the PDA method for discrete data types was used for each of the 500 trials. To construct the SPDF, we first tabulated the number of signal and noise responses for each row of the simulated data matrix and divided this number by 1,000 (i.e., the number of model simulations per trial). The resulting values served as an estimated probability of a signal and noise response on trial i under the proposed parameter values (see Eq. 8). We then multiplied the predicted probabilities corresponding to the observed response for each of the 500 responses for a given subject together, forming the pseudo-likelihood in Eq. 5.

We ran both algorithms for 2,500 iterations following a burn-in period of 500 iterations using 16 chains, resulting in 40,000 samples of the joint posterior distribution. We used Gibbs steps to update the hyper parameters as discussed in Turner and Van Zandt (2013) and DE-MCMC to update the lower-level parameters for each subject (Turner & Sederberg, 2012; Turner, Sederberg, et al., 2013). Convergence of the chains was assessed using the CODA package in R (Plummer, Best, Cowles, & Vines, 2006), and the mixing properties of the chains were assessed through visual inspection.

Having fit the model using both likelihood-informed and likelihood-free (i.e., PDA) algorithms, we can now compare the estimated posterior distributions obtained from each method. Figure 4 shows the estimated marginal posterior distributions obtained using the PDA method (histograms) and the true likelihood (densities) for a randomly selected subject (i.e., subject 8). The figure shows that all of the parameters for this subject were adequately recovered, although there still exists some small amount of error in the estimated posteriors. Comparing the estimates in the remaining 15 subjects, we observed similarly close alignments of the estimated posterior distributions.

We can also examine the estimates obtained using the PDA method at the group level. If the PDA method produced estimates that systematically mismatched the true posteriors at the subject level, the errors would propagate to the hyper parameter estimates. Thus, it is important that the PDA method matches the true posterior at both the group and subject levels. Figure 5 shows the estimated marginal posterior distributions for each of the hyper parameters of the ECC model using the PDA method (histograms) and the true likelihood (densities). The figure shows that the two methods produce virtually identical posterior estimates for the hyper parameters.

On a modeling level, Fig. 5 shows that the posterior estimates for the hyper mean parameter for correct adjustments $\Delta_\mu^{(C)}$ has a mean of 0.04, whereas the hyper mean parameter for incorrect adjustments is larger with an estimate of 0.33.[8] Not surprisingly, the hyper mean parameter θ_μ is centered at 50.01, a value that is consistent with the experimental instructions. The variability of the adjustment parameters Δ_σ is small and similar across adjustment types, and the variability in the initial criterion location is also small.

In this section, we have shown that the PDA method can accurately recover the posterior distributions for a hierarchical version of the ECC model. However, the model fit performed in this section is not as easy as it may seem. Because the ECC model assumes a dynamic adjustment of the criterion that is based on both the type of stimulus presented and the response, to correctly estimate the parameters we must evaluate the model predictions relative to the data on every trial. For least squares estimation, one might naïvely compare the simulated model predictions with the observed responses by way of a root mean squared error. However, this metric would not produce accurate parameter estimates, because it implicitly assumes that the discrepancies between the simulated and

[8] The estimates differed by less than 0.001 for both parameters across the two methods.

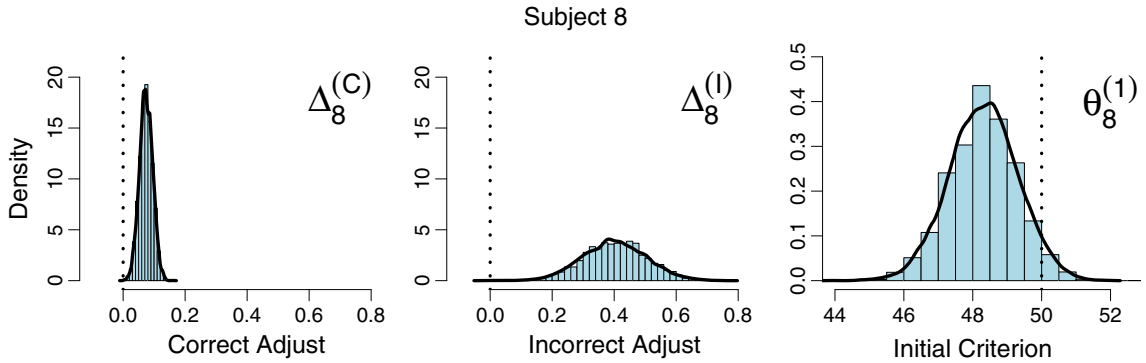

Fig. 4 Estimated marginal posterior distributions for each of the three subject-specific parameters of the error-correcting criterion model for a randomly selected subject (subject 8) obtained using the probability density approximation method (histograms) and the true likelihood (densities). The dashed vertical lines in the left and middle panels represent the parameter values that produce no criterion adjustment. The dashed vertical line in the right panel represents the location of the optimal initial criterion

observed data are equally diagnostic of a bad fit on every trial. Such a procedure would ignore the mismatch between an adjustment following an incorrect or correct response, which would produce inaccurate parameter estimates. Thus, methods for fitting models like the ECC model must respect the discrepancy between simulated and observed data on each trial, because such models do not make an i.i.d. distributional assumption about the responses. Furthermore, methods like the synthetic likelihood algorithm could not be applied to the responses on each trial because they are discrete measures, making the multivariate normal assumption inappropriate.

Real-world example 2: hierarchical LBA and LCA models

For our final example, we will investigate how our method performs on experimental data consisting of mixed measures (i.e., continuous and discrete). Data of mixed type present perhaps the most difficult challenge for our method because the distribution of continuous measures (e.g., response times) must be estimated accurately for each discrete measure (e.g., each response alternative). In addition, the data we will examine consist of three speed emphasis conditions per subject, which requires that we properly estimate a total of

