bes$_{i}ntro_{p}aper.bib$

Bayesian Evidence Synthesis for Informative Hypotheses: An introduction

Irene Klugkist, Thom Volker

Methodology and Statistics, Social and Behavioral Sciences, Utrecht University

January 28, 2022

Abstract

To establish a theory one needs cleverly designed and well executed studies with appropriate and correctly interpreted statistical analyses. Equally important, one also needs replications of such studies and a way to combine the results of several replications into an accumulated state of knowledge. An approach that provides an appropriate and powerful analysis for studies targeting pre-specified theories is the use of Bayesian model selection for informative hypotheses. Furthermore, it is claimed that an additional advantage of the use of this Bayesian approach is that combining the results from multiple studies is straightforward. In this paper we will discuss the behavior of Bayes factors in the context of evaluating informative hypotheses with multiple studies. By using simple models and (partly) analytical solutions we will compare two different approaches to combine evidence of multiple studies and by doing so clarify how different replication or updating questions can be evaluated.

Keywords and phrases: Bayesian evidence synthesis, Bayes factors, Bayesian updating, Informative hypotheses, Replication

1 Introduction

For any empirical science, replicability is an essential topic. There are several papers on the need for replication, including explanations on reasons for lack of replication studies and recommendations on how to increase replicability asendorpf$_{r}eplication_{2}013(e.g., Asendorpfetal., 2013; Simonsohn, 20$ $General)$.

It is important to distinguish between different types of replication studies, based on how similar the studies are in their design. *Direct or exact replications* aim for as much similarity with the original study as possible, such that the only difference is that in the replication study new data has been collected. Aggregation of results from exact replications is relatively straightforward. If a study is a strictly exact replication of the initial study and the raw data from both studies are available, then the data can be combined and analyzed as if it was one large study. In practice, usually other approaches are used, for instance, within the Bayesian framework, one could apply Bayesian sequential updating. This provides a summary of results after the initial study, an updated summary after adding one replication, a further updated summary after adding another replication, etcetera. The final result of Bayesian updating of data from multiple studies is exactly the same as the result of one Bayesian analysis of all data combined.

Another common approach to aggregation of multiple studies is (Bayesian) meta-analysis. One advantage of the meta-analysis approach is that one does not need the raw data of the studies. The aggregation is at the level of summary statistics (e.g., effect sizes and standard errors) which are often available in the publications of the separate studies. Another difference making meta analysis more flexible is that studies do not have to be strictly exact replications. With random effects meta-analysis and the option of adding moderators to explain differences between studies, the model accounts for and potentially helps understanding heterogeneity in the results. However, to be able to aggregate results with a meta-analysis still a relatively high level of similarity between studies is required. Since the aggregation is at the level of effect sizes, comparable effect sizes must be available for all studies to be synthesized. For studies that are theoretically related but methodologically highly diverse meta-analyzing the results may not be feasible.

In the context of *indirect or conceptual replications* the studies may indeed be highly diverse. One common theory may be investigated in different contexts, with different study designs, using different instruments, variables, and statistical analyses. An advantage of performing conceptual replications is that results that agree across different methodologies and contexts jointly provide stronger support for the underlying central theory. A disadvantage is that the aggregation of results of conceptual replications is not straightforward.

Kuiper (YEAR) proposed a method based on combining evidence for informative hypotheses on the level of Bayes factors. The underlying idea is that the central theory of interest is allowed to be operationalized differently in each study. The study specific informative hypothesis, that represents the central theory, is evaluated using the Bayes factor. A Bayes factor is a measure that represents the change (based on observed data) from the prior odds of two competing hypotheses to the posterior odds of those hypotheses. Aggregation of evidence from multiple studies is done by using the posterior odds after the first data set as the prior odds for the next (i.e., a replication study). This provides updated (with each new replication) relative support measures for the two hypotheses that are compared.

Using this approach, each study provides a level of evidence for the central theory despite the diversity in study design. Although it has been applied successfully (REFS), a paper describing the correct interpretation of the combined evidence and the advantages and limitations of this approach is currently lacking. The main goal of this article is therefore to provide a clear and correct understanding of this Bayesian Evidence Synthesis (BES) approach, that is based on combining Bayes factors that result form evaluating informative hypotheses.

We will first demonstrate the behavior of Bayes factors for an inequality constrained hypothesis in one study. In this section, the use of Bayesian model selection for informative hypothesis is shortly outlined and the performance of the resulting Bayes factors is demonstrated using a simple binomial example. In the next section, we will consider the synthesis of results from multiple studies. The starting point is an example where a set of exact replications is available, again using the binomial model as illustration. This is followed by an example in the context of conceptual replications, where BES is applied for a set of highly diverse studies. The paper will end with a discussion of results and recommendations for potential users of BES, as well as for future methodological research.

