Motivation

1

What I cannot create, I do not understand.

Richard Feynman

The ultimate endeavor of any good cognitive scientist is to build a model that mimics the essential dynamics of the (human) mind. We should not hope to create a model that is perfectly accurate, as the only model that could perfectly reproduce any dynamic is one that is as complex as the original system; a philosophical argument known as Bonini's paradox. We seek out and develop models because they are our best hope for generalizing complex decision making processes across individuals, tasks, or time. Our criterion for evaluating the utility of a model is not only in what it provides in terms of understanding, but also in how well it can capture essential trends in behavioral data.

A psychological model is a representation of a psychological process, a representation that quantifies or provides a mechanism for how a behavioral task is performed. When we write down a model, we write down a statement of the form

$$y = f(x, \theta).$$

where y is a set of dependent variables or observed measurements, x is a set of independent variables or experimental manipulations, the function f is the mathematical structure dictated by our theory of how the data y are generated, and θ is a set of parameters that relate the independent variables to the model structures. When we want to explore how well the model explains our data y, or how well it predicts new data y', we fit the model to the data y by estimating the parameters θ.

These two endeavors—explanation and prediction—are often considered the foundational pillars of cognitive modeling [1]. Both endeavors are facilitated by accurate cognitive models, and both require detailed knowledge of estimated parameters. Hence, one of the major priorities of any would-be cognitive modeler

J.J. Palestro et al., *Likelihood-Free Methods for Cognitive Science*,
Computational Approaches to Cognition and Perception,
https://doi.org/10.1007/978-3-319-72425-6_1

should be accurate parameter estimation—the method for finding the parameter value or values that best fit the observed data.

Consider a recognition task, in which people are asked to discriminate between items that were presented to them earlier in an experiment (targets) from items that were not (distractors). An old, well-known model of how recognition might be accomplished is the "high threshold" model [2, 3]. This model is based on the idea that if an item were presented to a person earlier in the experiment, it will have left behind a memory trace. What is the function f, or model structure that relates our experimental manipulations x and parameters θ to the dependent variable y, which, in this case, is the probability of identifying an item as "old"?

The high-threshold model has two parameters. The first, R, is the probability that an old target item leaves behind a trace. If a trace is present, then a person will always respond "old." The second, g, is the probability that a person responds "old" by guessing alone. If a trace is not present, then the person will respond old with probability g. Therefore, the probability of responding "old" to a target item is

$$R + g(1 - R).$$

When an item is new (i.e., a distractor), there is no trace, and the process reduces to guessing alone. So the probability of responding "old" to a distractor item is g. The function relating these parameters to recognition performance is therefore

$$f(x, \{R, g\}) = \begin{cases} R + g(1 - R) & \text{if } x \text{ is a target} \\ g & \text{if } x \text{ is a distractor,} \end{cases}$$

where x plays the role of an independent variable that identifies the item as a target or distractor.

The parameters R and g, together with the probability structure $f(x, \{R, g\})$, provide the explanatory mechanisms of the high-threshold model. Knowing how small or large R is tells us something about the efficiency of the memory process. Knowing how small or large g is tells us the tendency of the person to respond "old" or "new." Furthermore, we might expect R and g to be different over different individuals. Some people have better memories than others; some people are less willing to say "old" when they aren't certain. Information about these parameters, how they change over different experimental conditions, and how they differ over individuals, is critical to being able to test the model, decide if it is a good model or not, and in so doing learn about the psychological process that generated the data.

This brings us to the problem that is the focus of this book: how do we learn about the parameters of the model? The process of model fitting can be conducted in a number of ways, all of them "correct" in some sense, but some better than others in the context of different modeling goals. We can divide these methods into two general types: frequentist and Bayesian. The reader is probably already familiar with this distinction, but it is important to highlight the key differences between the two approaches.

Frequentist inference treats data as random and parameters as fixed quantities to be estimated. Parameters are assumed to be fixed within a group, condition, or block of experimental trials, and inference is therefore based on the sample space of hypothetical outcomes that might be observed by replicating the experiment many times. Inference about these unknown, fixed parameters takes the form of a null hypothesis test (such as a *t*-test), or estimating the parameters by determining the parameter values that minimize the difference between the model predictions and the data.

