Thesis Proposal

The Effect of Individual Study Power on the Performance of Bayesian Evidence Synthesis in the Context of Multiple Regression Analysis

Anita Lyubenova (3384022)

Supervisor: Irene Klugkist

Word Count: 730

FETC protocol number (approved): 22-1864

Methodology and Statistics for the Behavioral Biomedical and Social Sciences

Utrecht University

October 2022

Candidate journal: Behavior Research Methods

Introduction

Much focus has been given to the necessity to accumulate evidence across replications in order to gain certainty about the presence of an effect and its size. To this end, meta-analysis and Bayesian Sequential Updating (BSU; Schönbrodt, Wagenmakers, Zehetleitner, and Perugini, 2016) are among the available tools that can be used. However, both are problematic when aggregating effects over conceptual replications, i.e., studies that investigate the same effect but differ in design, operationalization of the dependent variable, or measurement of the variables (Klugkist & Volker, 2022). An alternative tool that could tackle this problem is Bayesian Evidence Synthesis (BES; Kuiper, Buskens, Raub, and Hoijtink, 2013). Instead of pooling effect sizes or data, BES aggregates evidence at the level of hypothesis. That is, a Bayes factor (BF; Kass and Raftery, 1995) is used to quantify the relative support for one hypothesis over another in each study and BES aggregates the BFs to quantify the overall support across all studies. However, the performance of BES has not been thoroughly investigated in realistic situations when the studies in the set have different evidential value, i.e., when some of them are underpowered, while others are highly powered.

Previous research has shown that BES performs very well if all included studies have adequate power (Klugkist & Volker, 2022; van Wonderen, 2022; Volker, 2022). However, there is evidence that that even a single underpowered study with small sample size can strongly reduce the support for the true hypothesis (Volker, 2022). Another study, however, showed that this effect diminished when the set of studies was larger and included more studies with adequate sample size (van Wonderen, 2022). Another finding was that if the set of studies included only studies with small samples, the support *against* the true hypothesis accumulated (Volker, 2022).

In a set of conceptual replications power can differ across studies not only because of the sample size but also because of other factors, such as the scale of the outcome and the analysis type (e.g., logistic, probit, linear regression,), or the complexity of the hypothesis, that is, the number of constraints imposed on the parameters (Klugkist & Volker, 2022; Volker, 2022). Thus, it is important first to adequately quantify the power of each study while taking all these factors into account. Only then can the effect of individual study power on the performance of BES be systematically evaluated.

Using such power quantification this paper aims to (1) elucidate how the distribution of power in the study set relates to the performance of BES, and in particular, to what extent can highly powered studies compensate for underpowered ones; (2) provide applied researchers with R code to calculate power for individual studies; and (3) provide power tables that can give impression of how reliable the aggregate support from BES would be, given certain study and population characteristics.

Analytical Strategy

To address these aims simulation studies will be performed, in which a study set is simulated such that data for individual studies is generated and analyzed according to multivariate linear, logistic or probit regression. The following aspects will be manipulated across simulation studies: the number of the studies in set (e.g., 10, 20, 40); the distribution of individual study power in the study set (e.g., spread and skewness); the number of the predictors in the regression model (2 and 3), as well as the correlation between the predictors (.1, .3 and .5).

In this paper the power of a study to test one hypothesis H_i against an alternative hypothesis H_a is defined and quantified as the probability of obtaining a BF that supports H_i (i.e., BF>1) if H_i is true or the probability of obtaining a BF that supports H_a (i.e., BF<1) if the H_a is true (whichever is smaller). This definition has been adapted from the concept of reliability of a BF (Hoijtink, 2011) that has also been employed in Fu's (2022) algorithm for sample size determination for Bayesian hypothesis testing.

The performance of BES will be evaluated as the proportion of times the BES-aggregate shows sufficient support for the true hypothesis across iterations. Different definitions of sufficient support will be investigated (e.g. BF larger than 3, 10 and 20).

All analyses will be performed in R (version 4.2.0; R Core Team, 2022). Approximate adjusted fractional Bayes factor (Gu, Mulder, & Hoijtink, 2018) will be computed with the R package BFpack (version 1.0.0; Mulder et al., 2019).

Cited References

Fu, Q. (2022). *Sample Size Determination for Bayesian Informative Hypothesis Testing.* https://doi.org/10.33540/1221

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists / Herbert Hoijtink*. *Chapman & Hall/CRC statistics in the social and behavioral sciences*. Boca Raton: CRC Press.

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

Klugkist, I., & Volker, T. B. (2022). *Bayesian Evidence Synthesis for Informative Hypotheses: An introduction*, Manuscript submitted for publication.

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining Statistical Evidence From Several Studies. *Sociological Methods & Research*, *42*(1), 60–81. https://doi.org/10.1177/0049124112464867

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., . . . van Lissa, C. (2019). *BFpack: Flexible Bayes Factor Testing of Scientific Theories in R.*

R Core Team (2022). R: A language and environment for statistical computing (Version 4.2.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2016). *Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences*. Center for Open Science. https://doi.org/10.31219/osf.io/w3s3s

Van Wonderen, E. (2022). *Comparing Bayesian evidence synthesis to meta-analysis: A simulation study and empirical application*, Manuscript submitted for publication.

Volker, T. B. (2022). *Combining support for hypotheses over heterogeneous studies with Bayesian Evidence Synthesis: A simulation study*, Manuscript submitted for publication.

Reading list

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013a). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*, *14*(8), 585–586. https://doi.org/10.1038/nrn3475-c4

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013b). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. [J.] (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Cohen, J. [Jacob] (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Etz, A., & Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, *11*(2), e0149794. https://doi.org/10.1371/journal.pone.0149794

Gu, X. (2021). Evaluating predictors' relative importance using Bayes factors in regression models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000431

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511–527. https://doi.org/10.1037/met0000017

Hoijtink, H. (2021). Prior sensitivity of null hypothesis Bayesian testing. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000292

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria*, *25*(3), 301–321.

Klaassen, F., Hoijtink, H., & Gu, X. *The power of informative hypotheses.*

Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H., & Hoijtink, H. (2018). All for one or some for all? Evaluating informative hypotheses using multiple N = 1 studies. *Behavior Research Methods*, *50*(6), 2276–2291. https://doi.org/10.3758/s13428-017-0992-5

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Lucas, J. W. (2003). Theory-Testing, Generalization, and the Problem of External Validity. *Sociological Theory*, *21*(3), 236–253.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171. https://doi.org/10.1111/bmsp.12013

Van de Schoot, R., Hoijtink, H., Hallquist, M. N., & Boelen, P. A. (2012). Bayesian Evaluation of inequality-constrained Hypotheses in SEM Models using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(4). https://doi.org/10.1080/10705511.2012.713267

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

Zondervan-Zwijnenburg, M. A. J., Veldkamp, S. A. M., Neumann, A., Barzeva, S. A., Nelemans, S. A., van Beijsterveldt, C. E. M., . . . 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