Utrecht University

Research Report

Bayesian Evaluation of Informative Hypotheses: Aggregating Evidence From Multiple Studies

Author:
Lion BEHRENS

Supervisor:
Dr. Rebecca KUIPER

January 5, 2018

1 Introduction

An increasing number of researchers aims at aggregating evidence for their claims over a variety of studies. Often, these claims are represented by informative hypotheses regarding the order of model parameters such as

$$\begin{aligned} H_1 :&\theta_1 > \theta_2 > \theta_3 > \theta_4, \\ H_2 :&\theta_2 > \{\theta_1, \theta_3\} > \theta_4 > 0, \\ H_3 :&\{\theta_1 - \theta_2\} > \{\theta_3 - \theta_4\}, \end{aligned} \tag{1}$$

In experimental settings, informative hypotheses might represent different expectations on the structure of group means μ_j. Analyzing observational data, behavioral researchers typically construct complex regression models and not seldomly aim to compare the relative importance of different regression coefficients β_j in the same manner.

A large field of scholars elaborated how to evaluate informative hypotheses using a Bayesian approach (Gu et al. [2014], Hoijtink [2012], Klugkist et al. [2005a,b]). Bayes factors (BF) compare the evidence of such hypotheses against an unconstrained alternative and contrast how much support they receive from the data at hand. Posterior model probabilities (PMP) quantify the degree of belief in every evaluated hypothesis on a scale from 0 to 1.

Their developers have shown that these measures perform well given that single studies are under consideration. Recently however, as serious doubts about the replicability of findings have been raised (Open Science Collaboration [2015]), calls for replication research are stronger than ever. Additionally, meta-analytic approaches are gaining prominence (Cumming [2014]). This rising interest in basing statistical inference on a variety of studies calls for the question of how to aggregate evidence for informative hypotheses.

Conceptually, Bayes factors and PMPs provide a well elaborated framework for doing so. However, how such aggregations empirically behave for different scenarios such as small samples, varying effect sizes or low statistical power is unclear. Thus, researchers currently applying this aggregation method cannot be confident in their results. The research presented here fills this gap. Section 2 provides a brief review of how Bayes factors can be constructed for a large class of statistical models that are evaluated via informative hypotheses. Based on this discussion, Section 3 derives an easy way of how evidence obtained from Bayes factors can be aggregated over studies using prior and posterior model probabilities. Section 4 presents simulations on how these aggregations behave for different sample and effect sizes if congruent and mixed evidence is combined across studies. Section 5 concludes with a discussion.

2 Bayes factor construction

In this section, we will shortly recap the general procedure with which Bayes factors are constructed. If $\theta = (\theta_1, \theta_2, ..., \theta_K)^T \in \boldsymbol{R}^{Kx1}$ represent the parameters of any general statistical model under evaluation for $j = 1, ..., K$, then Bayes factors can be constructed for (in)equality constrained hypotheses of the form of

$$H_m : S\theta = 0, R\theta > 0, \tag{2}$$

with hypotheses $m = 1, ..., M$. In Equation 2, S is a matrix representing equality constrains and R a matrix formulating inequality constrains of each hypothesis H_m. The aggregation method that we are assessing in this article allows for hypotheses specifications including equality constrains denoted by " $=$ ", inequality constrains denoted by " $>, <$ " and the absence of constrains denoted by "," regarding model parameters θ and combinations of model parameters such as $\{\theta_1 + \theta_2\}$. For a review of informative hypothesis formulation, see Hoijtink [2012].

