Bayesian explorations in mathematical psychology

Matzke, D.

Chapter 13

Summary and Future Directions

In what follows, I will summarize the results and the main conclusions presented in this dissertation, accompanied by suggestions about avenues for future development.

13.1 The Analysis of Response Time Distributions

Cognitive Interpretation of the Ex-Gaussian and Shifted-Wald Parameters

In Chapter 2, I investigated the cognitive interpretation of parameters of the ex-Gaussian and shifted Wald distributions. A growing number of researchers use descriptive distributions such as the ex-Gaussian and the shifted Wald to summarize response time (RT) data for speeded two-choice tasks. Some of these researchers also assume that the parameters of these distributions uniquely correspond to specific cognitive processes. We studied the validity of this cognitive interpretation by examining the extent to which the ex-Gaussian and shifted Wald parameters could be associated with the kind of psychological processes that are hypothesized by the Ratliff diffusion model (Ratcliff, 1978), a successful model whose parameters have a well-established cognitive interpretation (e.g.,Voss et al., 2004). In a simulation study, we fit the ex-Gaussian and shifted Wald distributions to data generated from the diffusion model by systematically varying its parameters across a wide range of plausible values. In an empirical study, we fit the two descriptive distributions to published data that featured manipulations of task difficulty (i.e., corresponding to changes in drift rate v), response caution (i.e., boundary separation a), and a priori bias (i.e., starting point z; Wagenmakers, Ratcliff, et al., 2008). The results were clear-cut: In the context of a two-choice task, the ex-Gaussian and shifted Wald parameters cannot be associated uniquely with the parameters of the diffusion model. We concluded that researchers should resist temptation to interpret changes in the ex-Gaussian and shifted Wald parameters in terms of cognitive processes. A possible reason for this unfortunate result may be that the descriptive distributions do not take response accuracy into account. Without any knowledge of response accuracy, it is very difficult to distinguish between the effects of task difficulty (or subject ability) and the effects of response caution.

Bayesian Estimation of Stop-Signal Reaction Time Distributions

In Chapter 3, I introduced a Bayesian parametric approach for the estimation of stopping latencies in the stop-signal paradigm. The stop-signal paradigm is frequently used to study response inhibition. In this paradigm, participants perform a two-choice RT task where, on some of the trials, the

primary task is interrupted by a stop signal that prompts participants to withhold their response. The dependent variable of interest is the latency of the unobservable stop response (stop-signal RT or SSRT). Based on the horse race model (Logan & Cowan, 1984), several methods have been developed to estimate SSRTs. Unfortunately, none of these approaches allow for the accurate estimation of the entire distribution of SSRTs. Here we presented a Bayesian parametric approach (BPA) that addresses this limitation. Our method is based on the assumptions of the horse race model and rests on the concept of censored distributions. The BPA treats response inhibition as a censoring mechanism, where the distribution of RTs on the primary task (go RTs) is censored by the distribution of SSRTs. The method assumes that go RTs and SSRTs are ex-Gaussian distributed and uses Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks et al., 1996) to obtain posterior distributions for the model parameters. The BPA can be applied to individual as well as hierarchical data structures. We presented the results of a number of parameter recovery and robustness studies and applied the new approach to published data from a stop-signal experiment. The WinBUGS (Lunn et al., 2012) and WBDev (Wetzels, Lee, & Wagenmakers, 2010) codes that implement the BPA are available online.

Releasing the BEESTS

In Chapter 4, I presented BEESTS, an efficient and user-friendly software implementation of the BPA 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. The software is open source and runs on Windows and OS X operating systems. BEESTS relies on Python for parameter estimation (Patil et al., 2010; Wiecki et al., 2013) and on R (R Core Team, 2012) for the post-processing of the output. For computational speed, the likelihood functions are coded in Cython (Behnel et al., 2011). We illustrated the use of the individual and the hierarchical BEESTS analysis with a published stop-signal data set.

Future Directions

The first part of the dissertation focused on modeling RTs —observed and unobserved— in two-choice tasks with descriptive RT models, such as the ex-Gaussian and the shifted Wald distributions. First, I showed that the parameters of these descriptive distributions should not be interpreted in terms of the cognitive processes assumed by the diffusion model. However, the parameters of the ex-Gaussian and shifted-Wald distributions need not be considered in isolation. Unlike the individual parameters, certain —possibly nonlinear— combinations of the ex-Gaussian or shifted Wald parameters might map uniquely onto parameters of the diffusion model. This possibility awaits further investigation.