2 Bayes factors for informative hypotheses in one study

Informative hypotheses are hypotheses that impose inequality and equality constraints on model parameters to reflect specific expectations that researchers may have when designing their study. Some background on the motivation to use informative hypotheses is provided in the first subsection. A summary of the approach for the evaluation of informative hypotheses using Bayes factors is provided next. This approach has been proven useful and intuitive, and has been described, investigated and applied for the analysis of single studies extensively in the past two decades (e.g. REFS). In the final subsection, a binomial example will illustrate the approach. With this simple model, analytical solutions are available for the Bayes factors of interest. The conclusions from the binomial example, however, also extend to other examples and statistical models.

2.1 Informative hypotheses

Researchers often initiate their study with specific expectations or theories about the outcomes in mind. For instance, in an experimental design, specific conditions are included because it is *a priori* expected that in certain conditions participants will score higher or lower than in other conditions. Such expectations are naturally represented by order constraints on the model parameters. As an example, consider a study that compares the effectiveness of two treatments and one control condition. The expectation of the researcher is that treatment A will lead to, on average, lower outcome scores (e.g., severity of complaints) than treatment B, but both treatments are expected to be more effective, and thus score lower on average, than the control group C. With μ_j denoting the group mean of group j ($j = A, B, C$), this can be expressed as the informative hypothesis:

$$H_i : \mu_A < \mu_B < \mu_C.$$

Informative hypotheses can also include equality constraints. A researcher could, for instance, state the expectation that treatment A and B are equally successful and that both are better than control condition C, that is: $(\mu_A = \mu_B) < \mu_C$. In addition, specific interaction patterns can also be expressed using inequality and equality constraints. For instance, in a 2×2 design, one could state the expectation that cell means 1 and 2 are larger than 3 and 4, and that the difference between 1 and 3 is larger than the difference between 2 and

4, that is:

$$(\mu_1, \mu_2) > (\mu_3, \mu_4) \text{ and } (\mu_1 - \mu_3) > (\mu_2 - \mu_4).$$

For some examples of applications of informative hypothesis evaluation in psychology, for instance, see: Cooper et al. (2014), Van den Hout et al. (2012), Hartendorp et al. (2012), Bullens et al. (2011).

There are several reasons why a traditional null hypothesis test based on p-values is not the optimal choice for the evaluation of informative hypotheses. First of all, the hypotheses included in the NHT approach are not the research hypothesis of interest. Several authors claimed that the null hypothesis can never be (exactly) true (e.g. Cohen, 1990, 1994; Krueger, 2001; Lykken, 1991) and therefore rejecting it does not tell us anything. In addition, the alternative hypothesis in NHT is not specific or informative (usually just stating 'not H_0'). Royall (1997) argues that the focus of a statistical analysis should not be on the question whether there is evidence against the null hypothesis but, instead, one should ask whether there are scientifically meaningful alternative hypotheses that are better supported (Royall, 1997, p. 81).

An informative hypothesis is an example of a scientific meaningful hypothesis. If one would evaluate it using the NHT approach follow-up tests like pairwise comparisons are required. How to control type 1 and 2 errors in the resulting multiple testing situation is not at all straightforward (e.g., Maxwell, 2004). There is a risk of over-interpreting patterns in the observed data that are not necessarily indicative for patterns in the population, i.e., have a small chance of being replicated in new data. Some researcher may even be tempted to HARKing (Kerr, 1998), that is, Hypothesizing After Results are Known, as if the observed patterns were the anticipated results patterns a priori. Finally, the power to find support for an informative hypothesis using NHT with follow-up testing is extremely low (Klugkist et al., 2015).

A final argument against NHT is that researchers want to know how much support the data provide for their hypothesis. However, the p-value resulting from NHT is not the probability that any hypothesis is true or false and therefore does not provide such information (e.g., Cohen, 1994). A better alternative for testing informative hypotheses has been found within the Bayesian framework and will be presented in the next section.

2.2 Bayesian model selection

Bayesian model selection can be used for the evaluation of informative hypotheses and is based on the Bayes factor. There are many references that explain the Bayes factor in general (e.g. Kass and Raftery, 1995; + meer toegankelijke REFS) as well as in the specific context of testing informative hypotheses (e.g., Béland et al., 2012; Klugkist et al., 2005; meer?). Shortly summarized, the Bayes factor compares two models or hypotheses H_1 and H_2, by:

$$BF_{1,2} = \frac{P(D|H_1)}{P(D|H_2)},$$

where $P(D|H_1)$ and $P(D|H_2)$ denote the marginal likelihood of the observed data under hypothesis 1 and 2, respectively. When the resulting value is larger than one, there is more support for H_1, whereas $BF_{1,2} < 1$ implies more support for H_2.