Bayesian inference treats both data and parameters as random, but after data have been obtained they are fixed. Inferences about parameters are based on the probability distributions of the parameters, distributions referred to as *posterior distributions*. It may be odd to think of a parameter as random; think, for example, about a parameter like the constant $\pi = 3.1416\ldots$, or $c = 2.99792458 \times 10^8$. Are these parameters truly random? Most people would say no. The variability in a parameter that is represented by its posterior distribution should be viewed not as true variability in the parameter's value, but as our uncertainty about its true value. But where there is uncertainty there is also information; the Bayesian approach makes use of whatever prior information is available about a parameter and incorporates that information into the inference.

Despite the very different viewpoints that frequentists and Bayesians have about parameters, their goals are the same. Both groups want to develop and test models that can explain and predict behavior. Understanding behavior means understanding how model parameters change with experimental conditions, so long as we can link those parameters to specific mechanisms.

Bayesian methods have become very popular in mathematical and computational psychology over the last decade [4–21]. The reasons for this growth in popularity are numerous, but can be linked to the wide availability of powerful computing resources and, most importantly, to the fact that Bayesian techniques work where frequentist methods cannot easily be applied. In particular, Bayesian inference can be performed in the context of models of theoretical interest, while frequentist methods often must depend on simplifying asymptotic assumptions (e.g., the central limit theorem). These models can be embedded in hierarchical structures that permit estimation of individual differences as well as overall effects of experimental manipulations, and posterior distributions permit us to examine relationships between parameters that would ordinarily be unobservable. Bayesian methods also permit comparisons across models that are very different: non-nested models, where models are more than just special cases of each other (with, say, certain parameters fixed at 0 or 1), can be compared and quantitatively evaluated; and models that differ in dimensionality, the number of unique parameters each have.

In the next sections in this chapter we will present the most common methods for parameter estimation and contrast them with Bayesian methods. We will then discuss the problem that is the central focus of this book: how do we estimate the parameters of a model whose predictions can't be written down mathematically? This will set the stage for the chapters to come.

1.1 Methods of Least Squares

Least-squares methods of parameter estimation (LSE) are so called because the goal is to choose the parameters that minimize the squared distance between the observations $y = \{y_1, y_2, \ldots, y_N\}$ and the predicted values given by the model. That is,

$$\text{SSE} = \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2 ,$$

where N is the sample size and, as before, x represents the independent variables in the experiment, f is the model that relates x to y, and θ are the model parameters.[1]

As an example, simple linear regression assumes a model that predicts $y = mx + b$, so $f(x, \{m, b\}) = mx + b$. Minimizing SSE in simple linear regression yields the least-squares estimates $\hat{m} = r_{xy}\frac{s_y}{s_x}$ and $\hat{b} = \bar{y} - m\bar{x}$, where r_{xy} is the correlation between x and y, $\bar{x}$ and $\bar{y}$ are the sample means, and s_x and s_y are the sample standard deviations.

Simple linear regression is called simple for more than one reason: finding the least squares estimates $\hat{m}$ and $\hat{b}$ can be done by putting pencil to paper. In many situations the function *SSE* cannot be easily minimized, which requires that we use a computer program that searches for the minimum by proposing values for $\hat{m}$ and $\hat{b}$, computing the resulting *SSE*, and then, by proposing new values, attempting to make it smaller. There are many efficient algorithms to do this.

1.2 Maximum Likelihood

Maximum likelihood methods, the frequentist standard for parameter estimation, form the basis for many inferential procedures and model comparison methods. They rely heavily on optimization algorithms because the computations necessary for parameter estimation are usually more complex than those for least squares. In contrast to least squares, the distance between the data and the model's predictions is defined by how closely the probability distribution of the data matches the distributional assumptions of the model. This distance can be minimized by maximizing the likelihood of the data under the model.

