Journal of Statistical Software

November 2021, Volume 100, Issue 18. doi: 10.18637/jss.v100.i18

BFpack: Flexible Bayes Factor Testing of Scientific Theories in R

Joris Mulder Tilburg University
Donald R. Williams UC Davis
Xin Gu East China Normal University
Andrew Tomarken Vanderbilt University
Florian Böing-Messing Jheronimus Academy of Data Science
Anton Olsson-Collentine Tilburg University
Marlyne Meijerink Tilburg University
Janosch Menke Utrecht University
Robbie van Aert Tilburg University
Jean-Paul Fox University of Twente
Herbert Hoijtink Utrecht University
Yves Rosseel Ghent University
Eric-Jan Wagenmakers University of Amsterdam
Caspar van Lissa Utrecht University

Abstract

There have been considerable methodological developments of Bayes factors for hypothesis testing in the social and behavioral sciences, and related fields. This development is due to the flexibility of the Bayes factor for testing multiple hypotheses simultaneously, the ability to test complex hypotheses involving equality as well as order constraints on the parameters of interest, and the interpretability of the outcome as the weight of evidence provided by the data in support of competing scientific theories. The available software tools for Bayesian hypothesis testing are still limited however. In this paper we present a new R package called **BFpack** that contains functions for Bayes factor hypothesis testing for the many common testing problems. The software includes novel tools for (i) Bayesian exploratory testing (e.g., zero vs positive vs negative effects), (ii) Bayesian confirmatory testing (competing hypotheses with equality and/or order constraints), (iii) common statistical analyses, such as linear regression, generalized linear models, (multivariate) analysis of (co)variance, correlation analysis, and random intercept models, (iv) using default priors, and (v) while allowing data to contain missing observations that are missing at random.

Keywords: Bayes factors, hypothesis testing, equality/order constrained hypotheses, R.

1. Introduction

This paper presents the software package **BFpack** which can be used for computing Bayes factors and posterior probabilities for statistical hypotheses in common testing problems in the social and behavioral sciences, medical research, and in related fields. Package **BFpack** (Mulder *et al.* 2021) is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=BFpack`. This new package is an answer to the increasing interest of the scientific community to test statistical hypotheses using Bayes factors in the software environment R (R Core Team 2021). Bayes factors enjoy many useful practical and theoretical properties which are not generally shared by classical significance tests (Jeffreys 1961; Berger and Delampady 1987; Sellke, Bayarri, and Berger 2001; Mulder and Wagenmakers 2016). These include its intuitive interpretation as the relative evidence in the data between two hypotheses (Wagenmakers 2007; Rouder, Speckman, Sun, Morey, and Iverson 2009), its ability to simultaneously test multiple hypotheses which may contain equality as well as order constraints on the parameters of interest (Hoijtink 2011), and its consistent behavior which implies that the true hypothesis will be selected with probability one as the sample size grows (Berger and Delampady 1987; Kass and Raftery 1995). This has resulted in an increasing literature where Bayes factors have been used for testing scientific expectations in many different fields of research including, but not limited to, gender research (Well, Kolk, and Klugkist 2008), management research (Braeken, Mulder, and Wood 2015), econometrics (Koop and Potter 1999), research on psychological contracts (De Jong, Rigotti, and Mulder 2017), criminal linkage (Porter 2016), opinion swing in political science (Irony, de Pereira, and Tiwari 2000), and clinical trial research (Van Ravenzwaaij, Monden, Tendeiro, and Ioannidis 2019). For a thorough overview of the merits of Bayesian inference in applied research, see for example Wagenmakers *et al.* (2018).

The Bayes factors that are implemented in **BFpack** are based on recent developments of Bayesian hypothesis testing of equality and order constraints on location parameters, such as (adjusted) means and regression coefficients (Mulder and Gu 2021; Gu, Hoijtink, Mulder, and Rosseel 2019; Gu, Mulder, and Hoijtink 2017; Mulder 2014), variance components, such as group variances and intraclass correlations (Böing-Messing, Van Assen, Hofman, Hoijtink, and Mulder 2017a; Mulder and Fox 2019), and measures of association, (Mulder 2016; Mulder and Gelissen 2019). These Bayes factors can be used for common testing problems in the social and behavioral sciences, and related fields, such as (multivariate) t testing, (multivariate) linear regression, (multivariate) analysis of (co)variance, or correlation analysis. The package allows users to perform (i) exploratory Bayesian tests of whether a model parameter is zero, negative, or positive, and (ii) confirmatory Bayesian tests where users manually specify a set of competing hypotheses with equality and/or order constraints on the parameters of interest. This will allow users to test their scientific expectations in a direct manner. Thus by providing Bayesian statistical tests for multiple hypotheses with equality as well as order constraints, **BFpack** makes important contributions to existing software packages, such as **lmtest** (Zeileis and Hothorn 2002) and **car** (Fox and Weisberg 2021), which contain key functions for classical significance tests of a single equality constrained hypothesis, e.g., `lmtest::coeftest()` and `car::linearHypothesis()`.

To ensure a simple and user-friendly experience, the different Bayes factors tests are implemented via a single function called `BF()`, which is the workhorse of the package. The function needs a fitted modeling object obtained from a standard R analysis (e.g., `lm`, `glm`; see Table 1

R function	Package	Test	Tested parameter	Parameter name	Bayes factor
`t_test`	**bain**	Student t test	mean (1-sample test)	`mu`	AFBF
			mean difference (2-sample test)	`difference`	AFBF
`bartlett_test`	**BFpack**	heterogeneity of variances	group variances	`g1`	AFBF
`aov`	**stats**	AN(C)OVA	group means	`g1`	AFBF
`manova`	**stats**	MAN(C)OVA	group means	`g1_on_y1`	AFBF
`lm`	**stats**	linear regression	regression coefficients	`x1`	AFBF
		multivariate regression	regression coefficients	`x1_on_y1`	AFBF
`cor_test`	**BFpack**	correlation analysis	measures of association	`y1_with_y2`,	uniform priors
				`y1_with_y2_in_g1`	uniform priors
`lmer`	**lme4**	random intercept model	group specific intraclass correlations	`g1`	uniform priors
`rma`	**metafor**	meta-analysis	between-study heterogeneity,	`I^2`,	uniform prior,
			effect size	`mu`	unit-information prior
`glm`	**stats**	generalized linear model	regression coefficients	`x1`	approx. AFBF
`coxph`, `survreg`	**survival**	survival analysis	regression coefficients	`x1`	approx. AFBF
`polr`	**MASS**	ordinal regression	regression coefficients	`x1`	approx. AFBF
`zeroinfl`	**pscl**	zero-inflated regression models	regression coefficients	`x1`	approx. AFBF

Table 1: R functions, packages, type of test, tested parameters, example name of a tested parameter, and the Bayes factor and prior that is used. Note: "AFBF" refers to the adjusted fractional Bayes factor (Appendix A.1) and "approx. AFBF" refers to the adjusted fractional Bayes factors using Gaussian approximations (Appendix A.3).

	Examples of hypotheses
Precise testing	$H_1 : \theta = 0$ vs $H_2 : \theta \neq 0$.
One-sided testing	$H_1 : \theta \leq 0$ vs $H_2 : \theta > 0$.
Interval testing	$H_1 : \lvert\theta\rvert \leq \epsilon$ vs $H_2 : \lvert\theta\rvert > \epsilon$, for given $\epsilon > 0$.
Exhaustive testing	$H_1 : \theta = 0$ vs $H_2 : \theta < 0$ vs $H_3 : \theta > 0$.
Precise testing	$H_1 : \theta_1 = \theta_2 = \theta_3$ vs H_2 : "not H_1"
Order testing	$H_1 : \theta_1 > \theta_2 > \theta_3$ vs $H_2 : \theta_1 < \theta_2 < \theta_3$ vs H_3 : "neither H_1, nor H_2".
Equality and order testing	$H_1 : \theta_1 < \theta_2 = \theta_3$ versus H_2 : "not H_1".

Table 2: Examples of hypothesis tests that can be executed using **BFpack**.

for a complete overview), and in the case of a confirmatory test a string that specifies a set of competing hypotheses (examples of hypotheses are provided in Table 2). Another optional argument is the specification of the prior probabilities for the hypotheses. By building on these traditional statistical analyses, which are well-established by the R community, we present users additional statistical measures which cannot be obtained under a frequentist framework, such as quantification of the relative evidence in the data between a broad class of statistical hypotheses.

When testing hypotheses using the Bayes factor, the use of arbitrary or ad hoc priors should generally be avoided (Lindley 1957; Jeffreys 1961; Bartlett 1957; Berger and Pericchi 2001). Therefore the implemented tests in **BFpack** are based on default Bayes factor methodology. Default Bayes factors can be computed without requiring external prior knowledge about the magnitude of the parameters. The motivation is that, even in the case prior information is available, formulating informative priors which accurately reflect one's prior beliefs under all separate hypotheses under investigation is a very challenging and time-consuming endeavor (Berger 2006).

Different default Bayes factors with default priors are implemented for testing different types of parameters, such as location parameters. For testing unbounded parameters, such as location parameters and group variances, adjusted fractional Bayes factors (O'Hagan 1995; Mulder 2014; Böing-Messing *et al.* 2017a) have been implemented. These Bayes factors have analytic expressions and are therefore easy to compute. The implied fractional priors contain minimal information so that maximal information in the data is used for hypothesis testing (O'Hagan 1995; Berger and Mortera 1995; Conigliani and O'Hagan 2000). For testing bounded parameters, such as measures of association, intraclass correlations, or between-study heterogeneity, proper uniform priors are implemented. When testing intraclass correlations under random intercept models or the between-study heterogeneity in a meta-analysis, a novel marginal modeling approach is employed where the random effects are integrated out (Mulder and Fox 2019; Fox, Mulder, and Sinharay 2017; Mulder and Fox 2013; Van Aert and Mulder accepted). Besides testing hypotheses based on substantive expectations, testing intraclass correlations is also useful for building multilevel models as the marginal model approach provides a more general framework for testing covariance structures than regular mixed effects models.

To also facilitate the use of Bayes factors for more general testing problems, an approximate Bayes factor is also implemented which is based on a large sample approximation of the posterior having an approximate Gaussian distribution. The approximate Bayes factor only requires the (classical) estimates of the parameters that are tested, the corresponding error covariance matrix, and the sample size of the data that was used to get the estimates and covariance matrix. The resulting approximated Bayes factor can be viewed as a Bayesian counterpart of the classical Wald test. This makes the approximate Bayes factor very useful as a general test for statistical hypotheses when exact tests are not available. In Section 4.4 we show how to obtain perform an approximate Bayesian hypothesis test using the output of `lmtest::coeftest()`. Table 1 shows for which models an exact Bayes factor is implemented and for which models we make use of the approximation.