Bayes theorem builds the basis for the construction of Bayes factors when evaluating informative hypotheses. It consists of merging the (multivariate) prior distribution $p(\boldsymbol{\theta})$ with the likelihood function $f(y|\boldsymbol{\theta}, \boldsymbol{X})$ to infer a (multivariate) posterior distribution $p(\boldsymbol{\theta}|y)$, where $\boldsymbol{\theta}$ is a vector of all model parameters and y and $\mathbf{X}$ are the data at hand:

$$p(\boldsymbol{\theta}|y) \propto f(y|\boldsymbol{\theta}, \boldsymbol{X})p(\boldsymbol{\theta}). \tag{3}$$

For the (multivariate) linear regression model, the likelihood function $f(y|\boldsymbol{\theta}, \boldsymbol{X})$ can be formulated as

$$f(y|\boldsymbol{\theta}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)} \exp\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{\beta X})^2\}, \tag{4}$$

where $\boldsymbol{\beta}$ is a vector containing the intercept β_0 and all regression coefficients β_j, y is a column vector of the dependent variable and $\boldsymbol{X}$ consists of a matrix of all predictor variables plus a column of ones to estimate the intercept. To construct the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$, a prior distribution needs to be specified for all model parameters. The choice of the latter has a crucial impact on the evaluation of the hypotheses at hand. First, we will look at why this is the case. Afterwards, we will further discuss its specification.

When constructing Bayes factors for informative hypotheses, scholars have proposed to compare each informative hypothesis to an unconstrained alternative H_u which does not impose any constrains on the parameters under evaluation:

$$H_u : \theta_1, \theta_2, \theta_3, \theta_4. \tag{5}$$

Together, all substantive hypotheses plus the unconstrained alternative form the set of hypotheses H_m. The Bayes factor of interest for every hypothesis H_m is denoted by BF_{mu}. Klugkist et al. [2005b] have shown that a simple form to obtain this Bayes factor is

$$BF_{mu} = f_i/c_i. \tag{6}$$

where c_i is the proportion of the prior distribution that is in agreement with H_m (the hypothesis' complexity) and f_i is the proportion of the posterior distribution that is in agreement with H_m (the hypothesis' fit). Naturally, for the unconstrained hypothesis H_u, the Bayes factor BF_{uu} is 1. For all substantive hypotheses, we can see from Equation 6 why the results of this hypothesis testing technique are inherently prior dependent. The Bayes factor BF_{mu} quantifies evidence from the data in favor of hypothesis H_m when comparing it to the unconstrained alternative H_u. Specifically, a value of 3 would indicate that the support for H_m is three times larger than the support for H_u, while balancing out the hypotheses' fit and complexity. To compare the performance of the substantive hypotheses H_1 and H_2 directly against each other, ratios of the hypothesis-specific Bayes factors BF_{mu} are constructed. Thus, H_1 is compared against H_2 via

$$BF_{12} = BF_{1u}/BF_{2u}. \tag{7}$$

Through this comparison, the resulting value directly quantifies the comparative support for each hypothesis H_m and filters out the one that fits the data best. Via the inclusion of H_u, it is prevented that the best one out of a set of bad fitting hypotheses is concluded to perform well.

The general procedure described above assumes that both prior and posterior distributions are normal. Normal prior distributions can be specified for any model in which parameter estimates are not bounded (Gu et al. [2014], p. 515). This applies to structural equation models, mixed models (like multilevel analysis) and any model that can be substituted under the general linear model framework (e.g. linear regression, logistic regression, repeated measures analysis). Since the classical linear regression model also takes on a normal likelihood function as depicted in Equation 4, resulting posterior distributions follow a normal form of

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = N(\hat{\boldsymbol{\theta}}, \hat{\Sigma}_{\theta}), \tag{8}$$

with the standardized parameter estimates $\hat{\boldsymbol{\theta}}$ as means and their covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ as variances. However, for all deviations of the linear model, the likelihood function is not normal, as for example in logistic regression, it will follow a logistic form. Resulting posteriors will thus not have the necessary form. Recently however, Gu et al. [2014, 2017] introduced a Bayes factor computation for any kind of general statistical model by approximating posteriors through normal distributions based on large sample theory (Gelman et al. [2004]). This computation is implemented in the software package Bain that is part of the R environment. Thus, with Bain, Bayes factors for any kind of statistical model can be constructed given that

- a vector of standardized parameter estimates $\hat{\boldsymbol{\theta}}$,
- their covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$

are provided. The following section will elaborate on how Bayes factors obtained from different studies can be combined into a single aggregate of evidence.