In order to evaluate an informative hypothesis with a Bayes factor it is therefore required to formulate at least one alternative hypothesis. In the context of informative hypotheses we will investigate three natural choices: the unconstrained alternative, the complement of the informative hypothesis, and the null hypothesis. In the following subsections, each of the options and some of their strengths and limitations are discussed.

2.2.1 Testing against the unconstrained model

From here, let the interest be to evaluate if and to what extent the data support the expectation $H_i : \mu_A < \mu_B < \mu_C$. Testing against the unconstrained alternative $H_u : \mu_A, \mu_B, \mu_C$ is proposed by Klugkist, Laudy, Hoijtink (2005), but see also Hoijtink (2012) and Hoijtink, Klugkist, Boelen (2008). It represents how the Bayes factor computation for informative hypotheses is implemented in what is called the encompassing prior approach. Each informative hypothesis can be seen as the unconstrained hypothesis plus a set of constraints. The encompassing prior approach uses this property by deriving an expression for the Bayes factor, $BF_{i,u}$, that requires only evaluation of the unconstrained model and determining the part of it that is in agreement with the constraints. Such evaluation of the posterior distribution of the parameters provides a measure of relative fit for the constrained versus the unconstrained model (denoted f_i). A similar evaluation of the prior distribution is required to determine the relative size, or complexity, of the constrained model (denoted c_i). The latter shows that a Bayes factor incorporates an automatic correction for model size to prevent from overfitting. Evaluation of prior and posterior distributions can be done by Markov chain Monte Carlo(MCMC) sampling. The resulting estimate of the Bayes factor is:

$$BF_{i,u} = \frac{f_i}{c_i}. \tag{1}$$

Evaluating hypotheses that include equality constraints, with the encompassing prior approach, requires some additional steps (see, Klugkist et al, 2005; Van Wesel, Klugkist, Hoijtink, 2010; Mulder et al., 2010) and will not be further discussed at this point. Finally, note that researchers may have multiple, competing informative hypotheses, say H_1 and H_2. The Bayes factor that mutually compares two informative hypotheses, that is, $BF_{1,2}$, is then easily computed by applying (1) twice providing $BF_{1,u}$ and $BF_{2,u}$ and the notion that

$$\frac{BF_{1,u}}{BF_{2,u}} = \frac{P(D|H_1)/P(D|H_u)}{P(D|H_2)/P(D|H_u)} = \frac{P(D|H_1)}{P(D|H_2)} = BF_{1,2}.$$

2.2.2 Testing against the complementary model

Testing against the complement of a constrained hypothesis is described by Hoijtink (2012), but see also Van Deun et al. (2009) and Van Rossum, van de Schoot, Hoijtink (2013). It provides the most powerful test when the interest lies in *one* informative hypothesis, because the two hypotheses describe mutually exclusive situations. For $H_i : \mu_A < \mu_B < \mu_C$, the complement H_c is the collection of all orderings of means that is not H_i. A potential disadvantage is that it is not straightforward to define the complement of a hypothesis including equality constraints. Pragmatically, Gu et al. (REF) propose that the unconstrained model (H_u) serves as the complement for any hypothesis that includes at least one equality constraint. In this paper, we will limit the examples and simulations to an informative hypothesis that expresses a simple ordering of means in which case the complement is clearly defined as 'any ordering other than the one specified in H_i'.

The computation of the Bayes factor comparing an order constrained hypothesis H_i with its complement H_c follows easily from (1) and the notion that $f_c = 1 - f_i$ and $c_c = 1 - c_i$, providing:

$$BF_{i,c} = \frac{BF_{i,u}}{BF_{c,u}} = \frac{f_i/c_i}{(1 - f_i)/(1 - c_i)}. \tag{2}$$

2.2.3 Testing against the null model

The third option is testing the informative hypothesis against the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C$. Since the null hypothesis is unrealistic ('the exact null is never true') and usually does not represent a theoretical expectation of the researcher, it could be argued that it is not a good competitor for the informative hypothesis. However, often researchers prefer to include and evaluate the option that all sample effects are likely to be chance results and therefore want to compare the theoretical expectation with the model stating that there are no effects at all. The Bayes factor of H_0 against the unconstrained model ($BF_{0,u}$) can also be estimated with (1), but an adjustment is necessary to estimate the fit f_0 and complexity c_0 for the null hypothesis. Technical details and a thorough investigation of the performance of the proposed estimator can be found in, for instance, REFS. Another approach for the estimation of $BF_{0,u}$ is the Savage-Dicky density ratio method as explained by Wagenmakers et al. (2010). This approach is used for the binomial example in this paper.