Before presenting the statistical methodology and functionality of **BFpack** it is important to understand what **BFpack** adds to the currently available software packages for Bayes factor testing. First, the R package **BayesFactor** (Morey, Rouder, Jamil, Urbanek, Forner, and Ly 2018) mainly focuses on precise and interval null hypotheses of single parameters in Student t tests, anova designs, and regression models. It is not designed for testing more complex relationship between multiple parameters. Second, the package **BIEMS** (Mulder, Hoijtink, and De Leeuw 2012), which comes with a user interface for Windows, can be used for testing various equality and order hypotheses under the multivariate normal linear model. The computation of the Bayes factors however is too slow for general usage when simultaneously testing many equality constraints as equality constraints are approximated with interval constraints that are made sufficiently small using a computationally intensive step-wise algorithm. Third, the **bain** package (Gu *et al.* 2021) computes approximated default Bayes factors by assuming normality of the posterior and a default prior. The package has shown good performance for challenging testing problems such as structural equation models. For approximate Bayes factors, **BFpack** package also builds on the functionality of **bain** for combinations of order constraints that cannot be written in matrix notation of full row-rank. Unlike **bain** however, **BFpack** utilizes existing R functions such as `dmvnorm()` or `pmvnorm()` from the **mvtnorm** package (Genz, Bretz, Miwa, Mi, and Hothorn 2021) and therefore does not contain numerical errors due to random sampling. Finally the free statistical software environment **JASP** (Love *et al.* 2019), which has contributed tremendously to the use of Bayes factors in psychological

research and other research fields, is specifically designed for non-R users by providing a user-friendly graphical user-interface similar to SPSS. Under the hood, **JASP** calls R functions for Bayes factor testing, such as several functions implemented in the R packages **BayesFactor** or **bain**. **BFpack**, on the other hand, is developed to give R users a flexible tool for testing a very broad class of hypotheses involving equality and/or order constraints on various types of parameters (means, regression coefficients, variance components, and measures of association) under common statistical models by building on standard R functions. Finally note that Bayes factors are also implemented in other R packages for various purposes, such as the **condir** package for human threat conditioning research (Krypotos, Klugkist, and Engelhard 2017), the **BAS** package for Bayesian variable selection and model averaging (Clyde, Ghosh, and Littman 2018), **BayesMed** for Bayesian mediation analysis (Nuijten, Wetzels, Matzke, Dolan, and Wagenmakers 2014), or **BayesVarSel** for objective Bayesian variable selection in linear models (Garcia-Donato, Forte, and Vergara-Hernandez 2020; Garcia-Donato and Forte 2018).

The paper is organized as follows. Section 2 describes the theoretical background of Bayes factors and posterior probabilities Section 3 gives a general explanation about the usage of the main function `BF()` in **BFpack**. Section 4 presents 8 different applications of the methodology and software for a variety of testing problems. The paper ends with some concluding remarks in Section 5.

2. Theoretical background of Bayesian statistical inference

Let us consider a statistical model which contains a vector of Q key parameters of interest, denoted by $\boldsymbol{\theta}$, while $\boldsymbol{\phi}$ contains the V nuisance parameters of the model. Depending on the goal of the analysis, the parameters of interest and nuisance parameters may vary. For example, in an analysis of variance, the group means are the parameters of interest and the within group variance is a nuisance parameter (Klugkist, Laudy, and Hoijtink 2005). On the other hand when one is interested in testing an equality or order of the group variances (Böing-Messing *et al.* 2017a), the group variances are the parameters of interest and the group means are treated as nuisance parameters.

BFpack consists of a variety Bayes factors for testing a set of T hypotheses with linear equality and/or order constraints on the key parameters $\boldsymbol{\theta}$ of the form

$$H_t : \mathbf{R}^e\boldsymbol{\theta} = \mathbf{r}^e \;\&\; \mathbf{R}^o\boldsymbol{\theta} > \mathbf{r}^o, \tag{1}$$

where $[\mathbf{R}^e \mid \mathbf{r}^e]$ is a $r^e \times (Q+1)$ augmented matrix specifying the equality constraints on $\boldsymbol{\theta}$ under H_T, and $[\mathbf{R}^o \mid \mathbf{r}^o]$ is a $r^o \times (Q+1)$ augmented matrix specifying the order (or one-sided) constraints on $\boldsymbol{\theta}$ under H_T, for $t = 1, \ldots, T$ constrained hypotheses. A hypothesis index is omitted in the restriction matrices $[\mathbf{R}^e \mid \mathbf{r}^e]$ and $[\mathbf{R}^o \mid \mathbf{r}^o]$ to simplify the notation. Examples of constrained hypotheses are given in Table 2. By default **BFpack** executes a standard (exploratory) hypothesis test. The hypotheses that are tested in the exploratory analysis depend on the class of the fitted model (Section 3 and Section 4). If a user is interested in executing a specific (confirmatory) hypothesis test based on scientific expectations, where hypotheses contain competing equality and/or order constraints as in Equation 1, the user can manually specify the constrained hypotheses (Section 3 and Section 4). Finally note that the notation H_0, for the traditional "null" hypothesis of "no effect", is not used as **BFpack**

builds on the idea that researchers can be flexible in terms of (i) the types of hypotheses to test and (ii) the number of hypothesis to test, instead of restricting to only testing the traditional (precise) null hypothesis against a two-sided alternative.

The Bayes factor between two constrained hypotheses, say, H_1 against H_2, is defined by the ratio of the marginal likelihoods under the two hypotheses, i.e.,

$$B_{12} = \frac{p_1(\mathbf{Y})}{p_2(\mathbf{Y})} = \frac{\int\int p_1(\mathbf{Y} \mid \boldsymbol{\theta}_1, \boldsymbol{\phi}_1)\pi_1(\boldsymbol{\theta}_1, \boldsymbol{\phi}_1)d\boldsymbol{\theta}_1 d\boldsymbol{\phi}_1}{\int\int p_1(\mathbf{Y} \mid \boldsymbol{\theta}_2, \boldsymbol{\phi}_2)\pi_2(\boldsymbol{\theta}_2, \boldsymbol{\phi}_2)d\boldsymbol{\theta}_2 d\boldsymbol{\phi}_2}, \tag{2}$$

where the marginal likelihood, denoted by $p_t(\mathbf{Y})$ under H_t, is computed as the integral over the product of the likelihood of the data $\mathbf{Y}$, denoted by $p_t(\mathbf{Y} \mid \boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$ under hypothesis H_t, and the prior, denoted by $\pi_t(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$ under H_t, over the respective constrained parameter spaces of the free parameters $\boldsymbol{\theta}_t$ and $\boldsymbol{\phi}_t$ under hypothesis H_t, which is given by $\{\boldsymbol{\theta} \mid \mathbf{R}^e\boldsymbol{\theta} = \mathbf{r}^e \;\&\; \mathbf{R}^o\boldsymbol{\theta} > \mathbf{r}^o\}$. As the marginal likelihood quantifies the predictive performance of a prior and hypothesis of the observed data, the Bayes factor B_{12} can be interpreted as a relative measure of the evidence in the data for hypothesis H_1 relative to hypothesis H_2 (Kass and Raftery 1995). For example, if $B_{12} = 1$, this implies that both hypotheses receive equal support from the data, or if $B_{12} = 30$, this implies that H_1 receives 30 times more evidence than H_2, which would suggest strong evidence in favor of H_1 against H_2. In **BFpack** Bayes factors are first computed of each constrained hypothesis against an unconstrained hypothesis, denoted by H_u, which does not assume any constraints under the model of interest, and subsequently the transitivity relationship is used to obtain Bayes factors between the constrained hypotheses, e.g., $B_{12} = B_{1u}/B_{2u}$.

When specifying prior probabilities for the hypotheses under investigation, Bayes factors can be used to obtain the posterior probabilities of the hypotheses using

$$\mathsf{P}(H_t \mid \mathbf{Y}) = \frac{\mathsf{P}(H_t)B_{tu}}{\sum_{t'=1}^{T}\mathsf{P}(H_{t'})B_{t'u}}, \tag{3}$$

where $\mathsf{P}(H_t)$ denotes the prior probability that H_t is true before observing the data, such that $\sum_{t'=1}^{T}\mathsf{P}(H_{t'}) = 1$ holds. The posterior probability of H_t quantifies the plausibility that hypothesis H_t is true after observing the data under the assumption that one of the hypotheses under investigation is true. Equation 3 shows how the posterior probability combines the evidence in the data (through the Bayes factors) and the prior probabilities. Equivalently, when considering two hypotheses, the Bayes factor is used to update the prior odds to obtain the posterior odds according to

$$\frac{\mathsf{P}(H_1 \mid \mathbf{Y})}{\mathsf{P}(H_2 \mid \mathbf{Y})} = B_{12} \times \frac{\mathsf{P}(H_1)}{\mathsf{P}(H_2)}. \tag{4}$$

In **BFpack** equal prior probabilities are specified by default, i.e., $\mathsf{P}(H_t) = \frac{1}{T}$. In this case, the posterior odds between hypotheses will be equal to the Bayes factor (Equation 4). Users can manually specify other choices for the prior probabilities depending on the context.

Guidelines for interpreting Bayes factors and posterior probabilities have been provided in the literature (Jeffreys 1961); see Table 3. It is important to note however that these guidelines should not be used in a strict sense as the evidence for a hypothesis lies on a continuous scale. In the end it is up to the scientific community to judge when enough evidence has been collected to conclude whether a hypothesis is true or not.

B_{12}	$\mathsf{P}(H_1 \mid \mathbf{Y})$	Evidence for H_1
1 to 3.2	0.5 to 0.76	Mild
3.2 to 10	0.76 to 0.91	Substantial
10 to 32	0.91 to 0.97	Strong
32 to 100	0.97 to 0.99	Very strong
> 100	> 0.99	Decisive

Table 3: Guidelines for interpreting the Bayes factors B_{12} (Jeffreys 1961) and posterior probability of H_1 assuming equal prior probabilities.

To facilitate the interpretation, posterior probabilities can also be interpreted in terms of conditional error probabilities given the observed data (Berger, Brown, and Wolpert 1994; Hoijtink, Mulder, Van Lissa, and Gu 2019c). For example, when one would conclude that hypothesis H_1 is true based on a posterior probability of 0.80, there would be a conditional probability of 0.20 of drawing the wrong conclusion given the observed data. This idea may guide researchers who are relatively new to Bayesian statistics when interpreting Bayes factors and posterior probabilities.