3 Aggregation of Evidence

Due to prominent calls for more replication studies (Open Science Collaboration [2015]) and an increasing popularity of meta-analytic approaches (Cumming [2014]), many researchers aim at evaluating the same informative hypotheses using multiple studies. Kuiper et al. [2013] formulated a method to aggregate evidence regarding hypotheses about one parameter of interest. This article translates their logic into the case of informative hypotheses regarding any number of model parameters.

3.1 Theoretical considerations

The method to be presented can be applied if a number of competing hypotheses H_m is formulated towards a number of conceptual parameters θ_j that represent the relation between *theoretical* concepts and a certain dependent variable in an *underlying* population.

Although every individual study aims at inferring towards this underlying population, different researchers are typically following different sampling strategies. Thus, every individual study used a sample from a specific *study population* that was possible to reach in the construction of the empirical sample. In every empirical sample, the conceptual parameters θ_j are thus estimated through specific parameters $\hat{\theta_j}$ that are based on i) different study populations and, often, ii) different measurements. Thus, estimated parameters $\hat{\theta_j}$ cannot be directly compared between studies, as their study populations differ and their effects might be measured on different scales. However, within every study, they are indicative for the same underlying effect and quantify evidence for the same underlying hypothesis.

Thus, the aggregation method presented here does not combine parameter *estimates*. Rather, it combines *evidence* for every informative hypotheses H_m regarding the population parameters θ_j.

3.2 The aggregation method

The last section has shown how Bayes factors for informative hypotheses can be constructed based on any kind of statistical model. Now, we will show how such evidence can be aggregated over multiple studies. At first, the evidence quantified through Bayes factors is transformed into posterior model probabilities. As a general rule, this is done in three steps.

First, prior probabilities π_m^0 have to be formulated for every hypothesis H_m. Since we don't want these prior probabilities to influence the posterior results, these are chosen to be

$$\pi_m^0 = 1/M. \tag{9}$$

Second, the whole probability mass under evaluation is constructed by summing up all Bayes factors BF_{mu} weighted by their (equal) prior probabilities π_m^0 through

$$\sum_m \pi_m^0 BF_{mu}. \tag{10}$$

Third, the posterior model probabilities are constructed by assessing how much percent of the probability mass falls on every hypothesis H_m via

$$\pi_m^1 = \frac{\pi_m^0 BF_{mu}}{\sum_m \pi_m^0 BF_{mu}}. \tag{11}$$

These posterior model probabilities quantify the support for each hypothesis H_m on a scale from 0 to 1 and automatically account for the hypotheses' fit and complexity.

From this construction, Kuiper et al. [2013] derived how evidence for informative hypotheses can be aggregated over multiple studies. Let us say a researcher aims at combining evidence from T. Given that every hypothesis H_m receives an equal prior probability $\pi_{t=1,m}^0 = 1/M.$ when evaluating the first study, posterior model probabilities $\pi_{t=1,m}^1$ can first of all be straightforwardly constructed like outlined above. Following, for every subsequent study under consideration, these posterior model probabilities serve as prior model probabilities when evaluating the next study. For every subsequent study, posterior model probabilities

$$\pi_{t,m}^1 = \frac{\pi_{t,m}^0 BF_{mu}^t}{\sum_{t,m} \pi_{t,m}^0 BF_{mu}^t} \tag{12}$$

can thus be constructed given that $\pi_{t,m}^0 = \pi_{t-1,m}^1$ for $t = 2, ..., T$. Iteratively, one can thus aggregate evidence over all studies that are oft interest. Once the last study has been included in this process, overall posterior model probability measures quantifying evidence for each hypothesis H_m being true in every of the evaluated studies have emerged. As Kuiper et al. [2013] show, the results of this process are independent of the order in which studies enter this aggregation process (p. 73).

4 Simulation Study

Although the aggregation of Bayes factors from separate studies into overall PMPs as introduced in Section 3 is conceptually well elaborated, little is known on how this procedure empirically behaves for different sample and effect sizes. The following simulation study is designed to fill this gap.