From the estimated $BF_{0,u}$ one can easily derive the Bayes factor of interest that compares H_i with H_0 using:

$$BF_{i,0} = \frac{BF_{i,u}}{BF_{0,u}}.$$

2.2.4 Prior sensitivity

Note that for the null hypothesis, or any informative hypothesis that includes one or more equality constraints, the results can be highly sensitive to the choice of the (encommpassing) prior. Van Wesel et al. (2010) and Mulder et al. (2010) investigated a method based on training sample data (Berger and Pericchi, 1996) for informative hypotheses, and Mulder et al. (2012) implemented this approach in the software BIEMS (see `http://informative-hypotheses.sites.uu.nl/software/biems` for free download and tutorial); later extended by Gu in the software bain (see XXX).

The focus of this paper is not on the effect of the prior on the resulting Bayes factor (but be warned that this is a relevant issue when using the Bayes factor in practical applications), but on the behavior of Bayes factors in replication. In all analyses of this paper we will explicitly state which prior we used but we will not investigate the sensitivity of results to that choice.

2.3 Bayes factors for a Binomial example

A simple example of an inequality constrained hypothesis is testing a success probability θ based on the number of successes x in a sample of n trials assuming $x \sim Bin(n, \theta)$, that is, a binomial distribution:

$$f(\theta|n,x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{3}$$

It is convenient to use the conjugate beta prior:

$$p(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{4}$$

where $B(\alpha, \beta)$ denotes the beta function. We will use (4) with $\alpha = \beta = 1$ as the prior distribution for a proportion θ without any constraints imposed, that is, $H_u : \theta$. This is equal to the uniform distribution on the interval [0,1], i.e. $p(\theta) = 1$. With this choice one states that, a priori, each value for θ between zero and one is considered equally likely.

The unconstrained posterior distribution using the binomial likelihood and the $Beta(\alpha, \beta)$ prior is $Beta(\alpha+x, \beta+n-x)$. For the $Beta(1, 1)$ prior this reduces to the $Beta(x+1, n-x+1)$ posterior distribution. It is easy to see that this distribution equals the likelihood in (3), that is, the constant prior does not add any information about θ, and therefore the posterior is determined by the data only.

To illustrate inequality constrained testing, we will consider the hypothesis stating that the success probability is larger than 0.6. This hypothesis will be evaluated against the unconstrained, the complement, and the null hypothesis. In this relatively simple model, the Bayes factors can be computed analytically instead of through MCMC sampling from the prior and posterior distributions. Using Figure 1, the calculation of each Bayes factor will be explained. The plot shows the prior distribution, $Beta(1, 1)$, as well as a posterior assuming that we observed a sample of size $n = 10$ with number of successes $x = 7$, providing $Beta(8, 4)$.

For $BF_{i,u}$ and $BF_{c,u}$ we need to evaluate the parts of both the prior and the posterior distributions in agreement with the constraints of H_i and H_c, respectively. To obtain probabilities in a Beta distribution, one can, for instance, use the function `pbeta` in R. For $BF_{0,u}$ we use the Savage-Dickey density ratio method. In Figure 1, the two large dots show the two densities that are required for this ratio: the prior and posterior density at $\theta = 0.6$.

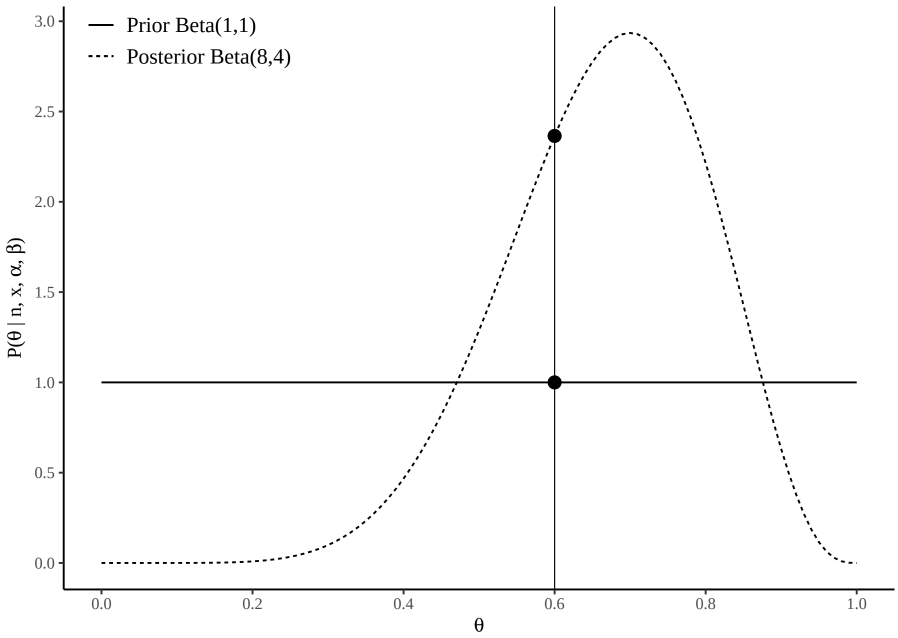

Figure 1: Prior Beta(1,1) and Posterior Beta(8,4)

These densities can, for instance, be obtained by the R function `dbeta`. The resulting values for fit, complexity and the Bayes factor (against the unconstrained model) are provided in Table **??**.

