Bayesian explorations in mathematical psychology

Matzke, D.

Chapter 1

Introduction

How can we best understand data obtained from psychological experiments? Throughout this dissertation, I will argue that this is best done by means of formal mathematical models. The goal of mathematical modeling is to capture regularities in the data using parameters that represent separate statistical of psychological processes. Mathematical models come in many flavors. Consider, for instance, the following two-choice task: You have to indicate with a button press whether an arrow presented on the computer screen points to the left or to the right. You perform this simple task, say, five times, and produce the following response times (RTs): 440, 409, 326, 482, and 511 ms. We may, for instance, model these five RTs with a normal distribution, $\text{RT}_i \sim \text{Normal}(\mu, \sigma)$, for $i = 1, ..., 5$, and use the sample mean (i.e., 433.6 ms) and the sample standard deviation (71.39 ms) as estimators for μ and σ; it is not a very sophisticated model, but it is model nevertheless.

If additional to the mean and the standard deviation, we also want to describe the —most likely right-skewed— shape of your RT distribution, we have to dig a bit deeper and rely on descriptive RT models such as the ex-Gaussian distribution. RT distributions, however, only provide a descriptive summary of the data and do not account for the psychological processes that underlie performance in a given experimental paradigm. If we want to relate the observed responses to psychological processes, we need to rely on more sophisticated —cognitive— models that go beyond statistical description. Performance in our hypothetical RT experiment, for instance, may be accounted for by the diffusion model (Ratcliff, 1978), one of the most prominent RT models with parameters that correspond to well-defined cognitive processes, such as the rate of information accumulation and response caution.

Let us now make the hypothetical RT experiment slightly more difficult: On some of the trials, the arrow is followed by a tone that tells you to withhold —inhibit– your response on that trial. We are no longer interested in the left-right responses, rather we would like to measure the time required to stop the primary response to the arrow. In this situation, we necessarily have to go beyond observable measurements; after all, stopping latencies cannot be observed directly. We may, for instance, formalize our assumptions about the cognitive processes that underlie successful response inhibition using the independent horse race model (Logan & Cowan, 1984). If we conceptualize response inhibition as a horse race between a go process and a stop process, we can derive and estimate the unobservable latency of stopping. Even better, if we are willing to make parametric assumptions about the distribution of the finishing times of the go and the stop process, we can estimate the entire distribution of stopping times.

Mathematical models can take a variety of forms and this dissertation mirrors this diversity. In addition to descriptive and process models of performance in two-choice RT tasks, I will focus on multinomial processing tree models for the analysis of categorical data as well as well-known statistical models, such as the t test, analysis of variance, (partial) correlations, structural equation models, and mediation analysis.

Once we chose a mathematical model for our data, we need to estimate the model parameters and assess the degree to which the chosen model provides an adequate description of the data. How can we best analyze psychological data sets that are described with mathematical models? Throughout this dissertation, I will argue that this is best done by means of Bayesian inference. In Bayesian inference, uncertainty is expressed in terms of probability. We start with a prior probability distribution that quantifies our existing beliefs about the state of the world. The prior is then updated by the incoming data using Bayes' rule to yield the posterior probability distribution. The posterior quantifies our updated beliefs about the state of the world and takes account of the prior as well as the observed data. In fact, the "...Bayesian approach is a common sense approach..." (Edwards, Lindman, & Savage, 1963, p. 195) that describes how rational decision makers should revise their opinion after considering incoming data. Although contemporary psychology heavily relies on frequentist inference, the application of Bayesian methods in psychological research has increased greatly over the past decade. This increase is partly fueled by the development of Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996), the introduction of Bayesian statistical software, such as WinBUGS, and the availability of the Bayesian equivalent of popular hypothesis tests (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Grasman, & Wagenmakers, 2012).