4.1 Hypotheses

In this simulation, the performance of the presented aggregation method is assessed using two informative hypotheses H_1 and H_2 that are evaluated against an unconstrained alternative H_u. Overall, three hypotheses are thus forming the set H_m that is under consideration. The hypotheses of interest are formulated as

$$\begin{aligned} H_1 :& \theta_1 > \theta_2 > \theta_3 > \theta_4 > 0, \\ H_2 :& \theta_2 > \{\theta_1, \theta_3\} > \theta_4 > 0, \\ H_u :& \theta_1, \theta_2, \theta_3, \theta_4, \end{aligned} \tag{13}$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ stands for four standardized predictors that are related to a dependent variable y. These hypotheses stand exemplary for a certain number of M inequality constrained hypotheses a researcher might specify to a certain set of parameters in a underlying general population of interest.

4.2 Combining evidence for equal study populations

4.2.1 Population values

In a first simulation run, evidence from four studies drawn from four equal study populations that obtain similar population values is aggregated. Each study population consists of four predictor variables x_1 to x_4 and a dependent variable y. In each of the study populations, the four predictors are normally distributed with mean $\mu_k = 0$ and variance $\sigma_k^2 = 1$:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.25 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 & 0.25 \\ 0.25 & 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 0.25 & 1 \end{pmatrix} \right]. \tag{14}$$

Since in behavioral research, explanatory concepts are often related to each other, each off-diagonal element is set to 0.25. From these predictor variables, a dependent variable y is constructed using

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i, \tag{15}$$

where β_0 to β_4 represent regression coefficients based on ordinary least squares estimation (OLS) and ϵ_i represents a distribution error with mean $\mu_\epsilon = 0$ and variance $\sigma_\epsilon^2 = 1 - R^2$. Following Cohen's work on effect sizes, we specify R^2 as $\frac{f^2}{1+f^2}$ and thus the error term as $\sigma_\epsilon^2 = 1 - R^2 = 1 - \frac{f^2}{1+f^2}$ (Cohen [1992]). We investigate scenarios in which f^2 ranges between small (0.02), medium (0.15) and large (0.35) effects (based on Cohen [1992], p. 157).

To construct y_i, we need to choose specific values for β_0 to β_4. Once the overall explanatory power of the model f^2 is set, values for these regression coefficients can be chosen. However, many sets of regression coefficients comply with a certain choice for f^2. To make sure that we look at meaningful differences between parameters, we choose them based on ratios $\boldsymbol{r}$. Regression coefficients β_k can thus be expressed as $\{\boldsymbol{r}c\}$, where c is a constant value which, multiplied with the vector $\boldsymbol{r}$, determines the concrete value of every coefficient. Ratios vary between 2:1 and 1.33:1. Under Hypothesis H_1, this leads to ratio vectors of $\boldsymbol{r_{2:1}} = (8, 4, 2, 1)$ and $\boldsymbol{r_{1.33:1}} = (2.35, 1.77, 1.33, 1)$. Under Hypothesis H_2, ratio vectors of $\boldsymbol{r_{2:1}} = (2, 4, 2, 1)$ and $\boldsymbol{r_{2:1}} = (1.33, 1.77, 1.33, 1)$ emerge. Since we also want to investigate situations in which neither of the substantial hypotheses is valid and the unconstrained alternative H_u should receive most support, a third hypothesis denoted by H_3 is defined by reversing the ordering of regression coefficients in H_1, leading to a coefficient order of

$$H_3 : 0 < \beta_1 < \beta_2 < \beta_3 < \beta_4 \tag{16}$$

and ratio vectors of $\boldsymbol{r_{1:2}} = (1, 2, 4, 8)$ and $\boldsymbol{r_{1:1.33}} = (1, 1.33, 1.77, 2.35)$. Table 1 summarizes these population parameters that are evaluated in the presented simulation study.