From these results, the Bayes factors of interest (for H_i against each of the alternatives) can easily be computed, as was explained before. In Table **??**, the first column provides the results for the current scenario (n=10, x=7). The other columns, demonstrate the behaviour of the different Bayes factors for increasing sample size (success rate fixed at 0.7

Table 1: Posterior fit f_j (for H_i and H_c) or density at $\theta = .6$ (for H_0), prior fit c_j (for H_i and H_c) or density at $\theta = .6$ and Bayes factor for each hypothesis against H_u (based on $Beta(1,1)$ prior for θ and $x = 7$ successes in $n = 10$ observations)

H_j	posterior	prior	$BF_{j,u}$
$H_u : \theta$	1	1	1
$H_i : \theta > 0.6$	.704	.400	1.76
$H_c : \theta < 0.6$	.296	.600	0.49
$H_0 : \theta = 0.6$	2.365	1.00	2.36

Table 2: Testing $H_i : \theta > .6$ against H_u, H_c, and H_0 for increasing sample sizes n and fixed observed success probability $(x/n = 0.7)$; all with prior $p(\theta) \sim Beta(1, 1)$

	10	20	40	80	100	500	1000
$BF_{i,u}$	1.76	2.00	2.24	2.41	2.45	2.50	2.50
$BF_{i,c}$	3.56	5.99	12,72	41.53	70.69	8.2E5	5.6E10
$BF_{i,0}$	0.74	0.77	0.95	1.73	2.42	6.3E3	2.2E8

and thus in agreement with our hypothesis of interest H_i).

The results show that all three Bayes factors generally behave well with increasing support for H_i when sample size increases. However, we also see that $BF_{i,u}$ is bounded at a maximum of 2.50. The explanation is straightforward when considering the formula $BF_{i,u} = f_i/c_i$. With a result in agreement with H_i and increasing sample size, at some point the entire posterior is in agreement with H_i, providing $f_i = 1$, i.e., the maximum value for perfect fit. The Bayes factor is then determined by the complexity measure c_i that does not depend on sample size and is equal to 0.4 in our example. The maximum BF value is then $1/0.4 = 2.50$.

Another observation is that testing against the null hypothesis suffers from 'power' problems when samples are small. In this example, for samples of 10, 20, or 40 observations, $BF_{i,0} < 1$, showing more support for H_0 than for H_i. The explanation is simple and very similar to what happens with NHT in the frequentist framework: the parsimoniousness of the null hypothesis (it has less parameters than any of the other models) benefits this model when deviations from the null are small compared to the amount of evidence (i.e., the sample size of the study).

These observations lead us to two general recommendations for two different situations. First, if the main goal is to evaluate one informative hypothesis and to what extent it is supported by the data, then it is recommended to use $BF_{i,c}$. It is the most powerful test and it does not have a maximum value, so, the more data in agreement with the hypothesis are observed, the more support the Bayes factor will show. Second, if multiple informative hypotheses are of interest, and specifically their mutual comparison, it is recommended to evaluate them against one another. One of these informative hypotheses could also be the null hypothesis but, when including the null, one needs to take into account that a sufficient sample size is required to have reasonable power to detect the true hypothesis if this is not the null.

3 Aggregating evidence from multiple studies

Replication is important and increasingly performed but it raises the question of how to aggregate results from multiple studies. We will discuss and compare two Bayesian options: Bayesian sequential updating (BSU) and updating at the level of Bayes factors (BES). In the context of exact replications, i.e., for data sets that have the same format, both can be applied but will give different results. In the first subsection this is discussed conceptually and in the next subsection the differences are illustrated using the binomial example again. The final subsection provides an illustration of the synthesis of evidence from a set of *conceptual* replications. The studies to be synthesized come from the same underlying population (with known effect size) but consist of different data formats, that are therefore analyzed by different statistical models. Aggregation of these studies is not feasible with BSU. Using an example with different types of regression models as the statistical analysis tools, we will demonstrate that the BES approach does provide a measure for the combined amount of evidence. A few simulation studies are presented to get a first impression what BES entails, how it works, and what the limitations are.

