PROPOSAL

Bayesian Evidence Synthesis

Thom Volker, 5868777

Supervisor: Irene Klugkist

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

September 21, 2020

Word count: 782

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.

As a consequence, 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. When the studies are highly similar, established methods as (Bayesian) meta-analysis and Bayesian updating can be used to pool the results (Glasauer, 2019; Lipsey & Wilson, 2001; Sutton & Abrams, 2001). glausauer als bron moet ik nog even aanpassen, maar vond ik nog geen prioriteit voor de presentatie However, when researchers conceptually replicate an earlier study, fundamental differences between the study-designs may occur. The same holds when researchers unintentionally make different data-analytic choices, a situation that is referred to as the garden of forking paths (Gelman & Loken, 2014). Under these circumstances, conventional synthesizing methods do not suffice, because these are restricted to combine parameter estimates that (i) share a common scale, and (ii) result from analyses with identical functional forms. Consequently, Kuiper, Buskens, Raub, & Hoijtink (2013) proposed to use Bayesian Evidence Synthesis (BES), which allows researchers to pool the 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 about a given relationship between two or more variables in the population. Since the studies might differ conceptually, the hypotheses could also vary over the studies, provided that the all hypotheses address the same general hypothesis (i.e., the same scientific theory). 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 the unconstrained or the 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).

Although BES has been applied in multiple studies (e.g., Zondervan-Zwijnenburg, Richards, et al., 2020; Zondervan-Zwijnenburg, Veldkamp, et al., 2020), research into the performance of the method is still limited. Besides Kuiper et al. (2013), the performance of BES has been investigated by Behrens, Ellerbrock, & Kuiper (2019), who showed that BES tends to perform better when only inequality constrained, that is, informative, hypotheses are considered, as compared to equality constraint hypothesis (for a thorough overview about the distinction, see Hoijtink, 2012). However, both studies predominantly serve as a proof of concept, and more complex simulations should validate and enhance the applicability of the method. Ik wil niemand beledigen natuurlijk, maar de simulatiesettings daar zijn relatief simpel, en wij willen juist verder de diepte in. Als dit te sterk uitgedrukt is hoor ik het graag. Namely, Bayes Factors are highly dependent on the complexity of the hypothesis (i.e., the number of parameters that are addressed by the hypothesis; Klugkist, Laudy, & Hoijtink, 2005), which may pose yet unaddressed problems when the study-specific hypotheses involve differing numbers of parameters. Additionally, since Bayes Factors are highly dependent on the sample size and effect size, the effect of having, at least, one underpowered study in the set of studies should receive considerable scrutiny. To address these questions, multiple statistical models, varying sample sizes and effect sizes as defined by Cohen (1988) will be adopted. The performance of BES will be evaluated by means of the True Hypothesis Rate (THR), which quantifies the proportion of times BES is able to identify the correct hypothesis.

2 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

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

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466.

Glasauer, S. (2019). Sequential bayesian updating as a model for human perception. In S. Ramat & A. G. Shaikh (Eds.), *Mathematical modelling in motor neuroscience: State of the art and translation to the clinic. Gaze orienting mechanisms and disease* (pp. 3–18). https://doi.org/10.1016/bs.pbr.2019.04.025

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.

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

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