For a specific combination of study population characteristics, the regression values underlying a study population are derived using

$$R^2 = \frac{f^2}{1+f^2} = \sum_{k=1}^{K} \{(r_i\beta_k)^2 c\} + 2 * \sum_{k<k'} \{(r_i\beta_k)c\}\{(r_i\beta_{k'})c\} * 0.25, \tag{17}$$

where it is solved for c using R's `uniroot` function (see Hoijtink [2012], p. 20). Table 2 displays the resulting study population structures for all evaluated scenarios.

Table 1: Simulation parameters when combining equal study populations

Parameter	Variations
Study population characteristics	
Coefficient struture H_m	H_1, H_2, H_3
Overall effect size f^2	0.02, 0.15, 0.35
Regression coefficient ratio $\boldsymbol{r}$	2:1, 1.33:1
Sample characteristics	
Sample size n	50, 100, 500

4.2.2 Simulation procedure

$S = 100$ samples of equal sample sizes varying between $n = 50$, $n = 100$ and $n = 500$ are drawn from each study population. This leads to 3x3x2x3=54 simulation scenarios. Studies are aggregated over each Nth sample. Thus, sample $s = 1$ from study population $t = 1$ is combined with sample $s = 1$ from study population $t = 2$ until study $t = T = 4$, leading to overall PMPs for the first simulation iteration. This is repeated for all S iterations. Thus, 100 overall PMPs are constructed for every simulation scenario.

4.2.3 Performance evaluation

Typically, the performance of a hypothesis testing technique is assessed using the criterion of statistical power, the probability that a test will reject a false null hypothesis. In the framework of informative hypotheses, the equivalent to power is the so-called true hypothesis rate (Kuiper et al. [2015]). This rate is defined as the percentage of times in which most support is rendered for the true hypothesis or the one that is closest to the correct one.

In this first simulation run, results summarize the THR, median, minimum and maximum aggregated PMP that was reported for the true hypothesis.

4.2.4 Simulation results

Table 3 summarizes the simulation results. Overall, we can see that the method performs well for the evaluated scenarios. Given that neither of the two substantial hypotheses H_1 and H_2 is true in the population and it was sampled from H_3, the method independently of sample sizes and effect ratios renders most support for the unconstrained hypothesis H_u, as THRs fall between 0.91 and 1 for all evaluated scenarios. If either H_1 or H_2 is imposed to the population structure, and effect sizes are medium ($f^2 = 0.15$) or large ($f^2 = 0.35$), THRs greater than 0.8 are rendered for 19 of the 24 evaluated

Table 2: Population values in each study population

		Effect Ratios 2:1				Effect Ratios 1.33:1			
H_i	f^2	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
H_1	.02	.102	.051	.026	.013	.076	.057	.043	.032
	.15	.264	.132	.066	.032	.195	.147	.11	.083
	.35	.372	.186	.093	.047	.275	.207	.155	.117
H_2	.02	.045	.090	.045	.022	.051	.068	.051	.039
	.15	.116	.231	.116	.058	.133	.176	.133	.100
	.35	.163	.326	.163	.082	.187	.248	.187	.14
H_u	.02	.013	.026	.051	.102	.032	.043	.057	.076
	.15	.032	.066	.132	.264	.083	.11	.147	.195
	.35	.047	.093	.186	.372	.117	.155	.207	.275

scenarios. Merely in the cases in which sample sizes are low ($n = 50, 100$) and effect ratios are small (1.33:1), performance drops to THRs of 0.69 and 0.71 when H_1 is set true and to 0.64 and 0.74 when H_2 is imposed. If effect sizes are small ($f^2 = 0.02$), THRs mostly lie below 0.70, but are positively affected by larger effect ratios and sample sizes.

Summarizing, the method renders correct support reliably if effect sizes are medium or large. When effect sizes are small, THRs are highest if sample sizes and effect ratios are large. When neither of the investigated hypotheses is true, the aggregation correctly and consistently renders most support for the unconstrained hypothesis, independently of all subsequent simulation parameters.