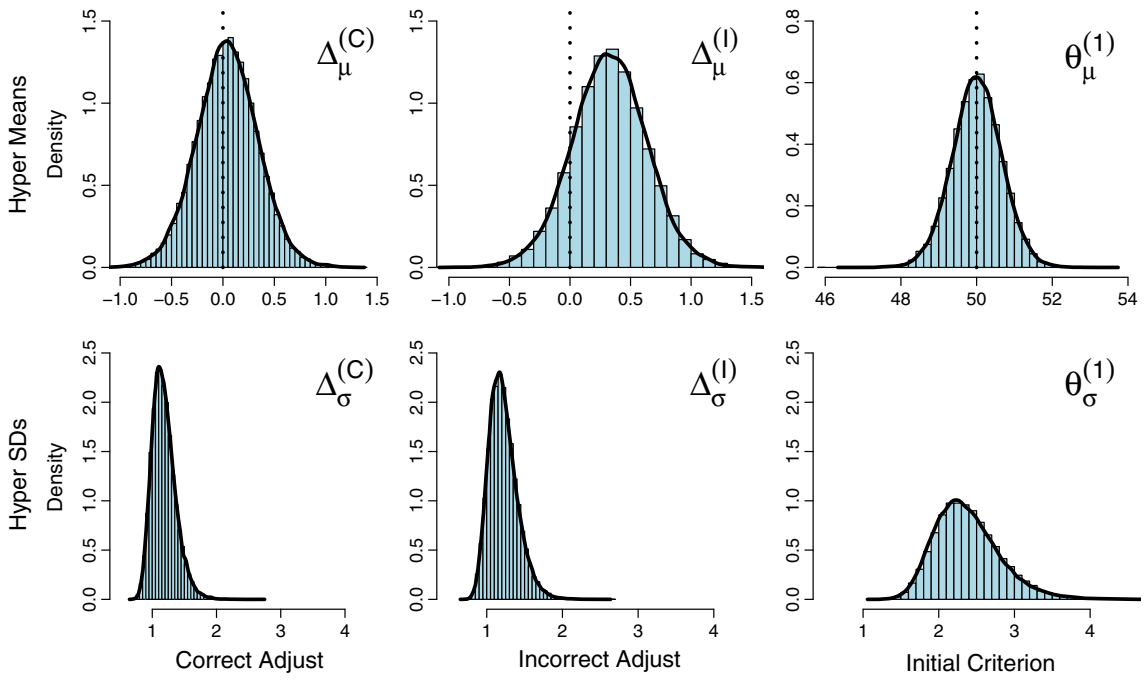

Fig. 5 Estimated marginal distributions of each of the hyper parameters of the error-correcting criterion model obtained using the probability density approximation method (histograms) and the true likelihood (densities). The dashed vertical lines in the top left and top middle panels represent the parameter setting where no adjustment is made. The dashed vertical line in the top right panel represents the location of the optimal initial criterion

six response time distributions. If one or more of these distributions is not estimated correctly, the error in the kernel estimate propagates through to the estimated posterior distributions, which would result in inaccurate estimates.

In this section, we will investigate two models of choice response time by fitting them to data. The first model is the LBA model, which was used above in the simulation study. Donkin, Averell, et al. (2009) provided a method for fitting the LBA model hierarchically to data within the program WinBUGS (Lunn et al., 2000). Despite this advancement, successfully fitting the LBA model hierarchically to data is a nontrivial problem, due to the extreme correlations between the model parameters (Turner, Sederberg, et al., 2013). Fitting a hierarchical LBA model allows us to further verify that our method works for data of mixed type. The second model is the LCA model, which has never been examined in the Bayesian context and has never been fit hierarchically to data.

The experiment

We chose data presented in Forstmann et al. (2011), which consisted of 20 young subjects and 14 elderly subjects. The study was a moving dots task where subjects were asked to decide whether a cloud of semirandomly moving dots appeared to move to the left or to the right. Subjects indicated their response by pressing one of two spatially compatible buttons with either their left or right index finger (e.g., a left button for a left index finger). Before each decision trial, subjects were instructed to respond quickly (the speed condition), accurately (the accuracy condition), or at their own pace (the neutral condition). Following the trial, subjects were provided feedback about their performance. In the speed and neutral conditions, the young subjects were told that their responses were too slow whenever they exceeded a response time of 400 and 750 ms, respectively. In the accuracy condition, subjects were told when their responses were incorrect. Each young subject completed 840 trials, equally distributed over the three conditions.

In this section, we will fit hierarchical versions of both the LBA and LCA models to a random subset of the data. Due to computational demands and for illustrative purposes, we randomly select 4 young subjects (i.e., subjects 1, 3, 7, and 11). For the LBA model, we again use both likelihood-informed and PDA methods to further verify our method, whereas for the LCA model, we use only the PDA method because, for reasons discussed in the Introduction, the LCA model does not have a tractable likelihood. We now discuss the details of each model in turn.

The linear ballistic accumulator model

Because we are now fitting the LBA model to experimental data, we will use more informative priors. First, we denote the threshold parameters for the accuracy (A), neutral (N), and speed (S) emphasis conditions as $b^{(A)}$, $b^{(N)}$, and $b^{(S)}$, respectively. We will reuse the notation for the remaining parameters from the simulation study above. Because this is a hierarchical model, we subscript each of the model parameters with a j to indicate that they are exclusive to the jth subject. Similarly, we subscript the hyper parameters with either a μ or a σ to indicate that they are either hyper mean or hyper standard deviation parameters, respectively. As in the ECC model above, we assume that each of the subject-specific parameters is a random perturbation of a common distribution according to the following specification:

$$
\begin{aligned}
log\left(b_j^{(k)}\right) &\sim \mathcal{N}\left(b_\mu^{(k)}, b_\sigma^{(k)}\right)\\
log\ \left(A_j\right) &\sim \mathcal{N}\left(A_\mu, A_\sigma\right) I\left(-\infty, min\left[b_j\right]\right)\\
log\left(v_j^{(c)}\right) &\sim \mathcal{N}\left(v_\mu^{(c)}, v_\sigma^{(c)}\right), \text{ and}\\
log\left(\tau_j\right) &\sim \mathcal{N}\left(\tau_\mu, \tau_\sigma\right) I\left(-\infty, log\left(min\left[RT_j\right]\right)\right),
\end{aligned}
$$

where $k \in \{A, N, S\}$, $c \in \{C, I\}$, RT_j is the set of response times for the jth subject and $I(a, b)$ is the indicator function that returns a zero for values that are outside of the interval (a, b) and a one otherwise. We use the indicator function as a simple way to censor proposals so that they obey certain restrictions. For example, the upper bound of the start point A must always be less than the threshold. Because we constrain the model to have the same start point for each speed emphasis condition, A must be less than the smallest of the three threshold parameters.

For the hyper parameters, we used informative priors similar to what was used in the ECC model:

$$
\begin{aligned}
b_\mu^{(k)}, A_\mu &\sim \mathcal{N}(1.5, 0.8)\\
v_\mu^{(c)} &\sim \mathcal{N}(0.75, 0.5)\\
\tau_\mu &\sim \mathcal{N}(-1.0, 0.5).
\end{aligned}
$$

For the LBA model, we based our decisions for the priors on previous work (Donkin et al., 2011). We assumed an inverse gamma prior distribution with shape parameter 4 and scale parameter 10 for each of the hyper standard deviation parameters. As in the simulated example, we conventionally set $s = 1$.

The leaky competing accumulator model

For the LCA model, we denote the rate of accumulation for the cth accumulator as ρ_c, the lateral inhibition parameter as β, the leakage parameter as κ, and the degree of noise in the accumulation process as ν, which, when simulated, is drawn from a normal distribution with a mean of zero and standard deviation ξ.[9] The activation of the cth

[9] In other words, at each time step t in the evidence accumulation process, $\nu_t \sim \mathcal{N}(0, \xi)$.

accumulator in the model is represented by the stochastic differential equation

$$dx_c = \left(\rho_c - \kappa x_c - \beta \sum_{j \neq c} x_j\right)\frac{dt}{\Delta_t} + \nu_t\sqrt{\frac{dt}{\Delta_t}}$$
$$x_c \rightarrow max(x_c, 0),$$

where Δ_t is a time-constant parameter. Once the degree of evidence for any accumulator reaches a threshold α, the process is terminated, and a response is elicited. Similar to the LBA, the LCA model also assumes a nondecision time parameter, which we will again denote τ.