Second, I showed that —despite its lack of theoretical underpinning in terms of specific cognitive processes— the ex-Gaussian distribution may be successfully used to describe and model the distribution of stopping latencies in the stop-signal paradigm. However, if the processes underlying response inhibition are of interest, cognitive models, such as the interactive race model (Boucher et al., 2007), the Linear Approach to Threshold with Ergodic Rate (Carpenter, 1981; Carpenter & Williams, 1995; Hanes & Carpenter, 1999), or a modified version of the linear ballistic accumulator model (Brown & Heathcote, 2008) are the appropriate choice. For other alternatives, see Logan et al. (2014). Nevertheless, the ex-Gaussian based BPA is certainly not a redundant tool in the growing arsenal of techniques targeted at estimating stopping latencies. To the contrary: the BPA

may be used to aid model development and evaluate the predictions of competing process models beyond the level of mean SSRT.

Third, I introduced BEESTS, a user-friendly software implementation of the ex-Gaussian based BPA. In order to assess the absolute goodness-of-fit of the BPA, BEESTS relies on posterior predictive model checks. To formalize the model checks, BEESTS computes posterior predictive p values using the median of the observed signal-respond RTs and the median of the signal-respond RTs predicted by the joint posterior of the model parameters. As I repeatedly stressed throughout the dissertation, this approach is not ideal; adequate analysis of RT data should not only focus on the median, but should consider the shape of the entire RT distribution. Accordingly, the posterior predictive model checks in BEESTS should preferably compare the entire distribution of observed signal-respond RTs to the distribution of signal-respond RTs predicted by the model. Unfortunately, this is easier said than done. The assessment of goodness-of-fit using the entire distribution of signal-respond RTs does not only involve the formal comparison of nonparametric distributions; it also involves the comparison of a single observed signal-respond RT distribution to multiple —often thousands of— predicted signal-respond RT distributions. Note also that BEESTS only allows user to assess the *absolute* goodness-of-fit of the model. The assessment of the *relative* goodness-of-fit of the BPA involves the specification of an alternative model and the application of formal Bayesian model selection. The improvement of the posterior predictive checks and the implementation of formal Bayesian model selection methods require further development.

13.2 Multinomial Processing Tree Models

Crossed-Random Effects Multinomial Processing Tree Models

In Chapter 5, I focused on a Bayesian approach that accounts for parameter heterogeneity as a result of differences between participants as well as items in multinomial processing tree (MPT) models. MPT models are theoretically motivated stochastic models for the analysis of categorical data. Traditionally, statistical analysis for MPT models is carried out on aggregated data, assuming homogeneity in participants and items (Hu & Batchelder, 1994). However, in many applications it is reasonable to assume that the model parameters differ both between participants and items. We should then treat both participant and items effects as random and base statistical inference on unaggregated data. Here we introduced a hierarchical 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. Our approach assumed that participant and item effects combine additively on the probit scale and postulated multivariate normal distributions for the random effects. We provided a WinBUGS implementation of the crossed-random effects pair-clustering model and an application to novel experimental data that featured the manipulation of word frequency.

Model Comparison for Multinomial Processing Tree Models

In Chapter 6, I discussed various procedures for model comparison in the context of MPT models. A careful model comparison procedure involves both qualitative and quantitative elements. Important qualitative elements include, for example, plausibility, consistency with known behavioral phenomena, and coherence of the underlying assumptions. The single most important quantitative element of model comparison relates to the tradeoff between parsimony and goodness-of-fit (Pitt & Myung, 2002). The topic of quantitative model comparison has received —and continues to receive— considerable attention in the field of statistics. Here we focused 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). We first provided a general description of the procedures and then applied them to three competing MPT models of memory interference (Wagenaar & Boer, 1987). The R codes (R Core Team, 2012) that implement the MDL and Bayes factor calculations are available online.

Future Directions

The second part of the dissertation focused on parameter estimation and model selection in MPT models. First, I introduced a hierarchical crossed-random effects extension to MPT models that assumes the additivity of participant and item effects. Although I focused exclusively on the pair-clustering model, the crossed-random effects approach may be extended to many other MPT models. The issue of model identification, however, must be carefully considered. The present approach deals only with models that are identified for each participant after collapsing across items and for each item after collapsing across the participants. In paradigms where items are restricted to certain category systems, model identification remains an issue that requires further development.

Second, I reviewed a number of procedures for model comparison in MPT models, with special emphasis on Bayes factors obtained from importance sampling. The chapter exclusively focused on MPT models that assume homogeneity in participants and items. For (crossed-) random effects hierarchical MPT models, however, the computation of Bayes factors using importance sampling is computationally infeasible. The development of more sophisticated model selection methods that are appropriate for hierarchical models is presently an active area of research; preliminary results indicate that reversible jump MCMC (Green, 1995) is a promising tool for the computation of Bayes factors in hierarchical MPT models.