As the computation of marginal likelihoods in Equation 2 can be expensive (Kass and Raftery 1995), **BFpack** utilizes a special form of the Bayes factor which avoids computing marginal likelihoods when possible. The special expression is referred to as the extended Savage-Dickey density ratio, which is defined by

$$\begin{aligned} B_{tu} &= B^e_{tu} \times B^o_{tu} \\ &= \frac{\pi_u(\boldsymbol{\theta}^e = \mathbf{r}^e \mid \mathbf{Y})}{\pi_u(\boldsymbol{\theta}^e = \mathbf{r}^e)} \times \frac{P_u(\boldsymbol{\theta}^o > \mathbf{r}^o \mid \boldsymbol{\theta}^e = \mathbf{r}^e, \mathbf{Y})}{P_u(\boldsymbol{\theta}^o > \mathbf{r}^o \mid \boldsymbol{\theta}^e = \mathbf{r}^e)}, \end{aligned} \tag{5}$$

where $\boldsymbol{\theta}^e = \mathbf{R}^e\boldsymbol{\theta}$ and $\boldsymbol{\theta}^o = \mathbf{R}^o\boldsymbol{\theta}$ (computational details can be found in Appendix A). The Bayes factor in Equation 5 holds when the prior under the constrained hypothesis is proportional to the prior under the unconstrained hypothesis, and zero elsewhere. The extended Savage-Dickey density ratio for an equality/order constrained hypothesis against an unconstrained alternative was reported in Mulder and Gelissen (2019), which builds on earlier work (Dickey 1971; Klugkist *et al.* 2005; Pericchi, Liu, and Torres 2008; Wetzels, Grasman, and Wagenmakers 2010; Mulder, Hoijtink, and Klugkist 2010; Gu *et al.* 2017; Mulder, Wagenmakers, and Marsman 2020b, among others). Interestingly, the four statistical measures in Equation 5 explicitly show how the Bayes factor balances between fit and complexity when evaluating constrained hypothesis (similar, in a way, as the AIC or BIC):

- The marginal posterior density evaluated at $\boldsymbol{\theta}^e = \mathbf{r}^e$ (numerator of first factor) is a measure of the *relative fit of the equality constraints* of H_t relative to H_u because a large (small) posterior value under the unconstrained model indicates that there is evidence in the data that $\boldsymbol{\theta}^e$ is (not) close to $\mathbf{r}^e$.
- The conditional posterior probability of $\boldsymbol{\theta}^o > \mathbf{r}^o$ given $\boldsymbol{\theta}^e = \mathbf{r}^e$ (numerator of second factor) is a measure of the *relative fit of the order constraints* of H_t relative to H_u because a large (small) probability under the unconstrained model indicates that there is evidence in the data that the order constraints (do not) hold.
- The marginal prior density evaluated at $\boldsymbol{\theta}^e = \mathbf{r}^e$ (denominator of first factor) is a measure of the *relative complexity of the equality constraints* of H_t relative to H_u because

a large (small) prior value indicates that the prior for $\boldsymbol{\theta}^e$ is (not) concentrated around $\mathbf{r}^e$, and thus there is little (big) difference between the precise formulation $\boldsymbol{\theta}^e = \mathbf{r}^e$ and the unconstrained formulation H_u.

- The conditional prior probability of $\boldsymbol{\theta}^o > \mathbf{r}^o$ given $\boldsymbol{\theta}^e = \mathbf{r}^e$ (denominator of second factor) is a measure of the *relative complexity of the order constraints* of H_t relative to H_u because a large (small) probability under the unconstrained model indicates that the order constrained subspace under H_t is relatively large (small), indicating that the constrained model is complex (simple).

As Bayes factors are known to be sensitive to the choice of the prior, arbitrary or ad hoc prior specification should be avoided. **BFpack** has several default Bayes factor and prior specification methods implemented depending on the nature of the statistical model and its key parameters. In the case of testing location parameters (Mulder and Gu 2021) and group variances (Böing-Messing *et al.* 2017a), generalized adjusted fractional Bayes factors are used based on minimal fractions under all groups where the implicit fractional prior is adjusted to the boundary of the constrained space (e.g., to the test value for a Bayesian t test). The fractional prior is located to the boundary of the constrained space to abide the rational that small effects are more plausible a priori than large effects (typical in applied research) and that negative effects are equally plausible as positive effects (Mulder 2014; Mulder and Olsson-Collentine 2019). The default Bayes factor is fully automatic for a given set of constrained hypotheses, and thus a prior scale of the effects does not need to be specified based on prior expectations about the anticipated effects. Adjusted fractional Bayes factors based on Gaussian approximations are used under statistical models when an exact expression of the fractional Bayes factor is unavailable (Gu *et al.* 2017). When testing measures of association (Mulder and Gelissen 2019) and intraclass correlations (Mulder and Fox 2019), which are bounded in an interval, proper uniform priors are used. An overview of the technical details of these Bayes factors can be found in Appendix A.

3. Bayes factor testing using the package

Bayes factor tests can be executed by calling function `BF()`. The function has the following arguments:

- `x`, a fitted model object that is obtained using a R function. An overview of R functions that are currently supported can be found in Table 1.

- `hypothesis`, a string specifying the hypotheses with equality and/or order constraints on the key parameters of interest. To get an overview of the key parameters that can be tested under a fitted model object.

  - The default setting is `hypothesis = NULL`, in which case only exploratory tests on the key parameters are executed. The tests that are executed in the exploratory test are discussed below.
  - The parameter names are based on the names of the estimated key parameters. An overview of the key parameters is given using function `get_estimates()` (i.e., `get_estimates(model1)`, where `model1` is a fitted model object). For example, if the coefficients in a fitted `lm` object, say, `fit1`, have the names `weight`,

height, and length, and the constraints in the hypothesis argument should be formulated on these character strings.

- Separate constraints within a hypothesis are separated with an ampersand "&". Hypotheses are separated using a semi-colon ";". For example hypothesis = "weight > height & height > 0; weight = height = 0" implies that the first hypothesis assumes that the parameter weight is larger than the parameter height and that the parameter height is positive, and the second hypothesis assumes that the two parameters are equal to zero. Note that the first hypothesis could equivalently have been written as weight > height > 0.
- Brackets, "(" and ")", can be used to combine constraints of multiple hypotheses. For example hypothesis = "(weight, height, length) > 0" denotes a hypothesis where the parameters weight, height, and length are positive. This could equivalently have been written as hypothesis = "weight > 0 & height > 0 & length > 0".
- In general we recommended not to specify order hypotheses that are nested, such as hypothesis = "weight > height > length; weight > (height, length)", where the first hypothesis (which assumes that weight is larger than height and that height is larger than length) is nested in the second hypothesis (which assumes that weight is largest and no constraints are specified between height and length). The reason is that the Bayes factor for the simpler hypothesis against the more complex hypothesis would then be bounded. Therefore the scale of the Bayes factor would become more difficult to interpret, and the evidence could not accumulate to infinity for the true hypothesis if the true hypothesis would be the smaller order hypotheses (e.g., see Mulder *et al.* 2010). If however a researcher has theoretical reasons to formulate nested order hypotheses these can be formulated and tested using the BF() function of course.
- When testing hypotheses on group variances on an object of class 'bartlett_htest' (Table 1), only simple constraints are allowed where a variance is equal to, greater than, or smaller than another variance.

- prior.hyp, a numeric vector specifying the prior probabilities for the hypotheses in the hypothesis argument in a confirmatory test. The default setting is prior.hyp = NULL which sets equal prior probabilities.

- complement, a logical value which specified if a complement hypothesis is included in the tested hypotheses specified under hypothesis. The default setting is TRUE. The complement hypothesis covers the remaining parameters space that is not covered by the constrained hypotheses. For example, if an equality hypothesis and an order hypothesis are formulated, say, hypothesis = "weight = height = length; weight > height > length", the complement hypothesis covers the remaining subspace where neither "weight = height = length" holds, nor "weight > height > length" holds.

In the case the class of the fitted model x is not supported, BF.default() is called which executes an approximate fractional Bayes factor test (Section A.3). In this case, the following (additional) arguments are required:

- `x`, a named numeric vector of the estimates (e.g., MLE) of the parameters of interest where the labels are equal to the names of the parameters which are used for the `hypothesis` argument.
- `Sigma`, the approximate posterior covariance matrix (e.g., error covariance matrix) of the parameters of interest.
- `n`, the sample size that was used to acquire the estimates and covariance matrix.

When running the `BF()` function on a fitted model, an exploratory test is always executed (when `hypothesis = NULL`, only an exploratory test is executed). In an exploratory test the models that are considered are based on the full model with different restrictions under the constrained hypotheses. The exploratory tests that are executed depend on the class of the object `x`. The choice of the tests is based on the standard tests that are executed for an object of this class. In (M)AN(C)OVA, when `x` is of class '`aov`', '`maov`', or '`manova`', and the factors are modeled as 0/1 dummy covariates, Bayes factors are computed for exploratory testing the separate main effects and the separate interaction effects (if present in the model). The motivation is that such tests are of main interest when performing (M)AN(C)OVA analyses. Section 4.2 presents for an example ANOVA analysis. When `x` is of class '`bartlett_htest`', i.e., when testing group variances, the exploratory analysis tests whether there is homogeneity of variances or not, again similar as in the standard analysis for a Bartlett test. Section 4.3 presents for an example analysis. For all other classes, the exploratory analysis executes exhaustive tests (Table 2) of whether each separate parameter is zero, negative, or positive (i.e., $\theta = 0$ versus $\theta < 0$ versus $\theta > 0$). Exhaustive tests are executed instead of a standard precise tests (i.e., $\theta = 0$ versus $\theta \neq 0$) because the exhaustive test also gives insight about the direction of a possible effect. In each exploratory test equal prior probabilities are used for the hypotheses.

A confirmatory hypothesis test is executed if one or more constrained hypotheses are specified using the `hypothesis` argument. A constrained hypothesis has equality and/or order constraints on the key parameters given in Equation 1. These constrained hypotheses can be based on prior expectations, scientific expectations, or formal substantive theories.

The output is an object of class '`BF`'. When printing the summary of the object, using the `summary()` function, the following is presented:

- The resulting posterior probabilities of the hypotheses in the exploratory tests.
- An extensive overview of the results of the confirmatory test if constrained hypotheses are formulated using the `hypothesis` argument, which includes
  - the posterior probabilities of the hypotheses in the confirmatory test,
  - the `Evidence matrix` containing the Bayes factors between all pairs of hypotheses in the confirmatory test,
  - the `Specification table` containing the quantities in the extended Savage-Dickey density ratio for the hypotheses in the confirmatory test, i.e.,
    - The first column "`complex=`" contains the relative complexity of the equality part ("=") of a constrained hypothesis (the denominator in the first factor of Equation 5).

- The second column "`complex>`" contains the relative complexity of the order (or one-sided) part (">") of a constrained hypothesis (the denominator in the second factor of Equation 5).
- The third column "`fit=`" contains the relative fit of the equality part ("=") of a constrained hypothesis (the numerator in the first factor of Equation 5).
- The fourth column "`fit>`" contains the relative fit of the order (or one-sided) part (">") of a constrained hypothesis (the numerator in the second factor of Equation 5).
- The fifth column "`BF=`" contains B^e_{tu} in Equation 5.
- The sixth column "`BF>`" contains B^o_{tu} in Equation 5.
- The seventh column "`BF`" contains the Bayes factor of a constrained hypothesis against an unconstrained alternative, $B_{tu} = B^e_{tu} \times B^o_{tu}$, in Equation 5.
- The eighth column displays the posterior probabilities of the hypotheses which combines the evidence in the data (quantified by the Bayes factor) and the prior probabilities.

– and the hypotheses that are tested in the confirmatory test.

For Bayes factors that cannot be expressed as a Savage-Dickey density ratio (which is the case when testing equality constraints on variance components), `NA`s are printed in the first and third column of the `Specification table`.

When printing an object of this class, using the `print()` function, the results of the exploratory test is printed when `hypothesis = NULL`, and the results of the confirmatory test is printed when hypotheses are specified in the `hypothesis` argument. As the number of separate tests that are executed in an exploratory analysis can be very large, only the posterior probabilities of the hypotheses are printed, and not all separate evidence matrices, not to overwhelm the user with too much output. For the exploratory tests and the confirmatory tests, all Bayes factors for each constrained hypothesis against the unconstrained (full) model can be extracted from the `BF` object by taking the element `BFtu_exploratory` or the element `BFtu_confirmatory`, respectively. These can be used to compute the Bayes factors between the constrained hypotheses using the transitive relationship, e.g., $B_{12} = B_{1u}/B_{2u}$.

