PROPOSAL

Bayesian Evidence Synthesis

Thom Volker, 5868777

Supervisor: Irene Klugkist

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

October 13, 2020

Word count: 767

1 Introduction

In recent years, a meta-analytic way of thinking has been advocated in the scientific community, an approach that is grounded in the belief that a single study is merely contributing to a larger body of evidence (Cumming, 2014). Additionally, the importance of replication has been legitimately supported (e.g., Open Science Collaboration, 2015; Baker, 2016; Brandt et al., 2014). However, most of the attention has been focused on studies that are highly similar, using an identical methodology and research design. These studies, commonly referred to as exact, direct or close replications, are merely concerned with the statistical reliability of the results. Unfortunately, if the results of these studies depend on methodological flaws, inferences from all studies will lead to suboptimal or invalid conclusions (Munafò & Smith, 2018). A safeguard against this deficiency is available in the form of conceptual replications, which primarily assess the validity of a study. That is, conceptual replications are a way of investigating whether the initial conclusions hold under different conditions, using varying measurement instruments or choosing different operationalizations.

Consequently, multiple studies regarding the same hypotheses arise and as per the cumulative nature of science, synthesizing the results is required to build a robust and solid body of evidence. As these conceptual replications employ fundamental between-study differences, established synthesizing methods as (Bayesian) meta-analysis and Bayesian updating are not applicable (Lipsey & Wilson, 2001; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Sutton & Abrams, 2001). These methods are restricted to combining parameter estimates that (i) share a common scale, and (ii) result from analyses with identical functional forms. To overcome these difficulties, Kuiper, Buskens, Raub, & Hoijtink (2013) proposed to use Bayesian Evidence Synthesis (BES), which allows researchers to pool evidence for a specific hypothesis over multiple studies, even if the studies have seemingly incompatible designs.

The use of Bayes Factors (Kass & Raftery, 1995) is at the very heart of BES. First, one proceeds by constructing study-specific hypotheses that reflect a more general hypothesis (i.e., scientific theory). Since the studies might differ conceptually, the hypotheses are allowed to vary over the studies, provided that the hypotheses address the same general hypothesis. Subsequently, the support for each of these study-specific hypotheses can be expressed in terms of a Bayes Factor. Bayes Factors render the support of the hypothesis at hand, relative to an alternative hypothesis, for which conveniently an unconstrained or complement hypothesis can be selected. Loosely speaking, the Bayes Factor expresses how much more likely the hypothesis at hand is, compared to the chosen alternative. Ultimately, the individual Bayes Factors can be multiplied, to express the support for the overall hypothesis in one measure of evidence (Kuiper et al., 2013).

2 Approach

In this study, we will build upon previous work on BES by Kuiper et al. (2013) and Behrens, Ellerbrock, & Kuiper (2019) who presented a proof of concept that has been followed by actual implementation of the method (e.g., Zondervan-Zwijnenburg, Richards, et al., 2020; Zondervan-Zwijnenburg, Veldkamp, et al., 2020). Behrens et al. (2019) showed that BES functions adequately when inequality constrained, that is, informative, hypotheses are evaluated, but shows problems when interest is in evaluation of equality constrained hypotheses (for a thorough overview of the distinction, see Hoijtink, 2012). These problems become especially apparent when a given study lacks statistical power, that is, when the sample size is small relative to the effect size. Additionally, Bayes Factors are known to be highly dependent on the complexity of a given hypothesis (i.e., the number of parameters that are addressed by that hypothesis; Klugkist, Laudy, & Hoijtink, 2005; Mulder, Hoijtink, & Klugkist, 2010). If conceptually similar studies assess hypotheses that differ in the number of addressed parameters, synthesizing the results of the studies may become problematic.

In the current study, we will aim at revealing under which circumstances BES performs adequately, and under which circumstances the method performs unsatisfactorily. We will do so by employing multiple simulations, in which samples are generated and analysed by means of multiple statistical models. Specifically, data will be generated and analysed according to multivariable linear, logit and probit models, driven by actual sociological research problems (e.g., Behrens et al., 2019; Buskens & Raub, 2002; Kuiper et al., 2013). In accordance with Behrens et al. (2019), we will let sample sizes vary between $n = 25$ to $n = 500$ in steps of 25, and small, medium and large effects as defined by Cohen (1988) will be considered. The performance of BES will be evaluated through the true hypothesis rate (THR). Desirably, the THR will converge towards the upper boundary, which is dependent upon the exact specification of the hypotheses. Ethical consent has been provided by the FERB.

3 References

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353–366. https://doi.org/10.1038/533452a

Behrens, L., Ellerbrock, S., & Kuiper, R. M. (2019). *Bayesian evaluation of informative hypotheses for synthesizing evidence from diverse statistical models.*

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Buskens, V., & Raub, W. (2002). *Embedded trust: Control and learning* (pp. 167–202; E. Lawler, S. Thye, & T. H. Woodhouse, Eds.). Amsterdam: Amsterdam: Elsevier.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second). Lawrence Erlbaum Associates.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* CRC Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/0.1037/1082-989X.10.4.477

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research*, *42*(1), 60–81.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* SAGE publications, Inc.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature*, *553*(7689), 399–401. https://doi.org/10.1038/d41586-018-01023-3

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303. https://doi.org/10.1177/096228020101000404

Zondervan-Zwijnenburg, M. A. J., Richards, J., Kevenaar, S., Becht, A., Hoijtink, H., Oldehinkel, A., … Boomsma, D. (2020). Robust longitudinal multi-cohort results: The development of self-control during adolescence. *Developmental Cognitive Neuroscience*, *45*, 100817.

Zondervan-Zwijnenburg, M. A. J., Veldkamp, S. A. M., Neumann, A., Barzeva, S. A., Nelemans, S. A., Beijsterveldt, C. E. M. van, … Boomsma, D. I. (2020). Parental age and offspring childhood mental health: A multi-cohort, population-based investigation. *Child Development*, *91*(3), 964–982. https://doi.org/10.1111/cdev.13267