4.3 Combining mixed evidence

4.3.1 Population values

In the first simulation run, one true order of regression coefficients H_m has been set equal for every evaluated study population that samples were drawn from. Similarly, in reality, there is only one true order of regression coefficients and one true Hypothesis H_m. Often, one would thus expect that to be combined evidence based on samples from the same underlying population has a similar structure. However, in applied research, evidence is usually mixed. For instance, this can be the case because different groups of researchers had access to specific parts of the underlying population that function differently. One research group could for example analyze undergraduate students of an US-American university, while another study draws inferences from a sample of male Dutch adults. Additionally, this can be a result of different studies specifying different statistical models with some being

Table 3: Descriptive statistics of overall PMPs for the correct hypothesis based on S=100 simulation iterations and four aggregated studies drawn from equal study populations.

			Effect Ratios 2:1				Effect Ratios 1.33:1			
H_i	f^2	n	THR	Med.	Min	Max	THR	Med.	Min	Max
H_1	.02	50	**.69**	.06	.00	.98	**.54**	.02	.00	.95
		100	**.69**	.25	.00	1	**.55**	.06	.00	.99
		500	**.92**	.76	.01	1	**.68**	.15	.00	.97
	.15	50	**.91**	.65	.00	1	**.69**	.13	.00	.98
		100	**.96**	.95	.00	1	**.71**	.41	.00	1
		500	**1**	1	.92	1	**.93**	.95	.04	1
	.35	50	**.96**	.97	.00	1	**.77**	.37	.00	.99
		100	**.99**	.99	.00	1	**.88**	.71	.00	1
		500	**1**	.99	1	1	**1**	1	.47	1
H_2	.02	50	**.65**	.08	.00	.94	**.59**	.04	.00	.90
		100	**.67**	.08	.00	.98	**.63**	.03	.00	.96
		500	**.9**	.56	.00	1	**.78**	.10	.00	.92
	.15	50	**.85**	.52	.00	1	**.64**	.09	.00	.93
		100	**.91**	.76	.00	1	**.74**	.15	.00	.99
		500	**.99**	1	.26	1	**.91**	.73	.02	1
	.35	50	**.91**	.85	.00	1	**.74**	.25	.00	.98
		100	**.91**	.76	.00	1	**.87**	.48	.00	.99
		500	**1**	1	.98	1	**.96**	.79	.08	1
H_u	.02	50	**.95**	1	.01	1	**.91**	.97	.01	1
		100	**.97**	1	.17	1	**.94**	.99	.04	1
		500	**1**	1	.99	1	**1**	1	.26	1
	.15	50	**1**	1	.42	1	**.97**	1	.06	1
		100	**1**	1	.64	1	**.98**	1	.26	1
		500	**1**	1	1	1	**1**	1	1	1
	.35	50	**1**	1	1	1	**1**	1	.06	1
		100	**1**	1	1	1	**1**	1	.65	1
		500	**1**	1	1	1	**1**	1	1	1

Note: True hypothesis rate (THR) represents percentage of times that most support is rendered for the true hypothesis. Median, Min and Max refer to overall PMPs that are aggregated over the four studies.

divergent from the data generating process in the population. Lastly, in specific samples of the same underlying population, varying hypotheses can obtain most support due to pure sampling error.

Thus, we deem it highly relevant to systematically assess the behavior of the aggregation method if mixed evidence is combined. One could be interested in assessing the method's performance if effect sizes vary over studies.

In the simulation run presented in Subsection 4.2, evidence for constant effect sizes has been aggregated. Thus, we have obtained the upper and lower bounds of the method's performance when a Hypothesis H_m is true in all study populations if effect sizes would be varied. Scenarios in which f^2 was set to 0.35 represent the upper bound of the method's performance if all studies obtain high effect sizes. When f^2 was set to 0.02, the lower bound if all studies obtain a low effect size can be inferred. Since the method overall performed well for both extremes, we will not investigate additional scenarios in which effect sizes vary between these bounds.