To satisfy mathematical scaling properties, we fixed the noise of evidence accumulation $\xi = 3.0$. We also fixed $dt = 0.01$ (with the unit of seconds) and $\Delta_t = 0.1$. We used similar notation and made similar assumptions about the priors as in the LBA model:

$$log\left(\alpha_j^{(k)}\right) \sim \mathcal{N}\left(\alpha_\mu^{(k)}, \alpha_\sigma^{(k)}\right)$$
$$log\left(\rho_j^{(c)}\right) \sim \mathcal{N}\left(\rho_\mu^{(c)}, \rho_\sigma^{(c)}\right)$$
$$\text{logit}\left(\kappa_j\right) \sim \mathcal{N}\left(\kappa_\mu, \kappa_\sigma\right)$$
$$\text{logit}\left(\beta_j\right) \sim \mathcal{N}\left(\beta_\mu, \beta_\sigma\right), \text{ and}$$
$$log\left(\tau_j\right) \sim \mathcal{N}\left(\tau_\mu, \tau_\sigma\right) I\left(-\infty, log\left(min\left[RT_j\right]\right)\right),$$

where $k \in \{A, N, S\}$, $c \in \{C, I\}$, and $\text{logit}(x) = log(x/(1-x))$. We use the logit function here to enforce the constraint that β, $\kappa \in [0.0, 1.0]$, to preserve the model's neurological plausibility. Specifically, values of β and κ greater than 1.0 would imply that the effect of lateral inhibition and/or leak would be greater than the activation of the accumulator itself, which would give rise to unstable network behavior and difficulty in interpreting parameters. For the hyper parameters, we used the following informative priors:

$$\alpha_\mu^{(k)} \sim \mathcal{N}(2.5, 0.5)$$
$$\rho_\mu^{(c)} \sim \mathcal{N}(0.20, 0.6)$$
$$\kappa_\mu, \beta_\mu \sim \mathcal{N}(0, 1.5)$$
$$\tau_\mu \sim \mathcal{N}(-1.0, 0.5).$$

Because the LCA has never been examined in the Bayesian context, we chose priors for κ_μ and β_μ that produced uniform priors on the interval [0, 1] for the subject parameters κ_j and β_j. As in the LBA model above, we assumed an inverse gamma prior distribution with shape parameter 4 and scale parameter 10 for each of the hyper standard deviation parameters.

Results

All implementation details of the PDA method were equivalent to those used for the simulation study above, except that the SPDF for each proposal was estimated with 30,000 model simulations, and we ran the algorithm for 5,000 iterations following a highly efficient burn-in period of 1,000 iterations for 36 chains (i.e., resulting in 180,000 samples; see Turner & Sederberg, 2012, for details on the burn-in period). Implementation of the PDA method required only two selections: the kernel density function and the number of model simulations per proposal. The kernel density function—the Epanechnikov kernel—was selected on the basis of minimization of the Kullback–Leibler divergence (see the simulation studies above). We chose the number of model simulations on the basis of parameter recovery on some preliminary simulation studies and by examination of the autocorrelation functions (Gelman et al., 2004). When the number of model simulations are too few, the autocorrelation functions tend to produce long tails, which is indicative of large rejection rates and chain dependency. Smaller rejection rates are obtained with a higher number of model simulations because, as the number of model simulations increases, the SPDF converges to the true PDF.

We present the results in two sections. First, we present the results of the hierarchical LBA model with an evaluation of the accuracy of the estimates obtained using our PDA method on a single subject. Second, we present the first Bayesian treatment of the LCA model on the same subject and provide some interpretations of the model parameters.

The LBA model

Figure 6 shows the estimated marginal posterior distributions for each of the LBA model parameters for Subject 1, who was randomly chosen. Figure 6 shows that the estimates obtained using the PDA method (histograms) closely match the estimated posterior distributions obtained using the likelihood-informed method (black densities) for each of Subject 1's parameters. The results for the other 3 subjects were similarly accurate.

On a modeling level, the estimated posterior distributions are consistent with standard speed–accuracy manipulations. Namely, the estimated response threshold in the accuracy condition is higher than in the speed condition, which implies that as the pressure to respond quickly increases (i.e., moving from accuracy to speed instructions), subjects require less accumulated information before eliciting a response, an adaptation that produces faster—yet less accurate—responses. The estimated posterior distribution for the correct drift rate is higher in magnitude than the incorrect drift rate, a result that is consistent with the raw accuracy of the choice response time data. See Turner, Sederberg, et al. (2013) for a more extensive examination of these data using a hierarchical LBA model (and the true likelihood equations).

As in the ECC model above, we also must examine the degree of match between the true and PDA-estimated

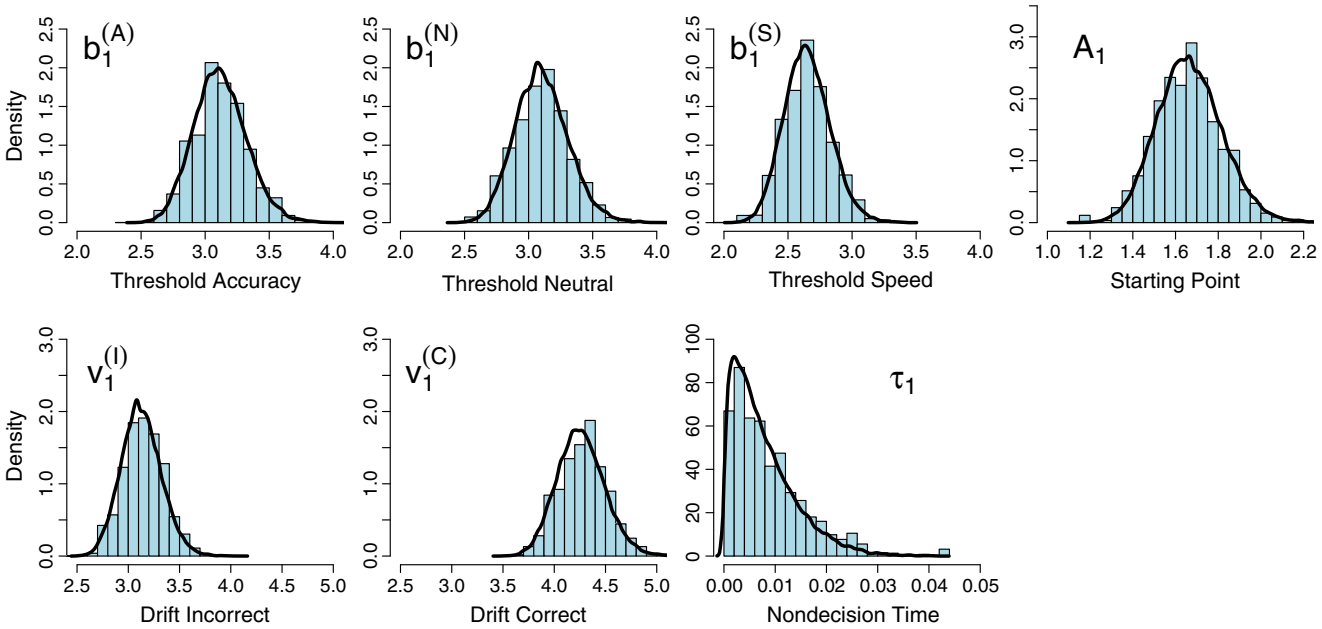

Fig. 6 Estimated marginal posterior distributions obtained using the true likelihood function (black densities) and the probability density approximation method (histograms) for each of the seven parameters in the linear ballistic accumulator model for Subject 1