13.3 Correlations, Partial Correlations, and Mediation

Power to Reject the Hypothesis of Perfect Correlation

In Chapter 7, I examined the power to reject the hypothesis of perfect correlation in the context of higher-order structural equation models (SEM). 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. However, the results of studies that addressed the equivalence of g and lower-order factors are inconsistent. We suggested that this inconsistency may partly be attributable to the lack of statistical power to detect the distinctiveness of the two factors. We therefore investigated the power to reject the hypothesis that g and a lower-order factor are perfectly correlated using artificial datasets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. The examination of published studies revealed that most case studies that reported a perfect correlation between g and a lower-order factor were severely underpowered, with power coefficients rarely exceeding 0.30. We concluded by emphasizing the importance of considering power in the context of identifying g with lower-order factors. The R code for the power calculation is available online.

Bayesian Correction for Attenuated Correlations

In Chapter 8, I discussed 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 outlined a Bayesian correction method for the attenuation of correlations proposed by Behseta et al. (2009) that is conceptually straightforward and easy to apply. We illustrated the Bayesian correction with two empirical data sets; in each data set, we first estimated posterior distributions for the uncorrected and corrected correlation coefficient and then computed Bayes factors to quantify the evidence that the data provided for the presence of an association. We demonstrated that correcting for measurement error can substantially increase the correlation between noisy observations. The WinBUGS and R codes that implement the Bayesian correction method and the Bayes factor calculations are available online.

A Default Bayesian Mediation Test

In Chapter 9, I described 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 estimation and hypothesis testing techniques. A recent exception is Yuan and MacKinnon (2009), who outlined a Bayesian parameter estimation procedure for mediation analysis. Here we completed the Bayesian alternative to frequentist mediation analysis by specifying a default Bayesian hypothesis test based on the Jeffreys-Zellner-Siow approach (Rouder et al., 2009). We further extend the default test by allowing the computation of directional or one-sided Bayes factors, using MCMC techniques implemented in JAGS (Plummer, 2009). All Bayesian tests are implemented in the `R` package `BayesMed`.

Future Directions

The third part of the dissertation focused on estimating and testing observed and unobserved (partial) correlations. First, I showed that the majority of studies that use SEM to evaluate the hypothesis of perfect correlation between g and a lower-order factor are underpowered. In contrast to previous chapters, here I relied on classical p value-based hypothesis testing. Second, I moved back to the domain of Bayesian inference, and described a method for correcting observed correlations for the uncertainty of the observations. Moreover, I illustrated a straightforward Bayesian procedure for testing the presence of a correlation using Bayes factors obtained with the Savage-Dickey density ratio method (Dickey & Lientz, 1970). The Bayesian correction method can be viewed as a simple Bayesian structural equation model with two latent variables, each with a single indicator. Third, I stayed within the Bayesian framework but moved away from latent variables, and described a default Bayesian hypothesis test for mediation. The mediation analysis relies on default Bayes factors for correlations and partial correlations (Wetzels & Wagenmakers, 2012).

Possible extensions for the techniques presented above are straightforward. Hypothesis test for assessing (perfect) correlations in SEMs may be implemented in a Bayesian setting. The simple Bayesian attenuation correction may be extended to (higher-order) SEMs featuring multiple latent

factors and indicators. The Bayesian mediation test may be adapted to handle latent variables. These extensions all rely on Bayesian parameter estimation and model selection in SEMs.

Bayesian parameter estimation can be easily implemented in standard statistical software, such as WinBUGS. Also, recent versions of Mplus (i.e, popular software for fitting and testing SEMs; Muthén & Asparouhov, 2012) support Bayesian parameter estimation and posterior predictive assessment of goodness-of-fit. Formal Bayesian model selection methods are also available for SEMs. The computation of Bayes factors, however, relies on sophisticated sampling methods, such as path sampling (S.-Y. Lee, 2007; Song & Lee, 2012) and reversible jump MCMC (Lopes & West, 2004), and is not yet implemented in standard statistical software. Hence it is all but impossible for most research psychologists to take advantage of the Bayesian developments. A notable exception is the work of van de Schoot, Hoijtink, and Deković (2010) that uses Mplus output to compute Bayes factors for inequality-constrained hypotheses. The development and implementation of Bayesian model selection in SEMs is an active and exciting area of research.