4. Applications

This section presents eight empirically motivated analyses using **BFpack** on different types of statistical models. Each subsection is discussed using the same format: first the statistical model and standard exploratory hypothesis test is presented, followed by a confirmatory test in applied research field, and finally the analysis is discussed using **BFpack**. In some subsections additional statistical elaborations are provided to give readers more insight about certain aspects of the method or to discuss possible extensions. On https://github.com/cjvanlissa/BFpack_paper, a R/Markdown version of these applications can be found to facilitate the reproducibility of the analyses.

4.1. Bayesian t testing

Statistical model and exploratory hypothesis test

The t test is one of the most commonly used statistical tests in applied research. In the case of a one-sample t test it is tested whether a normal mean μ is equal to a constant or not when data follows a normal distribution, $N(\mu, \sigma^2)$, where σ^2 is an unknown population variance which serves as nuisance parameter. In the case of a two-sample t test it is tested whether the mean difference between two normally distributed independent samples or dependent (paired) samples, i.e., $\theta = \mu_1 - \mu_2$, where μ_1 and μ_2 are the respective normal population means, equals zero or not.

In order to execute a Bayesian t test using **BFpack**, first a classical t test needs to be performed using the `t_test()` function. This function is equivalent to the standard `t.test()` function except that the fitted object also contains the sample size(s) and the sample variance(s), which are needed to compute Bayes factors. Next the fitted object is plugged into the `BF()` function to perform a Bayesian t test using the generalized fractional Bayes factor (Appendix A.1). The fractional prior contains minimal information and is centered at the null value.

In the case of a one-sample t test, **BFpack** executes an exhaustive exploratory test of whether the normal mean is equal to a pre-specified null value μ_0, whether it is smaller than μ_0, or whether it is larger than μ_0, i.e.,

$$H_1 : \mu = \mu_0 \text{ versus } H_2 : \mu < \mu_0 \text{ versus } H_3 : \mu > \mu_0.$$

For a confirmatory test, hypotheses can be formulated on the population mean which has label `mu` (Table 1). For the two-sample case, an exhaustive exploratory test is executed of whether the mean difference, i.e., $\theta = \mu_1 - \mu_2$, is zero, negative, or positive, i.e.,

$$H_1 : \theta = 0 \text{ versus } H_2 : \theta < 0 \text{ versus } H_3 : \theta > 0.$$

For a confirmatory test, hypotheses can be formulated on the mean difference which has label `difference` (Table 1). Below we illustrate the testing procedure for an independent two-sample t test.

Confirmatory hypothesis test in medical research

We consider a confirmatory two-sample t test for an application discussed in Venables and Ripley (2002). The data set `birthwt` contains information of 189 infants at a US hospital, and is available from the R package **MASS** (Ripley 2021). A number of variables are presented in the data set, with the main interest being the birth weight. In this example, we investigate whether the smoking status of the mother during pregnancy (`0` denotes nonsmoking and `1` denotes smoking) affects the average birth weights of the infants (in grams). There are $n_1 = 115$ infants whose mothers did not smoke during pregnancy and $n_2 = 74$ infants whose mothers smoked during pregnancy. Researchers expect that infants from the nonsmoking group have a larger average birth weight than the infants from the smoking group. This can be tested using a right one-sided t test on the mean difference θ:

$$H_1 : \theta = 0 \text{ versus } H_2 : \theta > 0,$$

where $\theta = \mu_1 - \mu_2$ is the mean difference, and μ_1 and μ_2 are the average birth weights in the nonsmoking group and the smoking group, respectively. Note that the hypotheses can

equivalently be written as $H_1 : \mu_1 = \mu_2$ and $H_2 : \mu_1 > \mu_2$. Hypothesis H_1 implies that the smoking status of mothers during pregnancy does not affect the average weight of their infants while H_2 implies that smoking of pregnant women has a negative effect on the average weight of their infants. As there is no medical reason that smoking can have a positive effect on the infants weights a third hypothesis $H_3 : \theta < 0$ is excluded from the test. Under the current model we assume that the within group variances are equal across the two groups.

Analyses using ***BFpack***

To perform the Bayesian t test using **BFpack**, first a classical t test is executed. Next the output of this analysis is plugged into the `BF()` function where hypotheses are formulated on the mean difference, which has label "`difference`":

```
R> library("MASS")
R> library("bain")
R> library("BFpack")
R> smoke0 <- subset(MASS::birthwt, smoke==0, select=bwt)
R> smoke1 <- subset(MASS::birthwt, smoke==1, select=bwt)
R> ttest1 <- bain::t_test(smoke0, smoke1, alternative = "greater",
     var.equal = TRUE)
R> print(ttest1)
R> constraints1 <- "difference = 0; difference > 0"
R> BF1 <- BF(ttest1, hypothesis = constraints1, complement = FALSE)
R> summary(BF1)
```

On the 1st, 2nd, and 3rd lines the R packages **MASS** (containing the data), **bain** (containing the classical t test function `t_test()` which is equivalent to the `t.test()` function but which also contains sample variances and sample sizes in the output), and **BFpack** (containing the Bayesian tests we present here) are loaded. In the 4th and 5th lines, two objects `smoke0` and `smoke1` are created which contain the birth weights of the smoking and nonsmoking group, respectively. On the 6th line the right one-sided two-sample t test is executed, and the output is printed in the 7th line. Next the two constrained hypotheses for the confirmatory test are specified on the parameter `difference` (which is the label of the mean difference between the groups in a two samples t test) (hypotheses are separated by a semi-colon "`;`"), and stored in the object `constraints1` on the 8th line. Subsequently, the Bayesian hypothesis tests are executed using the `BF()` function on the 9th line. The argument `complement = FALSE` is used so that the complement hypothesis, which in this case would assume that the difference in means is smaller than 0, is omitted in the confirmatory test. Furthermore equal prior probabilities of $\frac{1}{2}$ are set for the two hypotheses H_1 and H_2 in the confirmatory test (the default setting)[1]. Finally the output summary of the Bayesian tests are printed on the 10th line.

First we discuss the output of the classical test using the traditional significance level of 0.05 (controlling the unconditional type I error probability at 0.05):

[1]The same hypothesis test could be executed when using the `prior` argument instead of the `complement` argument where the complement hypothesis would have zero prior probability and the other two hypotheses would have prior probability of .5: `BF(ttest1, hypothesis = constraints1, prior.hyp = c(0.5, 0.5, 0), complement = TRUE)`.

```
        Two Sample t-test

data:  smoke0 and smoke1
t = 2.6529, df = 187, p-value = 0.004333
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 106.9528      Inf
sample estimates:
mean of x mean of y
 3055.696  2771.919
```

As the p value is smaller than the significance level of 0.05, there is enough evidence in the data to reject the null hypothesis in favor of the one-sided alternative which assumes that the average birth weight is larger for the nonsmoking group. Even though a small p value is obtained of 0.004333, the (unconditional) probability of incorrectly rejecting the null hypothesis is equal to the prespecified significance level of 0.05 as p values cannot be interpreted as error probabilities (Hubbard and Bayarri 2003).

Next we discuss the results of the exhaustive exploratory test (the first part of the summary output) which are given by

```
Bayesian hypothesis test
Type: exploratory
Object: t_test
Parameter: difference in means
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
           Pr(=0) Pr(<0) Pr(>0)
difference  0.204  0.003  0.793
```

In the exhaustive test the hypothesis assuming a positive difference (i.e., a larger average birth weight for the nonsmoking group) receives the largest posterior probability. This is in accordance with the positive t value (`t = 2.6529`, as can be seen from the classical test). The Bayes factor of a positive effect against no effect can be computed via

```
R> BF1$BFtu_exploratory[3]/BF1$BFtu_exploratory[1]
```

which yields $B_{31} = 3.896071$. Hence there is some positive evidence in the data in favor of a positive difference against no difference. Note that a negative difference receives hardly any posterior support, which confirms our prior intuition that this hypothesis cannot be explained by any medical arguments.

Next we discuss the results of the confirmatory test of H_1, which assumes that there is no group difference, versus H_2, which assumes a positive group difference (a larger mean for the nonsmoking group). The results of confirmatory tests are more extensive than the standard exploratory tests:

```
Bayesian hypothesis test
Type: confirmatory
```

```
Object: t_test
Parameter: difference in means
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
   Pr(hypothesis|data)
H1               0.204
H2               0.796

Evidence matrix (Bayes factors):
      H1    H2
H1 1.000 0.257
H2 3.896 1.000

Specification table:
   complex= complex> fit=  fit>   BF=   BF>    BF   PHP
H1        0      1.0    0 1.000 0.511 1.000 0.511 0.204
H2        1      0.5    1 0.996 1.000 1.991 1.991 0.796

Hypotheses:
H1: difference=0
H2: difference>0
```

Similar as in the exhaustive test there is most evidence for a positive difference (H_2) against no difference (H_1) with a Bayes factor of $B_{21} = 3.896$. Equivalently, the evidence for H_1 against H_2 equals $B_{12} = 1/B_{21} = 0.257$, which follows from the symmetry property of the Bayes factor. Because the prior probabilities for the hypotheses are equal, the Bayes factor is equal to the posterior odds: $\frac{\mathsf{P}(H_2|\mathbf{y})}{\mathsf{P}(H_1|\mathbf{y})} = \frac{0.796}{0.204} = 3.896 = B_{21}$. Notice that the posterior probability of the null hypothesis of 0.204 is considerably larger than the p value of 0.004333. This illustrates that classical p values tend to overestimate the evidence against a null hypothesis (Sellke *et al.* 2001). For this reason, there has been a recent call for using a smaller significance level in applied scientific research than the traditional cut-off value of 0.05 (Benjamin *et al.* 2018).

Statistical elaborations: Discussion of the ***`Specification table`***

The `Specification table` presents the different quantities of the extended Savage-Dickey density ratio in Equation 5. To understand how these values are computed, we plotted the unconstrained posterior for the `difference` parameter θ, which follows a Student t distribution with location 283.8, scale 107.0, and 187 degrees of freedom, and the unconstrained fractional prior, which follows a Student t distribution with location 0, scale 1407.7, and 1 degree of freedom, i.e., a Cauchy distribution, in Figure 1[2]. The unconstrained marginal posterior is obtained by updating the standard independence Jeffreys prior with the information in the observed data ($n_1 = 115$ and $n_2 = 74$). Note that this posterior results in identical Bayesian credible intervals as classical confidence intervals. The unconstrained fractional prior is obtained by updating the independence Jeffreys prior with a minimal fraction of the data which contains the information of three observations (as the three unknown parame-

[2]Technical details to derive these distributions can be found by following the steps in Appendix A.1.