posteriors at the group level (i.e., the hyper parameters). Figure 7 shows the estimated marginal posterior distributions obtained using the PDA method (histograms) and the likelihood-informed method (densities) for each of the seven hyper mean parameters. The figure shows that at the group level, the estimates obtained using the PDA method are close to the true estimated posteriors. The estimated posteriors were similarly accurate for the hyper standard deviation parameters. The results suggest that even for complicated designs involving multiple conditions and a hierarchical model, the

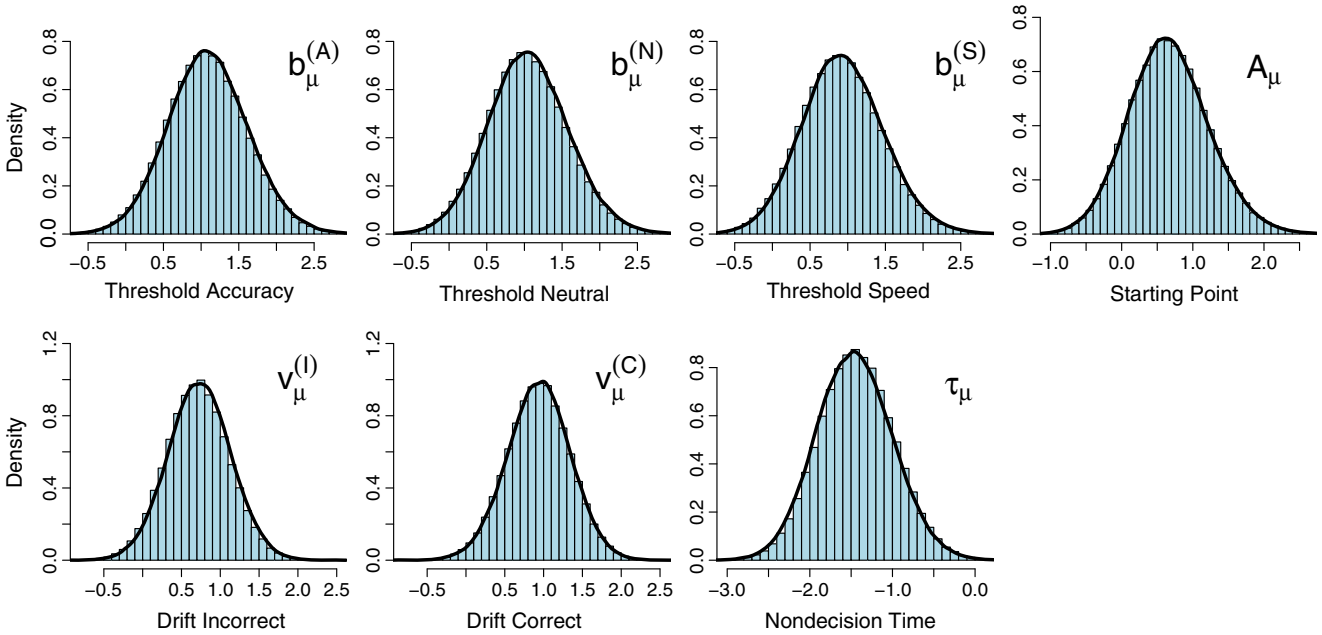

Fig. 7 Estimated marginal posterior distributions obtained using the true likelihood function (black densities) and the probability density approximation method (histograms) for each of the hyper mean parameters in the linear ballistic accumulator model. Note that the parameter estimates are shown on the log scale

PDA method can accurately recover the true posterior distribution for data of mixed type.

The LCA model

Because the LCA model is intractable, we have no method for obtaining the true posterior. Regardless, we can still examine the posterior distributions at the subject and group levels. Figure 8 shows the estimated marginal posterior distributions for each of the eight parameters for Subject 1. We chose to show Subject 1's parameter estimates as a comparison with the LBA model above. The figure shows a few interesting results. First, the estimates for the threshold parameter decrease as the pressure to respond increases, a result that is consistent with the LBA model. Second, the nondecision time parameter is appreciably different from the LBA model, which was centered near zero. Third, the drift rate for the correct response is considerably higher than for the incorrect response, a result that is also consistent with the LBA model. Finally, the estimates of leak and lateral inhibition are small in magnitude and approximately equal. After extensive analysis of the LCA model, Bogacz et al. (2006) concluded that when $\beta = \kappa$ and both parameters were large in magnitude, the LCA model performed optimally with respect to the rate of evidence accumulation (see also Bogacz, Usher, Zhang, & McClelland, 2012; Bogacz et al., 2007). Here, $\beta \approx \kappa$, but both parameters are near zero, so according to Bogacz et al. (2007), Subject 1 is performing suboptimally.

Figure 9 shows the estimated marginal posterior distributions for each of the hyper mean parameters in the model. At the hyper mean level, the leakage and lateral inhibition parameters tell a different story, as compared with the estimates for Subject 1 (see Fig. 8). Specifically, the lateral inhibition parameter is higher—having a mean of −2.537 on the logit scale (i.e., a mean of 0.073 on the standard scale)—than the leakage parameter, which has a mean of −3.101 on the logit scale (i.e., a mean of 0.043 on the standard scale). Thus, it would seem that for some of the subjects in these data, lateral inhibition played a slightly larger role than it did for Subject 1.

We can further examine the trade-off between the lateral inhibition parameter β and the leakage parameter κ. Figure 10 shows a scatterplot of the joint posterior distribution of (κ, β) for each of the 4 subjects in our data. In each panel, the diagonal line represents the setting $\beta = \kappa$. In the top left of each panel, we have calculated the probability that the lateral inhibition parameter is greater than the leakage parameter. Values near 1 indicate that lateral inhibition was a significantly stronger force than was leakage, whereas values near .5 indicate that neither force dominated the decision process. In each panel, the black dot represents the mean of each joint posterior. For Subjects 1 and 3 (top panel in Fig. 10), neither leakage nor lateral inhibition played a role in the decision process. However, for Subjects 7 and 11 (bottom panel in Fig. 10), lateral inhibition played a slightly larger role in the decision. These subjects can be described as exhibiting an "inhibition dominant" pattern (Tsetsos et al., 2011), which tends to create larger separations in the amounts of accumulated evidence between the (two) alternatives. The estimated posterior distributions for subjects 7 and 11 are indicative of suboptimal evidence integration according to Bogacz et al. (2006).