Rather, in a second simulation run, we exemplary examine three scenarios in which different Hypotheses H_m are imposed to be true over the four study populations. In this run, the method to construct population values follows the same procedure as outlined in 4.2.1. For constant sample sizes of $n = 100$, we examine scenarios in which samples are drawn from study populations that are defined by a variation of the hypotheses H_1 and H_3 and effect ratios $\mathbf{r}$ as presented in Table 4.

Table 4: Simulation scenarios for the evaluation of mixed results.

Parameter	Study ID			
	1	2	3	4
		Scenario 1		
H_m	H_1	H_1	H_1	H_3
f^2	0.15	0.15	0.15	0.15
$\boldsymbol{r}$	2:1	2:1	2:1	2:1
		Scenario 2		
H_m	H_1	H_1	H_2	H_3
f^2	0.15	0.15	0.15	0.15
$\boldsymbol{r}$	2:1	2:1	1.33:1	2:1
		Scenario 3		
H_m	H_1	H_2	H_3	H_3
f^2	0.15	0.15	0.15	0.15
$\boldsymbol{r}$	2:1	2:1	1.33:1	2:1

For Scenario 1, study populations 1 to 3 comply with one of the simulation scenarios investigated in the first simulation run, where H_1 was imposed to be true, f^2 was set to 0.15, ratios between the regression coefficients were $\mathbf{r} = 2$ and samples were of size $n = 100$. For this scenario, the aggregation method rendered most support for the true hypothesis in 96 per cent of the time, while rendering a median overall PMP of 0.95 for H_1. Now, by sampling from the fourth study population complying with H_3, mixed results will emerge. Following, these will be aggregated. For Scenario 2, H_2 is additionally imposed to the coefficient structure of study population 3. In Scenario 3, an additional study population is described by H_3.

4.3.2 Simulation procedure

Again, $N = 100$ samples of equal sample sizes are drawn from each study population and studies are aggregated over each Nth sample. We will thus evaluate every of the three exemplary scenarios described in Table 4 by 100 overall PMPs.

4.3.3 Performance evaluation

In the first simulation run, performance was primarily evaluated using the true hypothesis rate. When evaluating the method's performance based on an aggregation of mixed results, this concept is not applicable. This is because now, the true order of coefficients H_m is *varying* over the four studies. Rather, we will descriptively analyze the overall PMPs yielded by the aggregation.

4.3.4 Simulation results

Table 5 summarizes the method's performance for the evaluation of mixed results. As can be seen, the method turns out to behave highly sensitive in all of the evaluated scenarios. In Scenario 1, were three out of the four study populations are characterized by Hypothesis 1, 73 of the 100 simulation iterations rendered most support for the unconstrained hypothesis. With a mean of 0.74 and a median of 0.93, it clearly outperformed H_1. If more study populations are specified divergent from H_1, this outperformance becomes even more extreme with overall PMPs rendering most support for H_1 in 81 (Scenario 2) and 91 (Scenario 3) per cent of the times and the median of overall PMPs for H_u rising up to 0.99.

Table 5: Descriptive statistics of overall PMPs for each hypothesis based on S=100 simulation iterations and four aggregated studies.

	H_1			H_2			H_u		
	Perc.	Mean	Med.	Perc.	Mean	Med.	Perc.	Mean	Med.
Sc. 1	**0.27**	0.24	0.06	**0.00**	0.02	0.00	**0.73**	0.74	0.93
Sc. 2	**0.16**	0.17	0.02	**0.03**	0.06	0.00	**0.81**	0.77	0.94
Sc. 3	**0.04**	0.05	0.00	**0.05**	0.06	0.00	**0.91**	0.90	0.99

Note: Results reported for the three Scenarios (Sc.) described in Table 4. Perc. represents the per cent of times most support was rendered for the respective hypothesis. Mean and Med. report the arithmetic mean and median of all 100 overall PMPs.

5 Discussion

The present report introduced and evaluated a Bayesian method that can aggregate evidence for informative hypotheses over multiple studies of interest. The main goals were to guide the reader through the logic of Bayes factor construction, their aggregation using prior and posterior model probabilities and to assess the method's performance if evidence from multiple ordinary least squares regressions is combined.