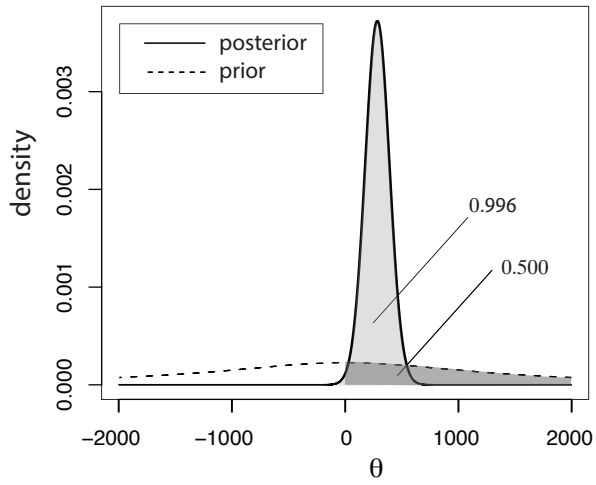

Figure 1: Unconstrained t distributed posterior for the difference in means θ based on the `birthwt` data (solid line), and the unconstrained fractional Cauchy prior (dashed line). Note that $P_u(\theta > 0) = 0.5$, $P_u(\theta > 0 \mid \mathbf{y}) = 0.996$, $\pi_u(\theta = 0) = 2.261\text{e}{-4}$, and $\pi_u(\theta = 0 \mid \mathbf{y}) = 1.156\text{e}{-4}$.

ters: group mean 1 μ_1, group mean 2 μ_2, and the common within group variance σ^2). The prior is adjusted so that the resulting Cauchy prior is centered around the null value $\theta = 0$. Under this adjusted fractional prior, (i) positive and negative effects are equally likely, and (ii) small effects are more likely than large effects. These are important properties for the prior in default Bayesian hypothesis testing (Jeffreys 1961). Because the prior is centered at the null value, the prior probability that the order (or one-sided) constraints hold under the unconstrained full model reflects the relative complexity (or relative size) of the constrained subspace under a hypothesis (see also Mulder *et al.* 2010).

The relative complexity of the one-sided hypothesis H_2 can be found in the column `complex>` which is equal to the prior probability that the one-sided constraint $\theta > 0$ holds under the unconstrained adjusted fractional prior, which is equal to 0.5 (as the prior is centered at the test value of 0; Figure 1, dashed line). This value quantifies the relative size of the constrained parameter space as the one-sided hypothesis covers half of the parameter space. Furthermore the relative fit of the one-sided hypothesis, in the column `fit>`, quantifies the relative fit of the one-sided constraint, which is calculated as the posterior probability that the constraint holds, which is equal to 0.996 (Figure 1, solid line). Consequently, following Equation 5, the Bayes factor of the one-sided hypothesis H_2 against the unconstrained full model equals $B^o_{2u} = \frac{.996}{.500} = 1.991$, as can be seen in the column labeled `BF>`. Note that $B^e_{2u} = 1$ as H_2 does not contain equality constraints.

Furthermore, the relative complexity and the relative fit of the equality hypothesis, in the column `complex=` and `fit=`, respectively, are equal to the probability densities at $\theta = 0$ for the unconstrained fractional prior and the unconstrained posterior, which are rounded to 0 with three decimals in the `Specification Table`. The unrounded density values are $1.156\text{e}{-4}$ and $2.261\text{e}{-4}$ for the posterior and prior, respectively. Consequently, following Equation 5, the Bayes factor of the equality hypothesis H_1 against the unconstrained full model equals $B^e_{1u} = \frac{1.156\text{e}{-4}}{2.261\text{e}{-4}} = 0.511$, as can be seen in the column labeled `BF=`. This Bayes factor of 0.511 for the equality hypothesis against the unconstrained hypothesis can also be seen from

Figure 1 as the posterior density at the null value is about half of the prior density at the null value. Note that $B^o_{1u} = 1$ as H_1 does not contain order constraints. We hope that these elaborations shed more insights about the nature of the procedure using Equation 5.

4.2. Analysis of variance

Statistical model and exploratory hypothesis test

(Multivariate) analysis of (co)variance ((M)AN(C)OVA) is performed when the interest is in testing group means using dummy group variables under a (multivariate) normal linear model, possibly by correcting for certain covariates. To perform statistical hypothesis tests under this class of models, first a model needs to be fit using the `aov()`, `maov()`, or `manova()` function. Subsequently, generalized adjusted fractional Bayes factors (Section A.1) are computed on the fitted object using the `BF()` function from **BFpack**.

If the factors are modeled using dummy (0, 1) variables[3], the exploratory analysis tests whether the effects of the dummy variables belonging to a certain factor equal to zero or not, and whether the effects of the dummy variables belonging to each interaction effect equal zero or. For example, if we consider a 2-way ANOVA model with two factors having two levels and an interaction effect,

$$y_i = \mu + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_{12} x_{i1} x_{i2} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2),$$

where x_{i1} and x_{i2} are dummy (0, 1) variables for the first and second factor, respectively, so that δ_1, δ_2, and δ_{12} capture the contribution of the first factor, the contribution of the second factor, and the contribution of the interaction effect of the two factors, respectively, and σ^2 is the common within groups variance, a nuisance parameter. The exploratory analysis then executes the following hypothesis tests:

$$\begin{aligned} \text{effect 1} &: H_1 : \delta_1 = 0 \text{ against } H_2 : \delta_1 \neq 0 \\ \text{effect 2} &: H_1 : \delta_2 = 0 \text{ against } H_2 : \delta_2 \neq 0 \\ \text{interaction effect} &: H_1 : \delta_{12} = 0 \text{ against } H_2 : \delta_{12} \neq 0. \end{aligned}$$

In the output H_1 and H_2 will be labeled as "`no effect`" and "`complement`", respectively. Note that in the case of factors with more levels (e.g., more than two groups), it would be tested whether all dummy effects belonging to a factor would be zero against an alternative full model. In the case of a multivariate model (e.g., MANOVA), the null hypothesis would assume that the dummy effects are zero across all dimensions of the dependent variable.

Confirmatory hypothesis test on numerical judgment

In experiment "4a" on numerical judgments of participants reported by Janiszewski and Uy (2008), the outcome variable was the amount by which a given anchor price for a television differed from the price estimated by a participant (expressed by means of a z score), and the two factors where (1) whether the anchor price was rounded, e.g., \$5000, or precise, e.g., \$4989 (`anchor` = `rounded` or `precise`, respectively); and (2) whether the participants

[3]The factors are modeled using dummy (0, 1) variables when `contrasts` are all set to `"contr.treatment"` in the `aov` object.

received a suggestion that the estimated price is close to the anchor value or whether they did not receive this suggestion (`motivation` = `low` or `high`, respectively). An example of a question, with `anchor` = `rounded` and `motivation` = `low`, was: "The retail price of a TV is $5000 (rounded). The actual price is only slightly lower than the retail price. Can you guess the price?". Alternatively, by changing "$5000" to "$4989" in the question a `precise` anchor price is obtained. By changing "slightly lower" to "lower" a question with a `high` motivation is obtained.

Given the above parameterization, let the reference group be `anchor = precise` and `motivation = high`, so that the combinations of the dummy variables correspond to the group means as follows:

dummy variables	`anchor`	`motivation`
$x_{i1} = 0,\ x_{i2} = 0$	`precise`	`high`
$x_{i1} = 1,\ x_{i2} = 0$	`rounded`	`high`
$x_{i1} = 0,\ x_{i2} = 1$	`precise`	`low`
$x_{i1} = 1,\ x_{i2} = 1$	`rounded`	`low`

It can be argued that participants who receive a `high` motivation are likely to give a lower estimate of the tv price, and that participants who see a `rounded` price may also give a lower price estimate. Moreover the estimated price may be even lower in the combined `high` and `rounded` condition. As the outcome of computed based on the anchor price minus the estimated price, we can translate this to the following confirmatory test where H_1 corresponds to the anticipated directional effect, H_2 is the traditional null hypothesis of no effect, and H_3 is the complement hypothesis:

$$\begin{aligned} H_1 &: \delta_1 > 0,\ \delta_2 < 0,\ \delta_{12} < 0 \\ H_2 &: \delta_1 = 0,\ \delta_2 = 0,\ \delta_{12} = 0 \\ H_3 &: \text{neither } H_1, \text{ nor } H_2. \end{aligned}$$

Analyses using ***BFpack***

First the 2×2 ANOVA model is fit using the `aov()` function. The fitted model is then plugged into the `BF()` function:

```
R> aov2 <- aov(price ~ anchor * motivation, data = tvprices)
R> summary(aov2)
R> constraints2 <- "anchorrounded > 0 & motivationlow < 0 &
+    anchorrounded:motivationlow < 0; anchorrounded = 0 &
+    motivationlow = 0 & anchorrounded:motivationlow = 0"
R> set.seed(1234)
R> BF2 <- BF(aov2, hypothesis = constraints2)
R> summary(BF2)
```

The output of the classical analysis (when running `summary(aov2)` in the second line) yields:

```
                Df Sum Sq Mean Sq F value   Pr(>F)
anchor           1  8.575   8.575  66.877 4.48e-11 ***
```

```
motivation          1 13.850  13.850 108.015 1.38e-14 ***
anchor:motivation   1  0.885   0.885   6.899   0.0112 *
Residuals          55  7.052   0.128
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a traditional significance level of 0.05, the p values suggest that there is enough evidence in the data to reject the null hypotheses of no effect of `anchor`, no effect of `motivation`, and no interaction effect of `anchor:motivation`. The first part of the output (when running `summary(BF2)`) gives the output of the exploratory hypothesis tests in a Bayesian framework:

```
Bayesian hypothesis test
Type: Exploratory
Object: aov
Parameter: group means
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
                  Pr(no effect) Pr(complement)
anchor                    0.000          1.000
motivation                0.000          1.000
anchor:motivation         0.251          0.749
```

Here `Pr(no effect)` denotes the posterior probabilities of the ("null") hypotheses of no effect of the dummy variables belonging to the effect of `anchor`, the effect of `motivation`, and the interaction effect `anchor:motivation`, denoted by H_1 above. Furthermore `Pr(complement)` denotes the posterior probabilities of the alternative (full) models, denoted by H_2 above. The results show clear support that an effect is present for the dummy effects of the `anchor` factor and the `motivation` factor (with posterior probabilities of approximately 1). Furthermore, there is some support that an interaction effect between the two factors is present (with a posterior probability of 0.749, equivalent to a Bayes factor of $\frac{0.749}{0.251} \approx 3$). As the evidence is very mild however, more data would need to be collected in order to draw a more decisive conclusion regarding the existence of an interaction effect. From a Bayesian point of view, this illustrates that classical p values tend to overestimate the evidence against a precise null hypothesis. See also Sellke *et al.* (2001) for an interesting discussion on this property. Due to this overestimation there has been a recent call to use smaller significance levels in classical significance tests (Benjamin *et al.* 2018).

In the second part of the output, the results of the confirmatory tests are given which yield:

```
Bayesian hypothesis test
Type: confirmatory
Object: aov
Parameter: group means
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
   Pr(hypothesis|data)
```

```
H1               0.999
H2               0.000
H3               0.001

Evidence matrix (Bayes factors):
      H1           H2       H3
H1 1.000 1.143559e+17 1963.526
H2 0.000 1.000000e+00    0.000
H3 0.001 5.824008e+13    1.000

Specification table:
   complex= complex> fit=  fit> BF=    BF>     BF   PHP
H1    1.000    0.083    1 0.994   1 11.933 11.933 0.999
H2    0.156    1.000    0 1.000   0  1.000  0.000 0.000
H3    1.000    0.917    1 0.006   1  0.006  0.006 0.001