Figure 10 also suggests that some caution must be taken when classifying subjects as either leak or inhibition dominant. If we were to classify the subjects into groups in a

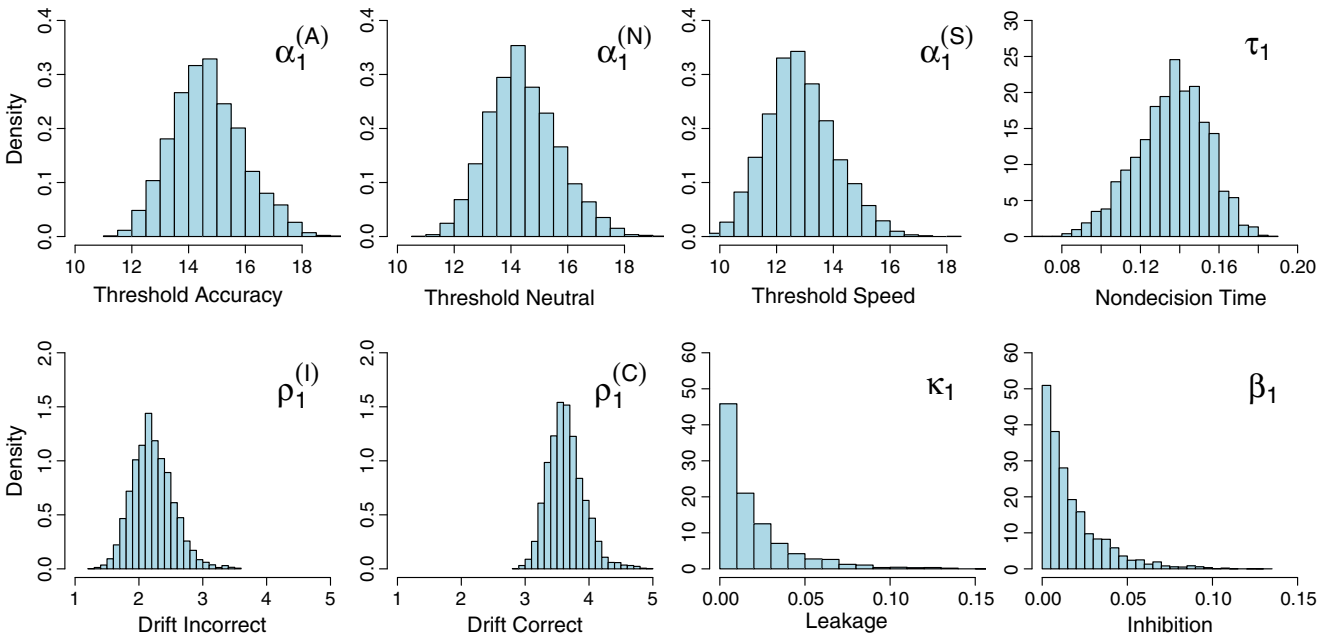

Fig. 8 Estimated marginal posterior distributions for each of the eight parameters in the leaky competing accumulator model for Subject 1

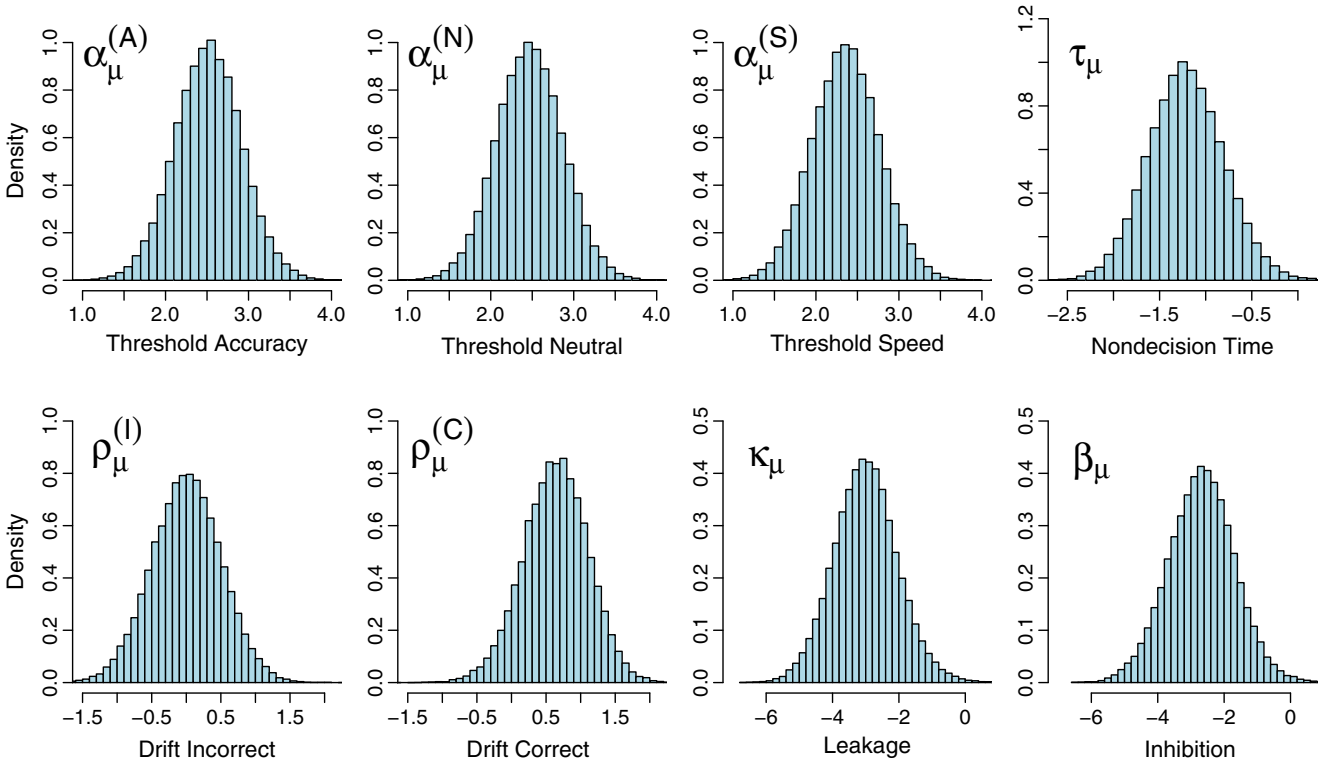

Fig. 9 Estimated marginal posterior distributions for each of the hyper mean parameters in the leaky competing accumulator model. Note that thresholds, drift rates, and the nondecision time parameter are shown on the log scale and leakage and lateral inhibition are shown on the logit scale

frequentist setting, we might use a best-fitting parameter estimate as an indication of which response strategy the subject was using. For example, a subject would be classified as inhibition dominant when $\widehat{\beta} > \widehat{\kappa}$. However, such a classification rule ignores the uncertainty in the parameter estimates. Figure 10 shows that for these 4 subjects, summarizing a subject's performance as one type or the other may be less clear than a simple discrete measure would suggest.