Returning to the high-threshold recognition memory model, the data we observe are the numbers of "old" responses O_T to targets and O_D to distractors, out of a total number of items N_T and N_D presented in the experiment. We have already seen that

[1] We will use the notational convention that a variable name without subscripts such as y or x may be either vector or scalar valued; context should make clear which. If a variable is subscripted, such as y_i or x_i, it represents either an element of a vector or a scalar.

$$f(x, \{R, g\}) = \begin{cases} R + g(1 - R) & \text{if } x \text{ is old} \\ g & \text{if } x \text{ is new} \end{cases}$$

gives us the predicted proportions of "old" responses to target and distractor items. Using these predicted proportions, the probability distributions of O_T and O_D are binomial with parameters $\{N_k, p_k = f(k, \{R, g\})\}$ for $k = T, D$. The likelihood of the data $\{O_T, O_D\}$ is the product of the two binomial distributions, which is proportional to

$$\begin{aligned} \ell(\{g, R\} \mid \{O_D, O_T\}) = g^{O_D}(1 - g)^{N_D - O_D} \left[R + g(1 - R)\right]^{O_T} \\ \times \left[1 - (R + g(1 - R))\right]^{N_T - O_T} . \end{aligned} \tag{1.1}$$

The maximum likelihood estimates of g and R are the values $\hat{g}$ and $\hat{R}$ that maximize $\ell(\{g, R\} \mid \{O_D, O_T\})$. This is another case where we can find $\hat{g}$ and $\hat{R}$ without complications and determine that

$$\hat{g} = O_D/N_D \text{ and } \hat{R} = \frac{O_T/N_T - O_D/N_D}{1 - O_D/N_D}$$

maximizes the function $\ell(\{g, R\} \mid \{O_D, O_T\})$. For more complicated likelihood functions, we will need numerical methods.

1.3 Bayesian Methods

Bayesian methods for parameter estimation do not just compute point estimates like least squares and maximum likelihood. As we described above, the final product of a Bayesian analysis is an estimate of a parameter's posterior distribution given the observed data. These methods incorporate the data's likelihood function, the same function used in maximum likelihood estimation, into Bayes' Theorem to arrive at this posterior distribution. Bayesian probabilities used to be called "inverse probabilities," a term that describes the problem of turning a likelihood into a probability distribution over parameters [22].

Bayes' Theorem is probably well known to most readers, but we will restate it here in terms of data, models, and parameters. A model, which we described earlier in terms of its predictions $f(x, \theta)$, states that data y will follow some probability distribution that is "tuned" according to its parameters θ. Using that probability distribution, we can write the likelihood L of y as a function of θ[2]:

[2]Don't confuse the probability (or density) function $f_Y(y \mid \theta)$ with the model structure $f(x, \theta)$. The predictions of the model, described by $f(x, \theta)$ are not necessarily the same as the probability of the data given by $f_Y(y \mid \theta)$, though they were the same for the high-threshold model above. For the simple regression model, however, $f(x, \{m, b\}) = mx + b$, while most applications of

$$L(\theta \mid y) = \prod_{i=1}^{N} f_Y(y_i \mid \theta). \tag{1.2}$$

Although the likelihood L is a function of θ (where y is given), we can think of it (generally) as the probability (or the density) of the sampled measurements y given the parameters θ. Applying Bayes' Theorem, we want to invert the likelihood to obtain a probability (or density) of θ given the data y.

To do this, we will need to specify a prior distribution over θ. This distribution might reflect our past experiences with the model as it was fit to similar data (an informed prior), or we might choose to avoid making strong a priori assumptions about θ and instead choose an objective distribution that is uninformative or relatively flat. Such prior distributions usually spread probability over a wide range of possible parameter values. There are a number of different criteria by which an objective prior might be selected [23], but lack of information is probably the most popular basis for an objective prior.

The choice of a prior gives us a probability or density function $\pi(\theta)$ that represents the variability in the parameter θ before any data are observed. Bayes' Theorem states that

$$\pi(\theta \mid y) = \frac{L(\theta \mid y)\pi(\theta)}{f_Y(y)},$$

where $f_Y(y)$ is the marginal distribution of the data, taken over all possible parameter values. Because $f_Y(y)$ does not depend on θ, it is only a normalizing constant. It is usually very difficult to compute for models of any real complexity, and so we usually write

$$\pi(\theta \mid y) \propto L(\theta \mid y)\pi(\theta);$$

the posterior of θ given y is proportional to the product of the prior and the likelihood.

If we know $\pi(\theta \mid y)$ exactly, then we have everything we need to make inferences about the parameter θ. Not only can we compute point estimates (such as the posterior mean, mode, or median), we can compute exact probabilities for different hypotheses. We can evaluate the probability of a null hypothesis such as $H_0 : \theta \leq 0$, or construct the Bayesian equivalent of a 95% confidence interval: a credible set (θ_0, θ_1) such that $P(\theta \in (\theta_0, \theta_1)) = 0.95$.