The goal of the first simulation run was to evaluate the aggregation performance if evidence from equal study populations was combined. Results showed that the method yielded correct support reliably in almost all evaluated scenarios. However, researchers need to be cautious if evidence based on small sample and effect sizes is combined.

The goal of the second simulation run was to evaluate the method's performance if combined evidence is mixed. In these scenarios, the method showed a behavior that is highly sensitive. If only one of the study populations was described by a different hypothesis than the three subsequent others, support for the unconstrained hypothesis clearly outperformed all other expectations. It is doubtful whether this behavior is suitable for researchers aiming at applying it to scientific problems, as support would be expected to be highest for the hypothesis that characterized three out of the four study populations that were under consideration. However, this result might be due to an extreme specification of the divergent study population(s) that might not come across in a large number of realistic scenarios. Future research should further examine the method's applicability if mixed results are under consideration.

As a closing comment, we furthermore note that all simulation studies were based on the assumption that correct models that are equivalent to or a simplified form of the data generating process of the respective study population are applied in the drawn samples. Since in applied research, this will often not be the case, the method's performance might be examined if estimated models deviate from the data generation in the respective populations. Furthermore, the research presented here exclusively focused on the combination of evidence that was drawn from ordinary least squares regressions. Assessing the method's behavior when evidence from a variety of statistical models is applied will be a crucial task for future endeavors.

References

Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. doi: 10.1037/0033-2909.112.1.155. URL https://www.ncbi.nlm.nih.gov/pubmed/19565683.

Geoff Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966. URL https://doi.org/10.1177/0956797613504966.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. 2004. ISBN 158488388X (alk. paper). doi: 10.1198/tech.2004.s199. URL http://www.loc.gov/catdir/enhancements/fy0646/2003051474-d.html{%}5Cnhttp://pubs.amstat.org/doi/abs/10.1198/tech.2004.s199.

Xin Gu, Joris Mulder, Maja Deković, and Herbert Hoijtink. Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19 (4):511–527, 2014. ISSN 1939-1463. doi: 10.1037/met0000017. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000017.

Xin Gu, Joris Mulder, and Herbert Hoijtink. Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, (024):1–43, 2017. ISSN 20448317. doi: 10.1111/bmsp.12110. URL http://onlinelibrary.wiley.com/doi/10.1111/bmsp.12110/pdf.

Herbert Hoijtink. *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. 2012. ISBN 9781439880524.

Irene Klugkist, Bernet Kato, and Herbert Hoijtink. Bayesian model selection using encompassing priors. *Statistica Neerlandica*, 59(1):57–69, 2005a. ISSN 1467-9574. doi: 10.1111/j.1467-9574.2005.00279.x. URL http://dx.doi.org/10.1111/j.1467-9574.2005.00279.x.

Irene Klugkist, Olav Laudy, and Herbert Hoijtink. Inequality constrained analysis of variance: a Bayesian approach. *Psychological methods*, 10(4): 477–493, dec 2005b. ISSN 1082-989X (Print). doi: 10.1037/1082-989X.10.4.477.

Rebecca M Kuiper, Vincent Buskens, Werner Raub, and Herbert Hoijtink. Combining Statistical Evidence From Several Studies. *Sociological Methods & Research*, 42(1):60–81, 2013. ISSN 0049-1241. doi: 10.1177/0049124112464867. URL http://journals.sagepub.com/doi/10.1177/0049124112464867.

Rebecca M. Kuiper, Tim Nederhoff, and Irene Klugkist. Properties of hypothesis testing techniques and (Bayesian) model selection for exploration-based and theory-based (order-restricted) hypotheses. *British Journal of Mathematical and Statistical Psychology*, 68(2):220–245, 2015. ISSN 20448317. doi: 10.1111/bmsp.12041.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–aac4716, 2015. ISSN 0036-8075. doi:

10.1126/science.aac4716. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716.