General discussion

In this article, we have presented a new method for likelihood-free posterior estimation that has considerable advantages over other current likelihood-free techniques. First, the PDA method is nonparametric, and it does not make any restrictive assumptions about the distribution of the data or summary statistics. Second, our method does not require the use of tolerance thresholds to evaluate the performance of individual proposals. Third, and perhaps most important, the PDA method is the first likelihood-free Bayesian method that does not require the use of summary statistics. Therefore, the PDA method is immune to the error introduced from using a set of summary statistics that are not sufficient.

After presenting the technical details of our method, we validated the method by way of a simulation study using the LBA model. The LBA model is a mathematically convenient model of choice response time with a tractable likelihood function. Having the likelihood function allowed us to compare the estimates obtained using our PDA method with the estimates obtained using standard Bayesian methods. In addition, we used a synthetic likelihood algorithm with conventional summary statistics (i.e., the quantiles and proportion correct). Figure 3 showed that the estimates obtained using the PDA method were excellent approximations to the true posterior, whereas the estimates obtained using the synthetic likelihood algorithm were inaccurate. We concluded that the failure of the synthetic likelihood approach was due to the use of summary statistics that were not sufficient for the parameters of the LBA model. These results reveal that quantiles are not appropriate for Bayesian estimation of the LBA model. At present, the relative merits of quantile-based and likelihood-based methods for parameter estimation in the frequentist setting remain unclear.

After demonstrating that the PDA method provided accurate estimates of the posterior distribution, we used the method to fit three hierarchical models to empirical data. The first model was the ECC model, which unlike other models of signal detection, does not assume that an observer's responses are independent or identically distributed across trials. To fit the model, we constructed an SPDF estimate for every trial of the experiment. While this is a challenging problem for least

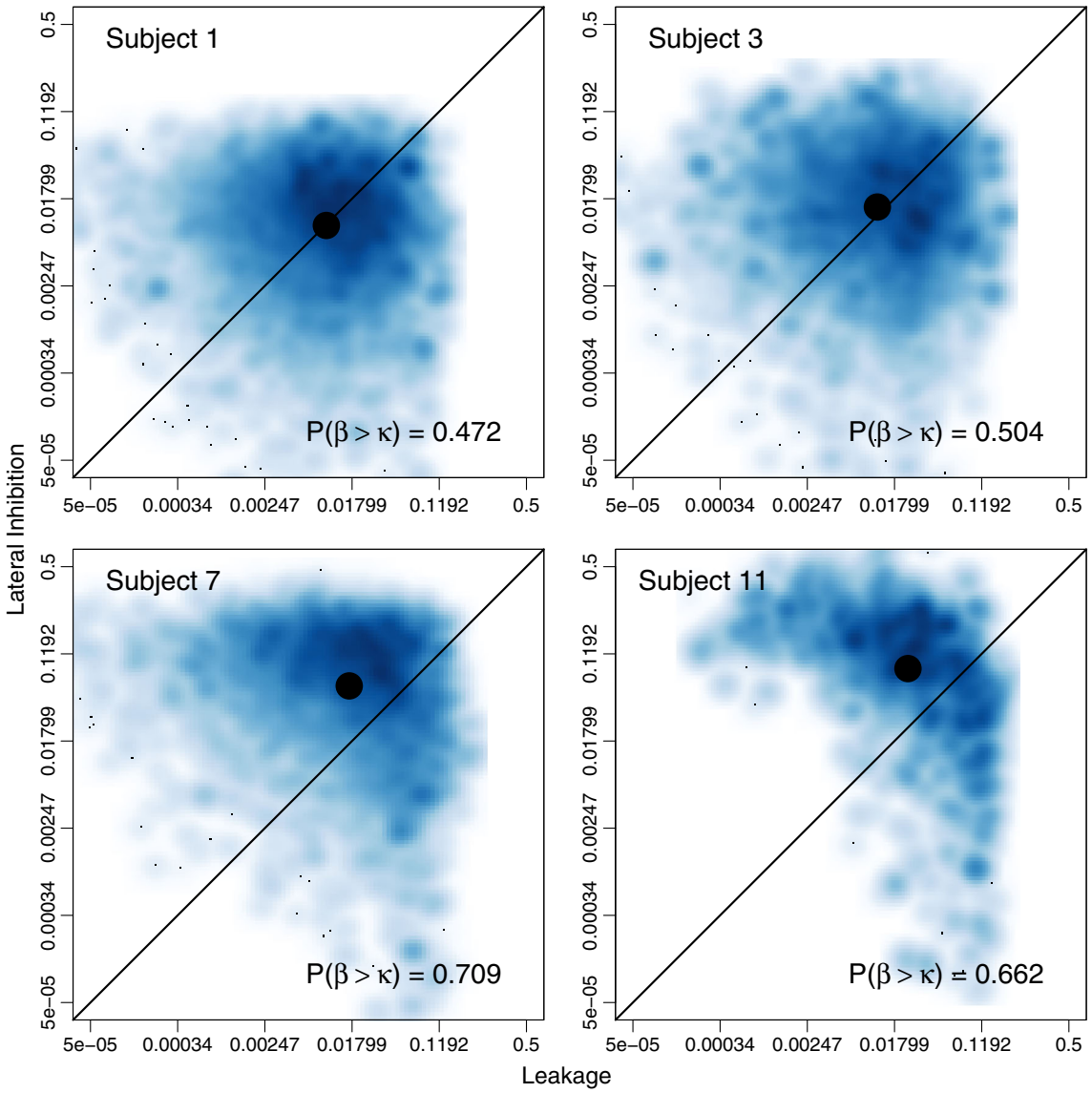

Fig. 10 Estimated joint posterior distributions of the lateral inhibition parameter β, and the leakage parameter κ for the 4 subjects in our data. Note that each joint posterior appears in the logit space

squares estimation, as well as other likelihood-free algorithms, we have shown that the PDA method was able to provide accurate estimates of the model parameters at both the subject and group levels.

We then fit hierarchical versions of both the LBA and LCA models to the same set of experimental data. We explained that because the data consisted of multiple speed emphasis conditions, it posed the greatest modeling challenge for the PDA method, due to the nature of the SPDF construction. For the LBA model, we showed that our method was able to accurately recover the true posterior distributions for Subject 1 and the hyper mean parameters. We then compared the estimates obtained from fitting the two models and found similar patterns among the threshold and drift rate parameters. We also examined the degree of lateral inhibition present in each subject's response strategy (see Fig. 10). The posterior estimates revealed that two of the subjects displayed no significant leakage or lateral inhibition, whereas the other two subjects displayed a slight inhibition dominant pattern.

While the PDA method is a substantial improvement over other likelihood-free methods, the current instantiation still produces only an approximation to the true posterior distribution. We will now discuss the limitations of our method.