Unfortunately, for most realistic models, we don't know $\pi(\theta \mid y)$ exactly, for one or two reasons. First, computing the normalizing constant $f_Y(y)$ is often complicated, preventing us from being able to write down closed-form solutions for $\pi(\theta \mid y)$. This problem, which was one major reason why Bayesian inference has lagged behind the development of frequentist techniques, has led to the development of algorithms that

regression state that y is normally distributed with mean $mx + b$ and some standard deviation σ. In this case, $f_Y(y \mid x, m, b, \sigma)$ is the normal density function that sketches out the bell curve.

permit us to sample values of θ from the posterior. These techniques, such as Gibbs sampling, Metropolis-Hastings sampling, Hamiltonian Monte Carlo sampling, and so forth, do not require explicit calculation of $f_Y(y)$, but instead approximate this marginalizing constant through Monte Carlo techniques.

The second reason we often don't know $\pi(\theta|y)$ exactly is because the likelihood function $L(\theta \mid y)$ may not have an explicit functional form. In psychology, neuroscience, and cognitive science, our goal is to develop a model that mimics human decision making, a process that is extremely complicated even for simple decisions. Often, while developing more complete explanations of behavioral data, models must grow in complexity to be able to account for different decision making patterns. For example, the high-threshold model may explain patterns of decisions from simple recognition memory experiments, but it is not equipped to handle more complex dynamics that appear in other memory experiments, such as those observed in free recall experiments [24,25]. The benefit of more complex models is the power of unifying explanations for many different patterns of behavioral data at once, but the cost is usually one of computational complexity. It is often the case that as models become more complex, it becomes more and more difficult to determine the likelihood of the models' outputs with a set of equations.

And here lies the purpose of this book. There is a growing emergence of successful computational models in psychology, neuroscience, and cognitive science for which the likelihood functions are either unknown or computationally difficult to evaluate. Because the likelihood function has yet to be derived, one must explore the predictions of such models through simulations, and inference procedures are limited to the methods of least squares described above. In other words, due to complications in evaluating the likelihood function, the aforementioned computational models are unable to enjoy the many benefits that Bayesian analyses provide.

1.4 Approximate Bayesian Computation

There are now influential models in the behavioral sciences that are constructed from the "bottom up." Relatively well-understood neural mechanisms are quantified and used as the building blocks of more complex structures that can generate simple responses to quantitative representations of stimuli. Many of these models are used in memory and vision research. These models are tested by repeated simulation of the models' responses using constrained values of the parameters suggested by findings in neuroscience.

Fitting such models to data is orders of magnitude more demanding than the methods we have just outlined for models with explicit likelihoods. The most common method of estimating a simulation model's parameters is called approximate least squares [26, 27]. To understand approximate least squares, refer again to the high-threshold model of recognition memory. If we were to use approximate least squares, the parameter estimates would be obtained by first proposing reasonable values for R and g. These initial values would be used to simulate a number of

responses to a sequence of target and distractor stimuli. For example, a "for" loop that cycles through the target items would first, by sampling from a Bernoulli distribution with probability parameter R_0, determine whether a trace had been laid down for each target item. All items with traces would be given an "old" response. All items without traces would then sample from another Bernoulli distribution with probability parameter g_0, and all items for which the sample was 1 would be given an "old" response. Another "for" loop would cycle through the distractor items, again sampling from a Bernoulli distribution with probability parameter g_0 to determine which distractors are given "old" responses. These two loops result in simulated values for $\hat{O}_{T,0}$ and $\hat{O}_{D,0}$, which can then be compared to the observed values O_T and O_D and evaluated as

$$\widehat{\text{SSE}}_0 = (\hat{O}_{T,0} - O_T)^2 + (\hat{O}_{D,0} - O_D)^2.$$

This would be the very first step in an optimization algorithm that would then select a new set of parameters $\{R_1, g_1\}$, perform a second simulation, the results of which would be used to compute $\widehat{\text{SSE}}_1$, and so on. Although the procedure is not difficult, it can demand enormous amounts of computing power to perform the simulation for each iteration, and, depending on the complexity of the problem, thousands of iterations may be necessary to find the optimal estimates of the model's parameters. Furthermore, because of the variability added by the simulated data, we shouldn't just simulate the data once for each proposed set of parameters, we really need to simulate the data many times, perhaps thousands of times, to reduce the influence of simulation variability on the value of SSE. However, the real reason that this approach is unsatisfactory is simply because it doesn't give us much information in the end: while we may have reasonably accurate point estimates for the parameters, we will not know how they are distributed, how they are correlated with each other, or what kinds of null hypothesis tests might be appropriate for determining if they are changing over experimental conditions.