3.1 Two updating schemes for exact replications

3.1.1 Bayesian sequential updating (BSU)

BSU has been well described in the literature (REFS) and is a procedure that pools data either case by case or study by study. The key idea is that the current state of knowledge about a parameter or hypothesis can be computed at any moment in the data collection period. In the context of replication, data from a first study provides (after specifying an initial prior) the posterior distribution, which is subsequently used as the prior distribution for a second study. After each study, the posterior reflects the current state of knowledge about the model parameters. Also, after each study, Bayes factors can be computed and reflect the current level of relative evidence for the hypotheses of interest.

It is important to note that, with the same initial prior distribution, the posterior after adding one set of 100 observations is exactly the same as the posterior after sequential updating, for instance, after every tenth observation. Also the order in which (subsets) of data come in does not affect the final result. Finally, in contrast with the NHT approach, no penalty for multiple testing is required when computing the Bayes factor after each updating step. (REF)

Sequential updating has some strengths and limitations. A first strength (compared to the NHT approach) is that a priori power analyses are not needed. Instead, one can specify a stopping rule, e.g., *the resulting Bayes factor should be larger than 10 for one of the two compared hypotheses.* Data collection and evaluation of the hypotheses then continues until this threshold is reached (or resources are exhausted). A second strength is that by pooling the data from multiple studies the overall power to detect true effects increases. Especially when the null hypothesis is included as a hypothesis of interest, with small samples there may be limited power to detect relatively small effects. Data pooling increases the overall power to detect non-null effects. The limitation of sequential updating is the high level of similarity that the studies to be aggregated must have. To use data pooling techniques (sequential updating is one example, meta analysis another), the data must have a similar format, because the synthesis takes place at the level of the model parameters or functions of parameters, like effect sizes. For highly diverse studies like those that could result from conceptual replications, we need a more flexible approach.

3.1.2 Bayesian evidence synthesis (BES)

BES aggregates evidence from multiple studies at the level of Bayes factors. The key idea is that a common theory of interest is investigated with multiple studies that are different in their design, e.g., using different variables and/or different statistical models. This makes data pooling complex or even impossible, because the studies provide data sets of different formats and apply models with different parameters. However, the assumption is that for each study an informative hypothesis can be formulated, that will reflect the common theory of interest. These study-specific operationalizations of the common theory are allowed to differ in their parameters and constraints. Irrespective of such differences, for each study, Bayes factors can be computed to reflect the support for the common hypothesis provided by this particular study.

The synthesis of evidence at the level of Bayes factors is done by updating model probabilities, using:

$$\frac{P(H_1|D)}{P(H_2|D)} = BF_{1,2}\frac{P(H_1)}{P(H_2)}. \tag{5}$$

This equation states that the prior odds of two hypotheses can be updated with information from the data through the Bayes factor, providing the posterior odds of the two hypotheses. Subsequently, the posterior odds after observing a first data set can be used as the prior odds for new data. The Bayes factor measuring the evidence for the hypotheses of interest in the second data set then again updates these prior odds to posterior odds. This process can be repeated for each new data set, or each additional replication study.

It is important to note, that the described BES approach is *not* a data pooling approach. It is a flexible tool for evidence synthesis but it measures evidence for a different research question compared to data pooling methods like BSU or meta-analysis. Where the latter investigates to what extent all studies together ('pooled') provide evidence, the former (BES) investigates to what extent the hypotheses of interest are supported in *each study*. These are distinct questions and therefore BSU and BES will provide different results when applied to the same data. This will be illustrated in the next subsection.

The fact that BES is not a data-pooling method can be seen as a limitation of the approach, because one of the goals of aggregating multiple data sets is increasing the power to find support for a hypothesized effect, especially when the individual studies are relatively small. BES does not solve power problems because it does not measure support for the pooled studies. However, in the context of conceptual replications the BES approach fits very well and provides answers to a useful question, that is, to what extent is the support for the common theory robust for different ways in which it can be investigated using diverse studies. If the hypothesized effect finds support in each of these studies, then it can be concluded that the results are robust with respect to (perhaps arbitrary) study design choices. So, BES provides a flexible tool that can synthesize evidence from studies that may be highly diverse and it answers a research question that is relevant for conceptual replications and enables robust scientific conclusions.