Hypotheses:
H1: anchorrounded>0&motivationlow<0&anchorrounded___X___motivationlow<0
H2: anchorrounded=0&motivationlow=0&anchorrounded___X___motivationlow=0
H3: complement
```

The posterior probabilities (which are based on equal prior probabilities of $\frac{1}{3}$ for each of the three hypotheses) show that the joint one-sided hypothesis H_1 which assumed directional effects of `anchor`, `motivation`, and `anchor:motivation` is clearly most plausible after observing the data. The Bayes factors that are shown in the `Evidence matrix` show an equivalent picture, i.e., $B_{12} = 1.1436e17$ and $B_{13} = 1963.5$, which implies decisive evidence for H_1 against both H_2 and H_3. The different quantities in the extended Savage-Dickey density ratio in Equation 5 can be found in the `Specification table`. For hypotheses with only order (equality) constraints the measures for `complex=` and `fit=` (`complex>` and `fit>`) are set to 1. As can be seen for instance the posterior and prior probability that the one-sided constraints of H_1 hold under the unconstrained full model are equal to 0.994 and 0.083, respectively. This implies a good fit of the order constrained of H_1 in combination with a relatively small subspace that is covered by the constraints. As Bayes factors function as an Occam's razor this suggests a large support for H_1 given the data. Similarly as the complement hypothesis H_3 covers the remaining subspace, the posterior and prior probability are equal to 1 minus the posterior and prior probabilities for H_1. Finally observe a poor relative fit of the equality constraints of H_2, which is approximately 0, as can be seen from the column labeled `fit=`.

Statistical elaborations: Analysis via the `lm()` *function*

Given the relationship between an ANOVA model and a linear regression model, it is also possible to perform the first step using the `lm()` function instead of the `aov()` function. In that case, the output object is of class '`lm`' instead of class '`aov`'. The above exploratory tests are then replaced by exhaustive tests of whether the separate coefficients in the model are zero, negative, or positive (which is done by default for `lm` objects). This can be done as follows (for illustrative purposes we omit the `hypothesis` argument so that only the exploratory tests are executed):

```
R> lm2 <- lm(price ~ anchor * motivation, data = tvprices)
R> summary(lm2)
R> BF2 <- BF(lm2)
R> print(BF2)
```

The classical analysis then gives the following significance results of the separate coefficients:

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.04000    0.09246  -0.433   0.6670
anchorrounded                1.02000    0.13075   7.801 1.84e-10 ***
motivationlow               -0.72000    0.13307  -5.411 1.41e-06 ***
anchorrounded:motivationlow -0.49000    0.18656  -2.627   0.0112 *
```

and the exploratory Bayesian analysis gives the posterior probabilities of the coefficients:

```
Bayesian hypothesis test
Type: exploratory
Object: lm
Parameter: regression coefficients
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
                            Pr(=0) Pr(<0) Pr(>0)
(Intercept)                  0.808  0.128  0.064
anchorrounded                0.000  0.000  1.000
motivationlow                0.000  1.000  0.000
anchorrounded:motivationlow  0.144  0.851  0.005
```

Because both factors only have two levels each, the results of the exploratory Bayesian analyses are very similar with the earlier exploratory analyses based using the `aov` object. An important difference is that when using the `lm()` function, we also test one-sided hypotheses on the separate coefficients, while the exploratory test on an object of class '`aov`' has a two-sided nature where *all* effects of the dummy variables belonging to a certain factor are either equal to zero or not. Another important difference is that in the exhaustive exploratory tests of a `lm` object, the prior probability of the "zero effect" hypotheses equals $\frac{1}{3}$ (because every test considers three hypotheses which are assumed to be equally likely) while the prior probabilities of the "zero effect" hypotheses in the exploratory test on an `aov` object equal $\frac{1}{2}$ (because every test considers two hypotheses which are assumed to be equally likely). Depending on the situation, users might prefer one analysis over the other. Of course users can utilize the `hypothesis` argument and the `prior.hyp` argument to perform more specific confirmatory hypothesis tests.

4.3. Testing group variances

Statistical model and exploratory hypothesis test

Besides (or in addition to) testing group means, there are also situations where the interest is in testing the heterogeneity across populations. Equality and order constraints can be tested

between group variances σ^2 under normally distributed groups, i.e., $N(\mu_q, \sigma_q^2)$, under group q, for $q = 1, \ldots, Q$, where the group means are treated as nuisance parameters, using the generalized adjusted fractional Bayes factor. First the `bartlett_test()` function is used to fit the model which is equivalent to the `bartlett.test()` function with the addition that the the sample variances are also contained in the fitted object. Next the generalized adjusted fractional Bayes factors (Section A.2) are computed when running `BF()` on the object, which is based on the methodology of (Böing-Messing, Van Assen, Hofman, Hoijtink, and Mulder 2017b).

For the exploratory test it is tested whether the group variances are homogeneous or not, i.e.,

$$H_1 : \sigma_1^2 = \ldots = \sigma_Q^2 \text{ versus } H_2 : \text{ not } H_1.$$

For a confirmatory test equality/order constrained hypotheses on the group variances can be specified using the `hypothesis` argument.

Confirmatory hypothesis test in neuropsychology

Silverstein, Como, Palumbo, West, and Osborn (1995) conducted a psychological study to compare the attentional performances of 17 Tourette's syndrome (TS) patients, 17 ADHD patients, and 17 control subjects who did not suffer from TS or ADHD. The participants were shown a total of 120 sequences of either 3 or 12 letters. Each sequence contained either the letter T or the letter F at a random position. Each sequence was presented for 55 milliseconds and afterwards the participants had to indicate as quickly as possible whether the shown sequence contained a T or an F. After a participant completed all 120 sequences, his or her accuracy was calculated as the percentage of correct answers. In this section, we are interested in comparing the variances of the accuracies in the three groups. Research has shown that ADHD patients tend to be more variable in their attentional performances than subjects who do not suffer from ADHD (e.g., Kofler *et al.* 2013; Russell *et al.* 2006). It is less well documented whether TS patients are less or more variable in their attentional performances than healthy control subjects. We will therefore test the following set of hypotheses to investigate whether TS patients are as variable in their attentional performances as either ADHD patients or healthy controls (C):

$$\begin{aligned} H_1 &: \sigma_C^2 = \sigma_{TS}^2 < \sigma_{ADHD}^2 \\ H_2 &: \sigma_C^2 < \sigma_{TS}^2 = \sigma_{ADHD}^2 \\ H_3 &: \sigma_C^2 = \sigma_{TS}^2 = \sigma_{ADHD}^2 \\ H_4 &: \text{ not } H_1, H_2, H_3. \end{aligned}$$

The complement is included to safeguard against the data supporting neither of the first three constrained hypotheses.

Analyses using ***BFpack***

Silverstein *et al.* (1995) reported the following sample variances of the accuracies in the three groups: $s_C^2 = 15.52$, $s_{TS}^2 = 20.07$, and $s_{ADHD}^2 = 38.81$. The object `attention` from **BFpack** contains hypothetically generated data having these descriptive statistics. First the data are analyzed using the `bartlett_test()` function, and next, the multiple hypothesis test is executed by plugging the fitted object in the `BF()` function together with the constrained hypotheses on the variances:

```
R> bartlett3 <- bartlett_test(x = attention$accuracy, g = attention$group)
R> print(bartlett)
R> get_estimates(bartlett3)
R> constraints3 <- "Controls = TS < ADHD; Controls < TS = ADHD;
+    Controls = TS = ADHD"
R> set.seed(358)
R> BF3 <- BF(bartlett3, hypothesis = constraints3)
R> summary(BF3)
```

The third line was added for users to see the labels of the variances, which yields `ADHD`, `Controls`, and `TS`, on which the constrained hypotheses can be formulated. We use equal prior probabilities for the hypotheses by omitting the argument `prior.hyp` in the call of the `BF()` function.

The output of the classical test (when calling `print(bartlett)`) looks as follows:

```
        Bartlett test of homogeneity of variances

data:  attention$accuracy and attention$group
Bartlett's K-squared = 3.6187, df = 2, p-value = 0.1638
```

The p value is equal to 0.1638. Thus the hypothesis of homogeneity of variances cannot be rejected using a significance level of 0.05. This result however does not imply that there is evidence in the data for the null hypothesis of homogeneity of variances because p values cannot be interpreted as measures of evidence in favor of a null. The reason is that p values are uniformly distributed in the interval $(0, 1)$ when the null is true.

The results of the Bayesian exploratory test of homogeneity of variances (first part when calling `summary(BF3)`) yield:

```
Bayesian hypothesis test
Type: Exploratory
Object: bartlett_htest
Parameter: group variances
Method: generalized adjusted fractional Bayes factor

   homogeneity of variances no homogeneity of variances
                      0.803                       0.197
```

Hence the posterior probability that the group variances are equal to each other is equal to 0.803 given the observed data. The Bayes factor would then be equal to $\frac{0.803}{0.197} \approx 4$, which (equivalently) suggests some positive evidence for homogeneity of variances.

The output for the confirmatory hypothesis test looks as follows:

```
Bayesian hypothesis test
Type: Confirmatory
Object: bartlett_htest
Parameter: group variances
Method: generalized adjusted fractional Bayes factor
```

```
Posterior probabilities:
   Pr(hypothesis|data)
H1               0.426
H2               0.278
H3               0.238
H4               0.058

Evidence matrix (Bayes factors):
      H1    H2    H3    H4
H1 1.000 1.530 1.791 7.315
H2 0.654 1.000 1.171 4.781
H3 0.558 0.854 1.000 4.083
H4 0.137 0.209 0.245 1.000

Specification table:
   complex= complex> fit=  fit>   BF=   BF>    BF   PHP
H1       NA    0.579   NA 0.970 4.363 1.677 7.315 0.426
H2       NA    0.423   NA 0.913 2.215 2.158 4.781 0.278
H3       NA    1.000   NA 1.000 4.083 1.000 4.083 0.238
H4       NA    1.000   NA 1.000 1.000 1.000 1.000 0.058

Hypotheses:
H1: Controls=TS<ADHD
H2: Controls<TS=ADHD
H3: Controls=TS=ADHD
H4: complement
```

Hypothesis H_1 receives largest posterior probability given the observed data, but H_2 and H_3 are viable competitors. It appears that even the complement H_3 cannot be ruled out entirely given a posterior probability of 0.058. As equal prior probabilities are used for the hypotheses, the same conclusions can be drawn based on the Bayes factors in the `Evidence matrix`. Thus, the results indicate that TS are as heterogeneous in their attentional performances as healthy control in this specific task, but further research would be required to obtain more conclusive evidence.

Finally note that Bayes factors on hypotheses with equality constraints on variances cannot be formulated as a Savage-Dickey density ratio (Appendix A.2). Therefore the numerator and the denominator of B^e_{tu} in Equation 5 are not available, the columns of `complex=` and `fit=` are left empty in the `Specification table`.

4.4. Logistic regression

Statistical model and exploratory hypothesis test

The generalized linear model is a flexible generalization of the normal linear regression model where the outcome variable has an error that follows a non-normal distribution (McCullagh and Nelder 1989). The logistic regression model is one of the most commonly used generalized linear models where the outcome variable is a binary variable. A logit function of the "success"

probability of the outcome variable is assumed to follow a linear function of the predictor variables, $\beta_0 + \beta_1 x_1 + \ldots + \beta_Q x_q$, where β_q denotes the effect of the q predictor variable, x_q, on the outcome variable, and β_0 denotes the intercept.

First the `glm()` function is used to fit the logistic regression model for a given data set. By plugging in the resulting `glm` object into `BF()`, adjusted fractional Bayes factors are computed based on Gaussian approximations and minimal fractions (Gu *et al.* 2017). For technical details we refer the interested reader to Section A.3.