Approximate Bayesian computation (ABC) was designed to overcome exactly this kind of problem. Originally developed by Pritchard et al. [28], ABC proceeds in a way similar to approximate least squares, replacing the computation of the likelihood with a simulation step. The simulation step produces a sample of simulated data X that is evaluated relative to the observed data Y. This evaluation is made on the basis of the distance between X and Y, and distance can be defined in a number of ways. The SSE is one example of a distance, in which the samples X and Y could be represented by sample statistics like their means and variances. However, ABC does not use distance minimization to generate point estimates of parameters, but rather to estimate the posterior distributions of the parameters.

The logic behind ABC is the following: if a proposed parameter value θ^* is able to generate a simulated data set X that is close to the observed data Y, then it must have associated with it a relatively high posterior probability. Therefore, for some distance function $\rho(X, Y)$, we will keep all values of θ^* that result in $\rho(X, Y) \leq \epsilon_0$ and discard the rest. If we choose $\rho(X, Y)$ and ϵ_0 correctly, then $\pi(\theta \mid \rho(X, Y) \leq \epsilon_0)$ will approximate $\pi(\theta \mid Y)$ [28].

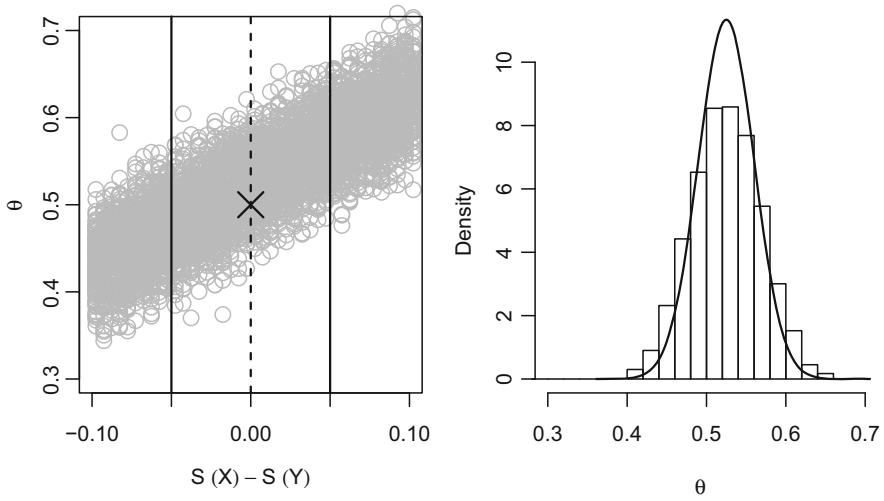

Fig. 1.1 Intuition behind approximate Bayesian computation. The left panel shows the joint distribution of the parameters of interest θ against the distance between the statistics of the observed data $S(Y)$ and the simulated data $S(X)$. The dashed vertical line represents the case where $S(Y) = S(X)$, and the solid black lines represent the degree of tolerance ϵ. The right panel shows the estimated posterior distribution (histogram) under the level of ϵ in the left panel, overlaid by the true posterior (black density)

Consider one last time the high-threshold model of recognition memory. Let

$$\widehat{\text{SSE}} = \left(\hat{O}_T^* - O_T\right)^2 + \left(\hat{O}_D^* - O_D\right)^2$$

be the distance function, where $\hat{O}_T^*$ and $\hat{O}_D^*$ are the simulated data generated by proposed parameter values $\{R^*, g^*\}$. We need to make sure that the number of simulated trials $N_D + N_T$ is the same as the number of trials in the experiment to ensure that the sampling distributions of $\hat{O}_T^*$ and $\hat{O}_D^*$ are comparable to those of O_T and O_D. If $\widehat{\text{SSE}}$ is less than ϵ_0, then we keep $\{R^*, g^*\}$ as a sample from the posterior. If it is greater than ϵ_0 we discard it and sample a new $\{R^*, g^*\}$, possibly from the prior or from some other proposal distribution, and repeat the simulation and computation of $\widehat{\text{SSE}}$. How the proposals are sampled and how ϵ_0 changes (or not) with repeated sampling are determined by the specific ABC algorithm that we choose for this particular problem. We will discuss these algorithms later in Chap. 2.