3.2 Aggregation in the binomial example

For the binomial example, and under the assumption that we have exact replications, we will examine some results for both BSU and BES. First, we will discuss the behaviour of BES by comparing it to BSU, that is, a data pooling approach. We will, again, evaluate the informative hypothesis $H_i : \theta > 0.6$ by aggregating evidence from 1 to 5 studies, where each study has a sample size of 20 and a success probability of 0.7. In Figure **??** results are plotted for evaluating H_i against the unconstrained model $H_u : \theta$ (panel A), against the complementary model $H_c : \theta < 0.6$ (panel B), and against the null model $H_0 : \theta = 0.6$ (panel C). Note that the resulting BF values are presented as logBF to better fit all results in one plot. A Bayes factor of one (i.e., equal support for both hypotheses) corresponds to a logBF of zero.

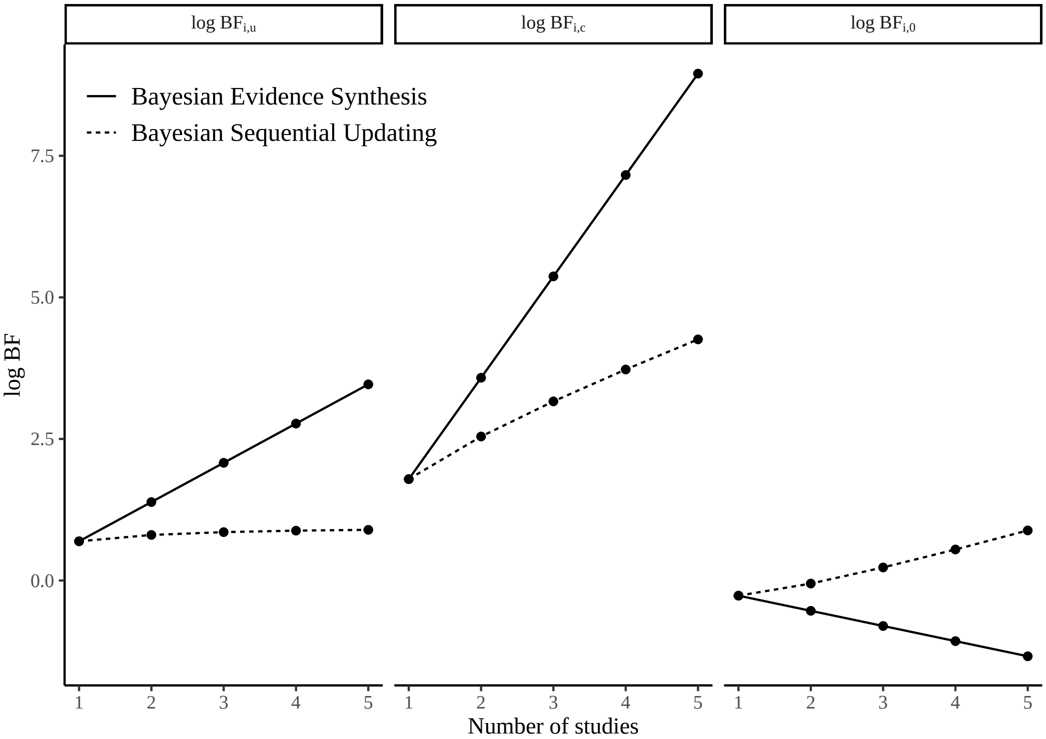

Figure 2: Bayes factors for BES (solid lines) and BSU (dashed lines) after combining 1-5 studies of $n = 20$ and $x/n = 0.7$ per study. All BFs are the result of evaluating $H_i : \theta > 0.6$ against one of the alternatives. In Panel A: $H_u : \theta$. In panel B: $H_c : \theta < 0.6$. In Panel C: $H_0 : \theta = 0.6$.

It is clear that the two approaches provide different results. When testing against the unconstrained model, BSU 'suffers' from the fact that the BF has a maximum. When there is enough evidence to reach a fit-value of one, the BF can not further increase than 1/complexity; in our example 1/0.4=2.5 (logBF=0.92). That explains the almost horizontal dashed line in Panel A. Instead, BES evaluates how much support is found for the hypothesis that H_i is the better hypothesis in each study. Since every single study shows a preference for H_i compared to H_u (BF=2; logBF=0.69), the synthesized evidence over multiple studies increases for each additional study, as can be seen by the steadily increasing solid line in panel A. In Panel B, both synthesis methods show an increase in the aggregated (log)BF with an increasing number of studies that support H_i. Again the differences in results between the approaches are caused by the fact that a different synthesis question is answered. This becomes even more clear in panel C, where H_i is evaluated against H_0, while a single study of n=20 does not have enough power to show preference for H_i over the more parsimonious model H_0. In this scenario, BSU shows an increasing amount of evidence for H_i when adding additional studies. The data pooling approach achieves that at some point there is enough power to find more support for the informative hypothesis. This is not the case for BES, because the question answered by BES is about the support for H_i in each individual study. Because each individual study prefers the simpler H_0, the aggregated evidence for H_0 becomes stronger with each additional study. This explains the decreasing solid line for the declining support for H_i in Panel C.