In this dissertation, I will focus on two applications of Bayesian inference: parameter estimation and model selection. With respect to parameter estimation, I will demonstrate that the Bayesian approach can be extremely valuable for hierarchical models where maximum likelihood estimation becomes practically difficult. With respect to model selection, I will rely on Bayes factors to measure statistical evidence in the context of competing cognitive models as well as standard statistical tests, such as the t test. I will make a case for the use of Bayesian model selection for the following reasons. First, Bayesian inference —as opposed to frequentist inference— does not depend on the intention with which the data were collected and hence does not require adjustments for sequential testing. Second, Bayesian hypothesis testing —as opposed to its frequentist counterpart— enables researchers to measure evidence in favor of the null hypothesis. Finally, Bayes factors —as opposed to p values— quantify the probability of the data under one hypothesis relative to the other; that is, Bayes factors tell researchers what they arguably would like to know in the first place when they engage in hypothesis testing.

The focus on Bayesian inference does not imply that frequentist solutions are always inappropriate nor that we are unable to reach valid conclusions relying on classical statistics. In fact, throughout the dissertation, the reader will stumble upon a number of p values and various other concepts related to classical null hypothesis testing. I will nevertheless argue that the Bayesian approach provides important theoretical and practical advantages that make it particularly suited for tackling problems that research psychologists face on a daily basis.

In the remainder of the introduction, I will give an overview and a brief description of the problems to be addressed in the coming chapters. The dissertation encompasses variety of topics; the glue that holds the diversity of topics together is the commitment to mathematical modeling and principled statistical inference.

1.1 Chapter Outline

Part I. The Analysis of Response Time Distributions

The first part of the dissertation focuses on modeling RTs —observed and unobserved— in two-choice tasks with descriptive RT models, such as the ex-Gaussian and the shifted Wald distributions.

In Chapter 2, I investigate the validity of the cognitive interpretation of the parameters of the ex-Gaussian and shifted Wald distributions. The ex-Gauss and the shifted Wald are popular RT distributions to summarize RT data for speeded two-choice tasks. The parameters of these distributions are often interpreted in terms of specific cognitive processes. We study the validity of this interpretation by relating the parameters of the ex-Gaussian and shifted Wald distributions to those of the Ratcliff diffusion model (Ratcliff, 1978), a successful model whose parameters have a well-established cognitive interpretation (e.g.,Voss, Rothermund, & Voss, 2004). The results clearly demonstrate that the ex-Gaussian and shifted Wald parameters do not correspond uniquely to parameters of the diffusion model.

In Chapter 3, I introduce a Bayesian parametric approach for the estimation of stopping latencies (SSRTs) in the stop-signal paradigm, a popular experimental paradigm to study response inhibition. Based on the horse race model (Logan & Cowan, 1984), several methods have been developed to estimate SSRTs. However, none of these approaches allow for the accurate estimation of the entire distribution of SSRTs. Here we introduce a Bayesian parametric approach that addresses this limitation. The new method assumes that SSRTs are ex-Gaussian distributed and uses MCMC sampling to obtain posterior distributions for the model parameters. We present the results of a number of parameter recovery and robustness studies and apply the approach to published data from a stop-signal experiment.

In Chapter 4, I present BEESTS, an efficient and user-friendly software implementation of the Bayesian parametric approach introduced in Chapter 3. BEESTS comes with an easy-to-use graphical user interface and provides users with summary statistics of the posterior distribution of the parameters as well various diagnostic tools to assess the quality of the parameter estimates. We illustrate the use of BEESTS with published stop-signal data.

Part II. Multinomial Processing Tree Models

The second part of the dissertation focuses on parameter estimation and model selection for multinomial processing tree (MPT) models. MPT models are theoretically motivated stochastic models for categorical data. Due to their simplicity, MPT models have been applied to a variety of areas in cognitive psychology (e.g., Batchelder & Riefer, 1999).

In Chapter 5, I introduce a Bayesian approach that accounts for parameter heterogeneity in MPT models as a result of differences between participants as well as items. Traditionally, statistical analysis for MPT models is carried out on aggregated data, assuming homogeneity in participants and items (Hu & Batchelder, 1994). In many applications, however, it is reasonable to treat both participant and items effects as random and base statistical inference on unaggregated data. Here we focus on a crossed-random effects extension of the pair-clustering model (Batchelder & Riefer, 1980), one of the most extensively studied MPT models for the analysis of free recall data. We apply the crossed-random effects pair-clustering model to novel experimental data featuring the manipulation of word frequency.