Figure 1.1 illustrates the logic of the ABC approach more generally. Let $S(X)$ and $S(Y)$ be functions that produce summary statistics (means, variances, quantiles, etc.) of the simulated data X and the observed data Y. For example, the statistics could be the number of "old" responses to target and distractor items for the observed (i.e., O_T and O_D, respectively) and simulated (i.e., $\hat{O}_T^*$ and $\hat{O}_D^*$) data. The distance function $\rho(X, Y)$ is $|S(X) - S(Y)|$. The left panel plots the joint distribution of the

parameter of interest θ against $S(X) - S(Y)$. Staying with our high-threshold model, the parameter θ could correspond to R or g. The "observed" data Y were sampled from a binomial distribution with "success" probability $\theta = 0.5$. This point in the joint distribution of θ and $S(X) - S(Y)$ is represented by an $\times$ at $\theta = 0.5$ and $S(X) - S(Y) = 0$. To generate the joint distribution in the left panel, we randomly selected many different values for θ^* ranging from 0.3 to 0.7. For each new value of θ^*, we simulated data from binomial model and computed the number of successes $S(X)$.

The dashed vertical line in the left panel of Fig. 1.1 is located at 0, when $S(X) = S(Y)$: a perfect match between $S(Y)$ and $S(X)$. If the likelihood were available, the marginal distribution of θ along the vertical line would be the true posterior distribution, which is shown as the black density in the right panel. We can't accept only those values of θ that produce $S(X) = S(Y)$ or $\rho(X, Y) = 0$; such a strict distance criterion would result in an extraordinarily heavy computational load. Instead, we specify the tolerance threshold $\epsilon_0 = 0.05$, which is shown as the solid vertical lines to the left and right of zero in the left panel. This value of ϵ_0 lets us retain enough samples of θ to be able to construct a relatively accurate estimate of θ's posterior. The right panel of Fig. 1.1 shows the histogram of the values of θ in the left panel that produced simulated data X such that $|S(X) - S(Y)| < \epsilon_0$; this is the region of the joint distribution that falls between the two solid vertical lines. The histogram estimate is close to the true posterior that would be obtained had a likelihood been known.

More generally, the relationship between the marginal posterior distribution of a parameter θ and the joint distribution of that parameter and the distance $S(X) - S(Y)$ shown in the left panel of Fig. 1.1 can be expressed as

$$\pi(\theta | Y) \propto \int_{\mathscr{X}} \pi(\theta) f(x, \theta) I(\rho(X, Y) \leq \epsilon) \, dx, \tag{1.3}$$

where $\mathscr{X}$ is the support of the simulated data and $I(a)$ is an indicator function returning one if the condition a is satisfied and zero otherwise. The integration in Eq. (1.3) expresses the marginalization over the random variable $\rho(X, Y)$ that was performed to provide an estimate of θ in the right panel of Fig. 1.1. All values of θ producing data that fell within the black vertical lines were accepted. Note that this marginalization does not take into account the obvious trend in the relation of θ to $S(X) - S(Y)$; this is an important aspect of some versions of ABC algorithms that we will discuss in the next chapter.

1.5 Outline

The focus of this book is to illustrate a variety of ABC techniques on psychological problems. As such, while we will review many different types of ABC algorithms, we will highlight a set of algorithms that have been developed for particular situations that arise regularly when doing cognitive modeling. In the next chapter,

we will outline several different ABC algorithms, focusing in particular on those approaches most similar to the ones we advocate for psychological models. This is not intended to be an exhaustive review of ABC algorithms. Interested readers may consult [29–35] for reviews, additional options, and more mathematical background. In the third chapter, we provide a worked example on the Minerva 2 model [36]. For this model, we provide simulations using two of the algorithms described in the second chapter, and compare their accuracy to a set of analytic expressions describing the limiting behavior of the model [37]. In the fourth and fifth chapters, we discuss a number of applications of ABC algorithms on interesting problems in psychology. In the sixth and final chapter, we provide an outlook on the ability of ABC techniques to advance the field of cognitive science, and discuss the role of mathematical tractability in the development of psychological theory.