To get further insight in the behaviour of aggregated evidence using BES, similar plots are constructed for data with success probabilities 0.8, 0.6, and 0.4. Each study still has a sample size of n=20, and 5 identical studies are aggregated. For the interested reader, also the BSU results are plotted. However, we will discuss only the BES results as presented in the top row of Figure **??**.

In the plot on the left, the data per study are in agreement with $H_i : \theta > 0.6$. The three different Bayes factors (logBF) are now plotted together; whereas BES and BSU are provided in separate plots. Contrary to the previous example, the combination of sample size and effect size is now large enough for the BF_{i0} per study to provide support for H_i (BF¿1; logBF¿0). Thus for all three alternative hypotheses, synthesizing over multiple studies gives increasing Bayes factors. As expected, testing against the complementary

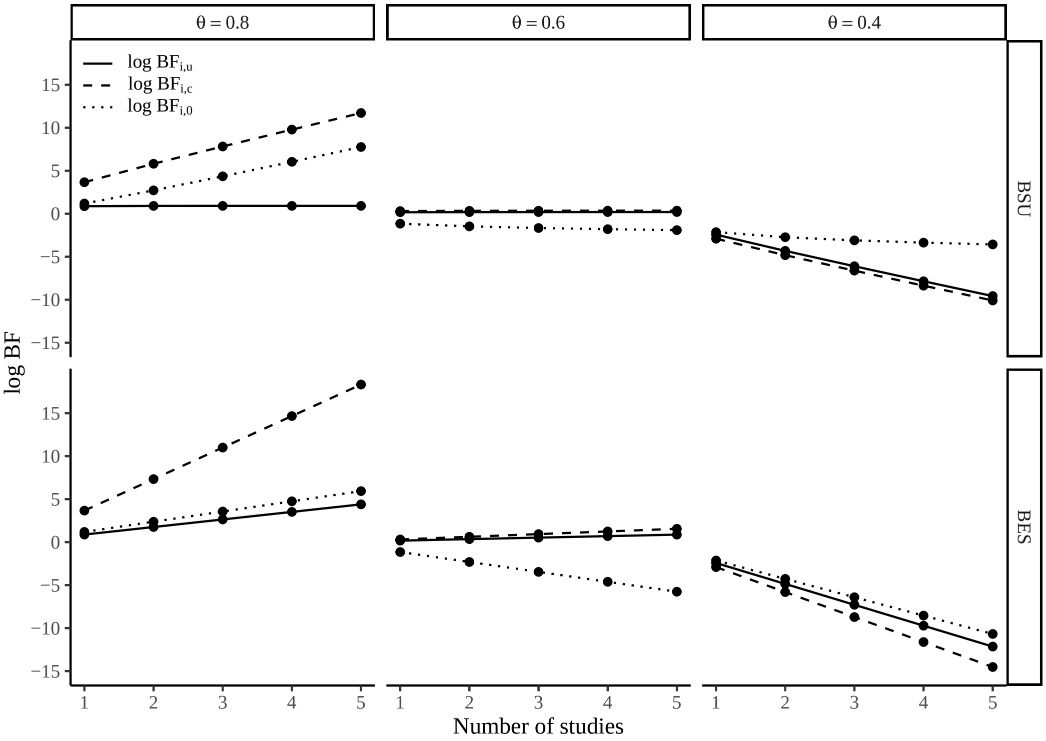

Figure 3: Performance of BES for different effect sizes ($x/n = 0.8$, i.e., in agreement with H_i; $x/n = 0.6$, i.e., in agreement with H_0; $x/n = 0.4$, i.e., in agreement with H_c) in the bottom row. For comparison, the top row shows BSU results for the same data.

hypothesis is most powerful and also leads to the biggest increase in support when more studies are aggregated. In the central plot of Figure **??** the observed effect in each sample is 0.6, that is, it is in agreement with the null hypothesis $H_0 : \theta = 0.6$. The decreasing dotted line shows that aggregation of multiple studies that show this result, indeed, provides increasing support against H_i, in favor of H_0. The last plot, on the right-hand side, presents the synthesized results when each study is not in agreement with H_i.

3.3 An example of conceptual replications

VOORBEELD VAN THOM VOOR VERSCHILLENDE DATA FORMATS; NU DUS ALLEEN BES; KORT LATEN ZIEN DAT HET KAN EN HOE HET ZICH IN BEPERKT AANTAL OMSTANDIGEHDEN GEDRAAGT (TOEWERKEN NAAR WAT WETEN WE AL EN WAT ZIJN OPEN METHODOLOGISCHE VRAGEN)

4 Conclusion and discussion