Limitations

Because the PDA method does not require the use of summary statistics, the approach is "error free" in the sense that there is no source of error as a result of using summary statistics. We stress however, that our algorithm does suffer from three

potential sources of error. The first is Monte Carlo error, which is intrinsic to all Monte Carlo approaches (see Robert & Casella, 2004, for a complete discussion). The second source of error arises in the estimation of the PDF by means of the kernel density estimate. This type of error has been studied extensively (see, e.g., Silverman, 1986), and we argue that, for most cases and as more sophisticated density estimation techniques become available, this source of error will become negligible. When developing the method, we tested a variety of different distributions that were representative of the types of distributions one might expect in a typical experiment in psychology: the gamma distribution, the Wald distribution, and the Gaussian distribution. We found that in all cases, constructing the SPDF with the Epanechnikov kernel provided a good approximation to the true PDF. For skewed distributions, we found that applying a transformation (e.g., a shifted logarithmic) before the kernel estimate resulted in even better approximations to the true PDF.

The third source of error, which is also related to the accuracy of the density estimate, stems from the number of simulations of the model per proposal. In the analyses reported here, we found that the number of model simulations had the greatest effect on the accuracy of the estimated posterior distributions. However, the degree of error induced is directly related to the cost of computation. For example, if we simulate a model only a few times, the resulting SPDF will not be reflective of the PDF, and this discrepancy manifests in the estimates of the posterior distribution. On the other hand, if we could simulate the model an infinite number of times, the SPDF would equal the PDF, assuming that our kernel estimate did not produce any error. Thus, we argue that as technology continues to improve, with regard to both computational speed and density estimation methods, the third source of error will also become negligible.

The error due to limited computational capacity can be studied prior to fitting the model to real data or even after a posterior has been estimated. For example, we could simulate the model with an arbitrarily selected parameter set θ_{TEST} a fixed number of times N, and we could call this simulated data the "base estimate." In the case of an analysis performed prior to estimating the posterior, θ_{TEST} might be a set of values on a fine grid in the parameter space. The set θ_{TEST} could also be sampled from an estimated posterior distribution to assess the degree of error in the posterior estimates. We would then generate a new data set under the same parameter value θ_{TEST} and then compare the new data set to the base estimate by means of some discriminant function statistic (e.g., the Kullback–Leibler divergence statistic). After generating and comparing many new data sets, we would obtain a Monte Carlo distribution for the discriminant function statistics, and the range of this distribution would inform our understanding of how well the number of model simulations N represents the PDF. In the limit as $N \rightarrow \infty$, the distribution of fit statistics will shrink to have zero variance. Thus, we would like to select a large enough value of N such that the mean and range of discriminant function statistics are small and the computational burden is manageable. This Monte Carlo testing approach could be used to assess the degree of error in all estimation problems—an assessment that is not possible with current likelihood-free methods.

A final limitation of the PDA method is one of computational complexity. Generating the SPDF for each proposal can be costly, especially for the complex stochastic models used in psychology. However, as technology continues to improve, the cost associated with using the PDA method (and likelihood-free techniques in general) will be dramatically reduced. Whereas the LBA model can be fit to data from a single subject in under 10 min using a multicore computer and our method, we used a graphics processor unit (GPU) for simulations of the LCA model to reduce the computational burden. For example, a typical LCA model fit to a single subject can be obtained in $\approx$ 45 min on an entry-level nVidia GPU running CUDA. We feel that this computational cost is a small price to pay for an accurate, likelihood-free estimate of the posterior distribution.

Although we demonstrated the PDA method with only a few cognitive models, it is important to emphasize that the method is applicable to *any* cognitive model in psychology and neuroscience, from models of memory, such as SAM (Gillund & Shiffrin, 1984), REM (Shiffrin & Steyvers, 1997), BCDMEM (Dennis & Humphreys, 2001), Minerva (Hintzman, 1988), TODAM (Murdock, 1982), SLiM (McClelland & Chappell, 1998), TCM (Howard & Kahana, 2002), TCM-A (Sederberg et al., 2008), and CMR (Polyn et al., 2009), to dynamic models of perceptual choice, such as DST (Pleskac & Busemeyer, 2010), dynamic SDT (Turner et al., 2011), RTCON (Ratcliff & Starns, 2009), and the self-regulating accumulator model (Vickers & Lee, 1998, 2000), to neurologically-based models, such as LEABRA (O'Reilly, 2001, 2006), FFI (Shadlen & Newsome, 2001), and the neural integrators model (Mazurek, Roitman, Ditterich, & Shadlen, 2003). In short, if one can simulate a model and it has parameters, using our PDA method it is now possible to obtain accurate Bayesian parameter estimates for the model, given the data. When used in conjunction with DE (ter Braak, 2006; Turner & Sederberg, 2012; Turner, Sederberg, et al., 2013), these estimates are possible at nearly the same computational cost as finding a single best-fitting parameter set.

Conclusions

In this article, we have presented a new method for likelihood-free (posterior) estimation. Our method improves on other likelihood-free methods by creating a general framework that requires no summary statistics or error terms. We believe that

our method is an improvement over rejection-based ABC samplers (Beaumont et al., 2009; Pritchard et al., 1999; Sisson et al., 2007; Toni, Welch, Strelkowa, Ipsen, & Stumpf, 2009) because it allows us to evaluate a proposal's fitness on a continuous scale. Our method improves upon kernel-based ABC approaches (Turner & Sederberg, 2012; Wilkinson, 2008) and synthetic likelihood approaches (Wood, 2010) because it does not require the use of summary statistics. Instead, the framework we propose constructs a nonparametric probability density estimate for the likelihood of each observed data point Y_i. While our approach may seem computationally demanding, we have shown that common assumptions about the distribution of the data (e.g., an i.i.d. assumption) can be exploited to obtain accurate estimates in a short amount of time.

When developing a model in any field, one often investigates many variants of a single base model. Likelihood-free estimation algorithms, which require only simulations of the model, afford us the opportunity to test and explore many variants of the base model without arduous derivations for each model variant. The posteriors that these algorithms provide can give us clear information about the relationships between the parameters—information that might otherwise be available only through extensive experience with each variant. Equipped with only a prior distribution and the PDA method, accurate likelihood-free posterior estimation is now feasible for any model.

Acknowledgements This work was funded by NIH award number F32GM103288. Portions of this work were presented at the 45th Annual Meeting of the Society for Mathematical Psychology. The authors would like to thank Chris Donkin, Cameron R. L. McKenzie, Mike Pratte, and Eric-Jan Wagenmakers for helpful comments that improved an earlier version of the manuscript.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Atkinson, R. C., & Kinchla, R. A. (1965). A learning model for forced-choice detection experiments. *British Journal of Mathematical and Statistical Psychology, 18,* 183–206.

Balakrishnan, J. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors, 40,* 601–623.

Balakrishnan, J. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods, 3,* 68–90.

Balakrishnan, J. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189–1206.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics, 41,* 379–406.