In Chapter 6, I present various procedures for model comparison in the context of MPT models. The topic of quantitative model comparison has received —and continues to receive— considerable attention (Pitt & Myung, 2002). Here we focus on two popular information criteria, the AIC ("an information criterion", Akaike, 1973) and the BIC ("Bayesian information criterion", G. Schwarz, 1978), on the Fisher information approximation of the minimum description length principle (MDL; Grünwald, 2007), and on Bayes factors as obtained from importance sampling (Hammersley & Handscomb, 1964).

Part III. Correlations, Partial Correlations, and Mediation

The third part of the dissertation focuses on estimating and testing observed and unobserved (partial) correlations.

In Chapter 7, I examine the power to reject the hypothesis of perfect correlation in the context of higher-order structural equation models. In higher-order factor models, general intelligence (g) is often found to correlate perfectly with lower-order common factors, suggesting that g and some well-defined cognitive ability, such as working memory, may be identical. Here we investigate the power to reject the equivalence of g and lower-order factors using artificial data sets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicate that most case studies that reported a perfect correlation between g and a lower-order factor were severely underpowered to detect the distinctiveness of the two factors.

In Chapter 8, I discuss a Bayesian method for correcting the correlation coefficient for the uncertainty of the observations. The Pearson product-moment correlation coefficient can be severely underestimated when the observations are subject to measurement noise. Various approaches exist to correct the estimation of the correlation in the presence of measurement error, but none are routinely applied in psychological research. Here we outline a Bayesian correction method for the attenuation of correlations proposed by Behseta, Berdyyeva, Olson, and Kass (2009). We illustrate the Bayesian correction with two empirical data sets and demonstrate that its application can substantially increase the correlation between noisy observations.

In Chapter 9, I discuss a default Bayesian hypothesis test for mediation. In order to quantify the relationship between multiple variables, researchers often carry out a mediation analysis. In such an analysis, a mediator (e.g., knowledge of healthy diet) transmits the effect from an independent variable (e.g., classroom instruction on healthy diet) to a dependent variable (e.g., consumption of fruits and vegetables). Almost all mediation analyses in psychology use frequentist parameter estimation and hypothesis testing techniques. Here we describe a default Bayesian hypothesis test based on the Jeffreys-Zellner-Siow approach (Rouder et al., 2009).

Part IV. Improving Research Practice

The fourth and final part of the dissertation focuses on suboptimal research practices in psychology.

In Chapter 10, I introduce a novel variant of proponent-skeptic collaboration that focuses on the association between horizontal eye movements and episodic memory. A growing body of research suggests that horizontal saccadic eye movements facilitate the retrieval of episodic memories in free recall and recognition memory tasks. Nevertheless, a minority of studies have failed to replicate this effect. Here we attempt to resolve the inconsistent results by introducing a novel variant of proponent-skeptic joint research. The proposed approach combines the features of adversarial collaboration (Kahneman, 2003) and purely confirmatory preregistered research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). As anticipated by the skeptics, the results of a series of Bayesian hypothesis tests indicate that horizontal eye movements did not improve free recall performance in the joint experiment.

In Chapter 11, I describe a comparison of the statistical evidence provided by p values, effect sizes, and default Bayes factors. Statistical inference in psychology heavily relies on p value significance testing. This traditional approach, however, has been widely criticized. Here we present a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. The comparison yields two main results. First, although p values and default Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support. That is, 70% of the p values that fall between .01 and .05 correspond to Bayes factors that indicate that the data are no more than three times more likely under the alternative hypothesis than under the null hypothesis. Second, effect sizes can provide additional evidence to p values and default Bayes factors.

In the twelfth and final chapter, I focus on multiway analysis of variance (ANOVA). Many empirical researchers do not realize that the common multiway ANOVA harbors a multiple comparison problem. We illustrate the use of the sequential Bonferroni (Hartley, 1955) correction for multiple comparison and show that its application often alters the conclusions drawn from ANOVA designs.