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: 746

Candidate journals: Statistical Science; Sociological Methods & Research

1 Introduction

In recent years, the importance of replications has received considerable attention (e.g., Open Science Collaboration, 2015; Baker, 2016; Brandt et al., 2014). However, emphasis has been placed primarily on exact, direct or close replication studies. These studies employ an identical methodology and research design as the initial study, and are thus merely concerned with the statistical reliability of the results. If these results depend on methodological flaws, inferences from all studies will lead to suboptimal or invalid conclusions (Munafò & Smith, 2018). To overcome these limitations, the use of conceptual replications has been advocated (e.g., Munafò & Smith, 2018; Lawlor, Tilling, & Davey Smith, 2017). Specifically, conceptual replications scrutinize the extent to which the initial conclusions hold under different conditions, using varying instruments or operationalizations.

However, established methods such as (Bayesian) meta-analysis and Bayesian updating are not applicable when studies differ conceptually. This is due to the fact that these methods require that the parameter estimates (i) share a common scale, and (ii) result from analyses with identical function forms (Lipsey & Wilson, 2001; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Sutton & Abrams, 2001). Consequently, Kuiper, Buskens, Raub, & Hoijtink (2013) proposed Bayesian Evidence Synthesis (BES), which is built upon the foundation of the Bayes Factor (BF; Kass & Raftery, 1995). This method allows researchers to pool evidence for a specific hypothesis over multiple studies, even if the studies have seemingly incompatible designs.

The work by Kuiper et al. (2013) and Behrens, Ellerbrock, & Kuiper (2019) has led to the implementation of the method in applied research (e.g., Zondervan-Zwijnenburg, Richards, et al., 2020; Zondervan-Zwijnenburg, Veldkamp, et al., 2020), and provided the building blocks for the current project. Under general conditions, BES adequately evaluates inequality constrained (i.e., informative) hypotheses, but shows problems when equality constrained hypotheses are evaluated (a thorough overview about the distinction is presented by Hoijtink, 2012). Equality constrained hypotheses become particularly problematic when, at least, one of the studies lacks statistical power. Additionally, BFs are highly dependent on the complexity of a given hypothesis (i.e., the number of parameters that are incorporated into the hypothesis; Klugkist, Laudy, & Hoijtink, 2005; Mulder, Hoijtink, & Klugkist, 2010). As studies differ conceptually, the complexity of the

hypotheses that address the same overarching theory in different studies may also differ. Currently, it is not known to what extent these issue affect the performance of BES, let alone how these might be overcome.

2 The current project

The foremost goal of the current project is to reveal under which circumstances BES performs inadequately. Additionally, we hope to propose adjustments to the method that improve its performance. We will do so by employing simulations in which data will be generated and analysed according to multivariable linear, logit and probit models, to reflect varying study-designs that are often encountered in sociological research (e.g., Kuiper et al., 2013; Buskens & Raub, 2002). Each of the datasets generated by one of these statistical models represents a single "study" in which a set of candidate hypotheses will be evaluated. We consider situations in which all "studies" assess sets of hypotheses with equal complexities, and situations in which the complexities of the hypotheses within the set differ between the "studies". Additionally, we distinguish between situations in which the set of candidate hypotheses contains the true hypothesis (i.e., the data-generating model) within all "studies", and simulations that only consider incorrect hypotheses. The BFs for all candidate hypotheses will be calculated by means of the R-package `bain` (Gu, Hoijtink, Mulder, & van Lissa, 2020), after which the study-specific BFs can be multiplied manually to obtain the combined BF.

We consider sample sizes that increase incrementally from $n = 25$ to $n = 500$ in steps of 25, although we distinguish between situations in which all studies employ identical sample sizes, and situations in which one study remains fixed at $n = 25$. Additionally, all simulations will be conducted for small, medium and large effect sizes (Cohen, 1988). By varying the complexity of the hypotheses and the sample sizes, we hope to pinpoint desirable and undesirable effects of between-study differences on the performance of BES. The actual performance will be deduced from the true hypothesis rate (THR; the proportion of all simulations in which the "true" hypothesis has the largest combined BF). Desirably, the THR will converge towards 1 as the sample size and effect size increase. Situations in which the THR deviates from this desirable effect will determine where additional simulation settings are required, and will guide the focus of potential improvements of the method. 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.

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2020). *Bain: Bayes factors for informative hypotheses.* Retrieved from https://CRAN.R-project.org/package=bain

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.

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2017). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, dyw314. https://doi.org/10.1093/ije/dyw314

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