Beaumont, M. A., Cornuet, J. M., Marin, J. M., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika, asp052,* 1–8.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics, 162,* 2025–2035.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences, 362,* 1655–1670.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. Theme issue on modeling natural action selection. *Philosophical Transactions of the Royal Society: B. Biological Sciences, 362,* 1655–1670.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. (2012). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. In A. K. Seth, T. J. Prescott, & J. J. Bryson (Eds.), *Modelling natural action selection* (pp. 91–119). Cambridge, UK: Cambridge University Press.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117–128.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178.

Chapeau-Blondeau, F., & Rousseau, D. (2009). The minimum description length principle for probability density estimation by regular histograms. *Physica A, 388,* 3969–3984.

Chhikara, R. S., & Folks, L. (1989). *The inverse Gaussian distribution: Theory, methodology, and applications*. New York, NY: Marcel Dekker, Inc.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Ranton, FL: CRC Press, Taylor and Francis Group.

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science, 4,* 135–150.

Craigmile, P., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika, 75,* 613–632.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution, 25,* 410–418.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator. *Behavioral Research Methods, 41,* 1095–1110.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial. *Journal of Mathematical Psychology, 55,* 140–151.

Donkin, C., Heathcote, A., & Brown, S. (2009). Is the Linear Ballistic Accumulator model really the simplest model of choice response times: A Bayesian model complexity analysis. In A. Howes, D. Peebles, & R. Cooper (Eds.), *9th international conference on cognitive modeling – ICCM2009*. Manchester, UK.

Dorfman, D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology, 8,* 264–284.

Dorfman, D., Saslow, C., & Simpson, J. (1975). Learning models for a continuum of sensory states reexamined. *Journal of Mathematical Psychology, 12,* 178–211.

Egan, J. P. (1958). *Recognition memory and the operating characteristic (Tech. Rep. No. AFCRC-TN-58-51)*. Bloomington, Indiana: Hearing and Communication Laboratory, Indiana University.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications, 14,* 153–158.

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categoriation decisions under uncertainty. *Psychological Review, 105,* 280–298.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. 1). New York: John Wiley.

Fermanian, J. D., & Salanié, B. (2004). A nonparametric simulated maximum likelihood estimation method. *Econometric Theory, 20,* 701–734.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., & Wagenmakers, E. J. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences, 107,* 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., & Ridderinkhof, K. R. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences, 105,* 17538–17542.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E. J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31,* 17242–17249.

Gao, J., Tortell, R., & McClelland, J. L. (2011). Dynamic integration of reward and stimulus information in perceptual decision-making. *PLoS ONE, 6,* 1–21.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. New York, NY: Chapman and Hall.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley Press.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavioral Research Methods, Instruments, & Computers, 36,* 678–694.

Heathcote, A., & Brown, S. D. (2004). Reply to speckman and rouder: A theoretical basis for QML. *Psychonomic Bulletin and Review, 11,* 577.

Heathcote, A., Brown, S. D., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald, and Weibull RT distributions with a parameter dependent lower bound. *Behavioral Research Methods, Instruments, & Computers, 36,* 277–290.

Heathcote, A., Brown, S. D., & Mewhort, D. J. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin and Review, 9,* 394–401.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269–299.

Kac, M. (1962). A note on learning signal detection. *IRE Transactions on Information Theory, IT-8,* 126–128.

Kac, M. (1969). Some mathematical models in science. *Science, 166,* 695–699.

Kontkanen, P., & Myllymäki, P. (2007). MDL histogram density estimation. In Proceedings of the 11th international conference on artificial intelligence and statistics. San Juan, Puerto Rico: Artificial Intelligence and Statistics.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Kubovy, M., & Healy, A. F. (1977). The decision rule in probabilistic categorization: What it is and how it is learned. *Journal of Experimental Psychology: General, 106,* 427–446.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review, 15,* 1–15.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science, 30,* 1081–1095.

Lee, M. D., Fuss, I. G., & Navarro, D. J. (2006). A Bayesian approach to diffusion models of decision-making and response time. In B. Scholkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing* (19th ed., pp. 809–815). Cambridge, MA: MIT Press.

Lee, M. D., & Wagenmakers, E. J. (2012). A course in Bayesian graphical modeling for cognitive science. Available from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf; last downloaded January 1, 2012.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing, 10,* 325–337.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States, 100,* 324–328.

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex, 13,* 1257–1269.

McClelland, J., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin and Review, 15,* 465–494.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology, 53,* 222–230.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 118,* 280–315.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13,* 1199–1242.

O'Reilly, R. C. (2006). Biologically based computational models of cortical cognition. *Science, 314,* 91–94.

Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. *Case Studies in Bayesian Statistics, VI,* 319–334.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two stage dynamic signal detection theory: A dynamic and stochastic theory of confidence, choice, and response time. *Psychological Review, 117*, 864–901.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11. http://CRAN.R-project.org/doc/Rnews/

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129–156.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution, 16,* 1791–1798.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R., & Starns, J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83.

Rice, J. A. (2007). *Mathematical statistics and data analysis*. Belmont, CA: Duxbury Press.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York, NY: Springer.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review, 12,* 573–604.

Rouder, J. N., Yue, Y., Speckman, P. L., Pratte, M. S., & Province, J. M. (2010). Gradual growth vs. shape invariance in perceptual decision making. *Psychological Review, 117,* 1267–1274.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavioral Research Methods, Instruments, & Computers, 33,* 457–469.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115,* 893–912.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology, 86*, 1916–1936.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin and Review, 4,* 145–166.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Sisson, S., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States, 104,* 1760–1765.

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin and Review, 11,* 574–576.

Stone, M. (1960). Models for choice reaction time. *Psychometrika, 25,* 251–260.

ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm Differential Evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing, 16,* 239–249.

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.

Treisman, M., & Williams, T. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

Tsetsos, K., Usher, M., & McClelland, J. L. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in Neuroscience, 5,* 1–18.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review, 120*, 667–678.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with Differential Evolution. *Journal of Mathematical Psychology, 56,* 375–385.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*, 368–384.

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

Turner, B. M., & Van Zandt, T. (2013). *Hierarchical approximate Bayesian computation*. (In press at Psychometrika).

Turner, B. M., Van Zandt, T., & Brown, S. D. (2011). A dynamic, stimulus-driven model of signal detection. *Psychological Review, 118,* 583–613.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108,* 550–592.

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E. J. (2012). Optimal decision making in neural inhibition models. *Psychological Review, 119,* 201–215.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review, 7,* 424–465.

Vickers, D., & Lee, M. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2,* 169–194.

Vickers, D., & Lee, M. (2000). Dynamic models of simple judgements: II. Properties of a self-organizing PAGAN (Parallel, Adaptive, Generalized Accumulator Network) model for mulit-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences, 4,* 1–31.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review, 14,* 779–804.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wilkinson, R. D. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Biometrika, 96,* 983–990.

Wood, S. (2010). Statistical inference for noise nonlinear ecological dynamic systems. *Nature, 466,* 1102–1107.