13.4 Improving Research Practice

A Preregistered Adversarial Collaboration

In Chapter 10, I introduced a novel variant of proponent-skeptic collaboration that focused 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 attempted to resolve the inconsistent results by introducing a novel variant of proponent-skeptic joint research. The proposed approach combined the features of adversarial collaboration (Kahneman, 2003) and purely confirmatory preregistered research (Wagenmakers et al., 2012). Prior to data collection, the adversaries reached consensus on an optimal research design, formulated their expectations, and agreed to submit the findings to an academic journal regardless of the outcome. To increase transparency and to secure the purely confirmatory nature of the investigation, the two parties set up a publicly available adversarial collaboration agreement that detailed the proposed design and all foreseeable aspects of the data analysis. As anticipated by the skeptics, a series of Bayesian hypothesis tests indicated that horizontal eye movements did not improve free recall performance. The skeptics suggested that the non-replication may partly reflect the use of suboptimal and questionable research practices in earlier eye movement studies. The proponents countered this suggestion and used a p-curve analysis to argue that the effect of horizontal eye movements on explicit memory does not merely reflect selective reporting. The preregistered adversarial collaboration agreement and the data are available on the Open Science Framework.

Bayes Factors, p Values, and Effect Sizes

In Chapter 11, I presented a comparison of the statistical evidence provided by p values, effect sizes, and default Bayes factors. Statistical inference in psychology has traditionally relied heavily on p value significance testing. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with complementary measures of evidence such as effect sizes. The second is to replace inference with Bayesian measures of evidence such as the Bayes factor. Here we provided 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 yielded

two main results. First, although p values and default Bayes factors almost always agreed about what hypothesis is better supported by the data, the measures often disagreed about the strength of this support; 70% of the p values that fell 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. We concluded that the Bayesian approach is comparatively prudent, preventing researchers from overestimating the evidence in favor of an effect.

Sequential Bonferroni Correction for Multiple Comparisons

In the twelfth and final chapter, I focused on the sequential Bonferroni correction in multiway analysis of variance (ANOVA). Many empirical researchers do not realize that the common multiway ANOVA harbors a multiple comparison problem. In the case of two factors, three separate null hypotheses are subject to test (i.e., two main effects and one interaction). Consequently, the probability of a Type I error is 14% rather than 5%. Here we described the multiple comparison problem and demonstrated that researchers seldom correct for it. We then illustrated the use of the sequential Bonferroni (Hartley, 1955) correction —one of several correction procedures— and showed that its application alters at least one of the substantive conclusions in 45 out of 60 articles considered. We argued that preregistration of the hypotheses provides an alternative method to combat the multiplicity problem in multiway ANOVA.

Future Directions

The fourth and final part of the dissertation focused on suboptimal research practices in psychology. First, I focused on questionable research practices and the replication crisis in psychology, and advocated the use of preregistered adversarial collaborations for scientific conflict resolution. I described a proponent-skeptic collaboration on the beneficial effects of horizontal eye movement on memory performance and illustrated how the Bayes factor can be used to quantify evidence *in favor of* the null hypothesis. Second, I showed that although p values and Bayes factors almost always agree about which hypothesis is better supported by the data, p values often overestimate evidence against the null hypothesis. Third, I revisited the frequentist approach, and described a hidden multiplicity problem in multiway ANOVAs and showed that the application of sequential Bonferroni correction often alters conclusions drawn from ANOVA designs.

My main goal was to highlight the advantages of adopting the Bayesian approach in original as well as replication research. Bayesian inference —as opposed to frequentist inference— does not depend on the intention with which the data were collected, it can be used to quantify evidence in favor of the null hypothesis, and enables researchers to assess what they would like to know in the first place when they engage in hypothesis testing, that is, the probability of the data under one hypothesis relative to the other. Various user-friendly default Bayesian procedures are now available for t tests (Rouder et al., 2009; Wetzels et al., 2009), ANOVAs (Masson, 2011; Wetzels et al., 2012), correlations and partial correlations, (Wetzels & Wagenmakers, 2012), mediation (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, submitted) and regression analyses (Liang et al., 2008; Rouder & Morey, 2012). Despite considerable progress over the past decade, the user-friendly Bayesian implementation of many popular techniques, such as structural equation models and contingency tables, awaits further development. Similarly, the further development and the implementation of Bayesian correction methods for multiple comparison (e.g., Berry & Hochberg, 1999; Marchini, Howie, Myers, McVean, & Donnelly, 2007; Scott & Berger, 2006, 2010) are exciting

research areas that will hopefully receive due attention from the statistical community in the near future.