By default exploratory exhaustive tests are executed of whether each effect is zero, negative, or positive, assuming equal prior probabilities, i.e.,

$$H_1 : \beta_q = 0 \text{ versus } H_2 : \beta_q < 0 \text{ versus } H_3 : \beta_q > 0,$$

for $q = 1, \ldots, Q$. Competing equality and order constraints on the β's can be tested in a confirmatory test using the `hypothesis` argument.

Confirmatory hypothesis test in forensic psychology

The presence of systematic biases in the legal system runs counter to society's expectation of fairness. Moreover such biases can have profound personal ramifications, and the topic therefore warrants close scrutiny. Wilson and Rule (2015) examined the correlation between perceived facial trustworthiness and criminal-sentencing outcomes (data available at `https://osf.io/7mazn/`, **BFpack** contains a simulated version of these data having the same descriptives). In Study 1 photos of inmates who had been sentenced to death (or not) were rated by different groups of participants on trustworthiness, "Afrocentricity" (how stereotypical "black" participants were perceived as), attractiveness and facial maturity. Each photo was also coded for the presence of glasses/tattoos and facial width-to-height ratio. A logistic regression with sentencing as outcome was fitted to the predictors.

Previous research had shown that the facial width-to-height ratio (fWHR) has a positive effect on perceived aggression and thus may also have a positive effect on sentencing outcomes. In addition, perceived Afrocentricity had been shown to be associated with harsher sentences (Wilson and Rule 2015). In the first hypothesis it was expected that all three predictors have a positive effect on the probability of being sentenced to death. Additionally, we might expect lack of perceived trustworthiness to have the largest effect. In the second hypothesis it was assumed that only trustworthiness has a positive effect. Finally, the complement hypothesis was considered. The hypotheses can then be summarized as follows

$$\begin{aligned} H_1 &: \beta_{trust} > (\beta_{fWHR}, \beta_{afro}) > 0 \\ H_2 &: \beta_{trust} > \beta_{fWHR} = \beta_{afro} = 0 \\ H_3 &: \text{neither } H_1, \text{ nor } H_2. \end{aligned}$$

*Analyses using **BFpack***

Before fitting the logistic regression we reverse-coded the trustworthiness scale and standardized it to be able to compare the magnitude the three effects. The data matrix has the name of `wilson` in **BFpack**. The exploratory tests and the confirmatory test are executed by plugging in the fitted `glm` object and the constrained hypotheses in the `BF()` function. First the full logistic regression model is fitted, and then the hypothesis tests are executed:

```
R> glm4 <- glm(sent ~ ztrust + zfWHR + zAfro + glasses + attract +
+    maturity + tattoos, family = binomial(), data = wilson)
R> summary(glm4)
R> set.seed(123)
R> constraints4 <- "ztrust > (zfWHR, zAfro) > 0; ztrust > zfWHR = zAfro = 0"
R> BF4 <- BF(glm4, hypothesis = constraints4)
R> summary(BF4)
```

The results of the classical test (when running `summary(glm4)`) yields (where we only print the table of estimates):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.78886    0.60300   1.308   0.1908
ztrust       0.38594    0.08496   4.542 5.56e-06 ***
zfWHR        0.37008    0.08450   4.380 1.19e-05 ***
zAfro       -0.18071    0.07183  -2.516   0.0119 *
glasses      0.39064    0.21155   1.847   0.0648 .
attract     -0.03131    0.14514  -0.216   0.8292
maturity    -0.14631    0.08732  -1.676   0.0938 .
tattoos      1.31579    0.81487   1.615   0.1064
```

Next we compare these results with the results of the Bayesian exploratory tests, which yield (first part when running `summary(BF4)`:

```
Bayesian hypothesis test
Type: Exploratory
Object: glm
Parameter: general parameters
Method: adjusted fractional Bayes factors using Gaussian approximations

Posterior probabilities:
            Pr(=0) Pr(<0) Pr(>0)
(Intercept)  0.853  0.014  0.133
ztrust       0.000  0.000  1.000
zfWHR        0.001  0.000  0.999
zAfro        0.365  0.631  0.004
glasses      0.712  0.009  0.278
attract      0.930  0.041  0.029
maturity     0.770  0.219  0.011
tattoos      0.787  0.011  0.202
```

These results show which hypothesis (either no effect, a negative effect, or a positive effect) is most plausible for each coefficient given the observed data. When comparing the posterior probabilities with the two-tailed p values from the classical analyses, we see that smaller p values correspond to smaller posterior probabilities for the null hypothesis of "no effect" (column `Pr(=0)`). As was argued by Sellke *et al.* (2001), however, classical p values tend to overestimate the evidence against an equality constrained null hypothesis. For example

based on the two-tailed p of `zAfro`, which is equal to 0.0119, it would be concluded that there is enough evidence to reject the null when using a significance level of 0.05. The posterior probability of a zero effect for `zAfro` however is (still) 0.365, and thus, concluding that the effect is nonzero would imply that there is a conditional error probability of about 0.365 to draw the wrong conclusion given the observed data, which is quite large. This is one of the motivations for the recent call for using more conservative significance levels when interpreting p values (Benjamin *et al.* 2018).

Now we discuss the results of the confirmatory test, which is of main interest. The output is given by

```
Bayesian hypothesis test
Type: Confirmatory
Object: glm
Parameter: general parameters
Method: adjusted fractional Bayes factors using Gaussian approximations

Posterior probabilities:
   Pr(hypothesis|data)
H1               0.071
H2               0.002
H3               0.927

Evidence matrix (Bayes factors):
       H1      H2    H3
H1  1.000  33.066 0.076
H2  0.030   1.000 0.002
H3 13.133 434.246 1.000

Specification table:
   complex= complex> fit=  fit>   BF=   BF>    BF   PHP
H1    1.000    0.037    1 0.003 1.000 0.079 0.079 0.071
H2    0.106    0.500    0 1.000 0.001 2.000 0.002 0.002
H3    1.000    0.963    1 0.997 1.000 1.035 1.035 0.927

Hypotheses:
H1: ztrust>(zfWHR,zAfro)>0
H2: ztrust>zfWHR=zAfro=0
H3: complement
```

In the output we see that the complement receives most support. This can also be concluded based on the posterior probabilities. The evidence matrix shows that the complement hypothesis (H_3) receives about 13 times more support than the second best hypothesis (H_1). The fact that none of the two anticipated hypotheses were supported by the data indicates that the theories are not yet well-developed. Closer inspection of the beta-coefficients reveals that this is largely driven by the negative effect between perceived Afrocentricity and sentencing harshness ($\hat{\beta}_{zAfro} = -0.18071$). This unexpected result is discussed further by Wilson and Rule (2015) in their supplementary materials (doi:10.1177/0956797615590992).

Statistical elaborations: Bayesian exploratory tests via classical analysis output

The exploratory Bayesian tests can also be executed using the results of a classical significance test via the `BF.default` (which calls the approximate adjusted fractional Bayes factor) by plugging in the estimates, the error variances, and the sample size (Section 3):

```
R> ct <- lmtest::coeftest(glm4)
R> BF(ct[,1], Sigma = diag(ct[,2]^2), n = attr(ct, "nobs"))
```

This is possible as the exploratory tests of the separate coefficients only needs the separate error variances, and not the entire error covariance matrix (which, by default, is not contained in the `ct` object), because for every separate tests the remaining parameters are treated as nuisance parameters and are thus integrated out. Thus the parameters cannot be tested against each other based on the default output of the function `coeftest()`.

4.5. Multivariate linear regression

Statistical model and exploratory hypothesis test

Multivariate normal linear regression models are useful for better understanding how a set of K predictor variables affect P outcome variances when the error follows a multivariate normal error with unknown covariance matrix:

$$y_{ip} = \mu_p + \beta_{1p}x_{i1} + \ldots + \beta_{Kp}x_{iK} + \epsilon_{ip},$$

for the i-th observation of the p-th outcome variable, for $p = 1, \ldots, P$, where $(\epsilon_{i1}, \ldots, \epsilon_{iP})' \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is an unknown error covariance matrix, and where β_{kp} denotes the effect of the k-th predictor variable on the p-th outcome variable, for $i = 1, \ldots, n$. First `lm` can be used to fit the multivariate normal linear regression model for a given data set. Subsequently Bayesian hypothesis tests are executed using generalized adjusted fractional Bayes factors using **BFpack** (Section A.1).

By default exploratory exhaustive tests are executed of whether each effect is zero, negative, or positive, assuming equal prior probabilities, i.e.,

$$H_1 : \beta_{kp} = 0 \text{ versus } H_2 : \beta_{kp} < 0 \text{ versus } H_3 : \beta_{kp} > 0,$$

for $k = 1, \ldots, K$, and $p = 1, \ldots, P$. In **BFpack**, the names of the parameters have the following form. If a predictor variable has name `x1` and a dependent variable has name `y1`, then the effect of this predictor variable on this dependent variable is labeled as `x1_on_y1`. By running the `get_estimates()` function on the multivariate `lm` object ('`mlm`'), a vector is obtained containing the parameter names. These parameter names can be used for formulated constrained hypotheses using the `hypothesis` argument, as discussed next.

Confirmatory hypothesis test in fMRI studies

It is well established that the fusiform facial area (FFA), located in the inferior temporal cortex of the brain, plays an important role in the recognition of faces. This data comes from a study on the association between the thickness of specific cortical layers of the FFA and individual differences in the ability to recognize faces and vehicles (McGuigin, Newton,

Tamber-Rosenau, Tomarken, and Gauthier 2020). High-resolution fMRI was recorded from 13 adult participants, after which the thickness of the superficial, middle, and deep layers of the FFA was quantified for each individual. In addition, individual differences in face and vehicle recognition ability were assessed using a battery of tests.

In this example, two hypotheses are of main interest. In a recent study, McGuigin, Van Gulick, and Gauthier (2016) found that individual differences in the overall thickness of the FFA are negative correlated with the ability to recognize faces but positively correlated with the ability to recognize cars. To elaborate, consider a multivariate multiple regression model model with cortical thickness measures for the superficial, middle, and deep layers as three repeated measures for each participant, and facial recognition ability and vehicle recognition ability as two dependent variables.

Hypothesis H_1 is a main effects only model specifying that only main effect terms for face and vehicle are sufficient to predict the thickness of layers. The absence of layer $\times$ face or layer $\times$ vehicle interaction terms means that the relations between face and vehicle recognition are invariant across cortical layers. This implies that regression coefficients between face recognition and cortical thickness measures are expected to be negative, coefficients between vehicle recognition and cortical thickness measures are expected to be positive, and no layer-specific effect is expected for either faces or vehicles.

Hypothesis H_2 is based on prior findings concerning the early development of facial recognition abilities and the more rapid development of the deep layer of the FFA. Here it is assumed that the negative effect between facial recognition and the cortical thickness would be more pronounced in the deep layer, relative to the superficial and middle layers.

A multiple hypothesis test is executed on these two hypotheses and the complement hypothesis, which can be summarized as

$$\begin{aligned}
H_1 &: \beta_{Face_on_Deep} = \beta_{Face_on_Middle} = \beta_{Face_on_Superficial} < 0 < \beta_{Vehicle_on_Deep} \\
&\quad = \beta_{Vehicle_on_Middle} = \beta_{Vehicle_on_Superficial} \\
H_2 &: \beta_{Face_on_Deep} < \beta_{Face_on_Middle} = \beta_{Face_on_Superficial} < 0 < \beta_{Vehicle_on_Deep} \\
&\quad = \beta_{Vehicle_on_Middle} = \beta_{Vehicle_on_Superficial} \\
H_3 &: \text{neither } H_1, \text{ nor } H_2.
\end{aligned}$$

Analyses using ***BFpack***

First a multivariate model is fitted with dependent variables `Superficial`, `Middle`, and `Deep` and predictor variables `Face` and `Vehicle`. Subsequently the fitted model and the constrained hypotheses on the effects (e.g., where `Face_on_Deep` refers to the effect of the predictor variable `Face` on the dependent variable `Deep`) are plugged into the `BF()` function:

```
R> mlm5a <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle,
+    data = fmri)
R> summary(mlm5a)
R> constraints5a <- "Face_on_Deep = Face_on_Superficial = Face_on_Middle
+    < 0 < Vehicle_on_Deep = Vehicle_on_Superficial = Vehicle_on_Middle;
+    Face_on_Deep < Face_on_Superficial = Face_on_Middle < 0 <
+    Vehicle_on_Deep = Vehicle_on_Superficial = Vehicle_on_Middle"
R> set.seed(123)
```

```
R> BF5a <- BF(mlm5a, hypothesis = constraints5a)
R> summary(BF5a)
```

The classical analyses (when running `summary(mlm5a)`) results in the following (slightly shortened) output:

```
Response Superficial :

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03553    0.07998  12.948 1.42e-07 ***
Face        -0.09314    0.10148  -0.918    0.380
Vehicle      0.13365    0.10782   1.240    0.243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Middle :

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.12380    0.09819  11.445 4.55e-07 ***
Face        -0.05877    0.12460  -0.472    0.647
Vehicle      0.20051    0.13237   1.515    0.161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Deep :

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3668     0.1416   9.655 2.19e-06 ***
Face         -0.6064     0.1796  -3.376  0.00705 **
Vehicle       0.2102     0.1909   1.101  0.29658
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we show the results of the exploratory Bayesian analyses using **BFpack**:

```
Bayesian hypothesis test
Type: Exploratory
Object: mlm
Parameter: regression coefficients
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
                           Pr(=0) Pr(<0) Pr(>0)
(Intercept)_on_Superficial  0.000  0.000  1.000
Face_on_Superficial         0.544  0.357  0.099
```

```
Vehicle_on_Superficial      0.475  0.079  0.447
(Intercept)_on_Middle       0.000  0.000  1.000
Face_on_Middle              0.609  0.257  0.134
Vehicle_on_Middle           0.404  0.063  0.533
(Intercept)_on_Deep         0.000  0.000  1.000
Face_on_Deep                0.053  0.939  0.008
Vehicle_on_Deep             0.507  0.087  0.406
```

As can be seen the exploratory tests provide the posterior probabilities of each of the 9 coefficients to be zero, negative, or positive. Based on the posterior probabilities we see that there is quite some posterior uncertainty regarding the effects of the two predictor variables on the three dependent variables.

Next we discuss the results of the confirmatory hypothesis test, which is given by (to keep the presentation concise we only print the posterior probabilities and the Bayes factors here):

```
Bayesian hypothesis test
Type: Confirmatory
Object: mlm
Parameter: regression coefficients
Method: generalized adjusted fractional Bayes factors

Posterior probabilities:
   Pr(hypothesis|data)
H1               0.023
H2               0.975
H3               0.002

Evidence matrix (Bayes factors):
       H1    H2     H3
H1  1.000 0.024  13.35
H2 42.391 1.000 565.93
H3  0.075 0.002   1.00
```

The evidence matrix reveals that there is clear evidence for H_2 against H_1 ($B_{21} = 42.391$) and extreme evidence for H_2 against H_3 ($B_{23} = 565.93$). The same conclusion can be drawn when looking at the posterior probabilities for the hypotheses. Based on these result we would conclude that hypothesis H_2 receives most support from the data. Moreover when we would conclude that H_2 is true, there is a probability of about 0.025 of drawing the wrong conclusion given the observed data.

Statistical elaborations: Comparison with other approaches

One could attempt to test and compare these hypotheses using linear mixed effects models software (e.g., the `gls()` function in the **nlme** package in R, Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2021) with an appropriate covariance structure on the residuals to account for within-subject dependence. Alternatively one could use a model selection framework like that embodied in the **BayesFactor** package in R. While these approaches can separately test

some components of a hypothesis, they are not well suited to jointly test all components of a constrained hypothesis, such as H_1 which specifies that all coefficients involving faces are smaller than 0 and that all coefficients involving vehicles are larger than 0. **BFpack** on the other hand allows researchers to test such hypotheses in a direct manner.

Statistical elaborations: Analyses using ***BFpack*** *with missing observations*

Missing data are ubiquitous in statistical practice. Properly handling missing data in model selection and hypothesis testing problems has not received a lot of attention in the literature as the focus generally lies on estimation problems in the presence of missing observations Rubin (1996). However, properly handling missing observations is specifically challenging in model selection and hypothesis testing problems due to model uncertainty and we need to avoid possible biases towards a certain model or hypothesis. In a Bayesian framework the natural approach is to use the posterior predictive distribution to impute missing data given the observed data, and subsequently compute each marginal likelihood by averaging over the imputed data. This can be computationally expensive when considering many models having complex equality and order constraints on certain parameters.

For Bayes factors which can be expressed as an extended Savage-Dickey density ratio as in Equation 5, the computation can be considerably cheaper. Interestingly we only need to generate imputed data using the posterior predictive distribution under the unconstrained full model (possibly including auxiliary variables) because the four elements in Equation 5 are all defined under the unconstrained model. Subsequently we can obtain each element given the observed data by averaging over the obtained element from the complete imputed data sets. This is explained in Appendix B. This procedure can be used when the missing data are missing at random, similar as in estimation problems (Little and Rubin 2002). For more details about the approach we refer the interested reader to Hoijtink, Gu, Mulder, and Rosseel (2019b) and Mulder and Gu (2021). Below we illustrate (i) how to compute Bayes factors in the case there are observations that are missing at random, and (ii) to illustrate that multiple imputation using the posterior predictive distribution is preferred over list-wise deletion as the latter results in more loss of statistical evidence.

Here we illustrate how Bayes factors can be obtained in the case of random missing observations in the fMRI data set when using the Bayes factors given in Equation 5 using **BFpack**. For illustration purposes a slightly simpler constrained hypothesis test is considered to reduce computation time, which is given by

$$\begin{aligned} H_1 &: \beta_{Face\_on\_Deep} = \beta_{Face\_on\_Middle} = \beta_{Face\_on\_Superficial} < 0 \\ H_2 &: \beta_{Face\_on\_Deep} < \beta_{Face\_on\_Middle} = \beta_{Face\_on\_Superficial} < 0 \\ H_3 &: \text{not } H_1, \text{ or } H_2. \end{aligned}$$

These hypotheses are specified as follows:

```
R> constraints5b <-
+    "Face_on_Deep = Face_on_Superficial = Face_on_Middle < 0;
+    Face_on_Deep < Face_on_Superficial = Face_on_Middle < 0"
```

First the Bayes factors and posterior probabilities are obtained for this hypothesis test for the complete data set:

```
R> mlm5b <- lm(cbind(Superficial,Middle,Deep) ~ Face + Vehicle, data = fmri)
R> BF5b <- BF(mlm5b, hypothesis = constraints5b)
R> print(BF5b)
```

This results in posterior probabilities of 0.050, 0.927, and 0.023 for the two constrained hypotheses and the complement hypothesis, respectively. The Bayes factor of the most supported hypothesis (H_2) against the second most supported hypothesis (H_1) equals $B_{21} = 18.443$.

Next 10 missing observations (out of 65 separate observations in total) are randomly generated (missing at random):

```
R> fmri_missing <- fmri
R> set.seed(1234)
R> for(i in 1:10) {
+    fmri_missing[sample(1:nrow(fmri), 1), sample(1:ncol(fmri), 1)] <- NA
+  }
```

This results in 7 rows with at least one missing observation. Therefore listwise deletion would leave us with only 6 complete observations (of the 13 rows in total). Even though list-wise deletion is generally not recommended (Rubin 1987, 1996), for this illustration we compute the Bayes factors and posterior probabilities based on these 6 complete data observations to illustrate the loss of evidence as a result of list-wise deletion.

```
R> fmri_listdel <- fmri_missing[!is.na(apply(fmri_missing, 1, sum)),]
R> mlm5b_listdel <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle,
+    data = fmri_listdel)
R> BF5b_listdel <- BF(mlm5b_listdel, hypothesis = constraints5b)
R> print(BF5b_listdel)
```

This results in posterior probabilities of 0.010, 0.820, and 0.170 for the two constrained hypotheses and the complement hypothesis, respectively. As expected the posterior probability of the hypothesis H_2 which received most evidence in based on the complete data set, decreased from 0.927 to 0.820.

Now we illustrate that the loss of evidence is less using the posterior predictive distribution which also uses the information of the partly observed cases. We first generate 500 imputed data sets using `mice` from the **mice** package (Van Buuren and Groothuis-Oudshoorn 2021), and then use `BF()` to get the measures of relative fit and relative complexity for the equality and order constraints for the three hypotheses. These are be obtained from the element `BFtable_confirmatory` of an object of class '`BF`'[4] as follows:

```
R> M <- 500
R> library("mice")
R> mice_fmri <- mice :: mice(data = fmri_missing, m = M, meth = c("norm",
+    "norm", "norm", "norm", "norm"), diagnostics = FALSE, printFlag = FALSE)
```

[4]Note that the measures of relative fit and relative complexity can also be found in the `Specification table` when calling the `summary()` function on an object of class '`BF`' in the case of a confirmatory test on the hypotheses specified in the `hypothesis` argument of the `BF()` function.

```
R> relmeas_all <- matrix(unlist(lapply(1:M, function(m) {
+    mlm5b_m <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle,
+    data = mice::complete(mice_fmri, m))
+    BF5b_m <- BF(mlm5b_m, hypothesis = constraints5b)
+    c(BF5b_m$BFtable_confirmatory[, 1:4])
+  })), ncol = M)
R> relmeas <- matrix(apply(relmeas_all, 1, mean), nrow = 3)
R> row.names(relmeas) <- c("H1", "H2", "H3")
R> colnames(relmeas) <- c("comp_E", "comp_O", "fit_E", "fit_O")
R> BF_tu_confirmatory <- relmeas[,3] * relmeas[,4] / (relmeas[,1] *
+    relmeas[,2])
R> PHP <- BF_tu_confirmatory / sum(BF_tu_confirmatory)
R> print(PHP)
```

In the 11th line the averages are computed for the four different measures of relative fit and relative complexity for the constrained hypotheses in Equation 5, across all 500 imputed data sets. Appropriate names are given on lines 12 and 13 for illustrative purposes. Subsequently in the 14th line, the Bayes factors of all constrained hypotheses against an unconstrained alternative are computed using the Equation 5 based on the four different quantities (which also holds for the generalized adjusted fractional Bayes factor; Section A.1). This results in posterior probabilities of 0.048, 0.900, and 0.52 for the two constrained hypotheses and the complement hypothesis, respectively. Thus the posterior probability for the most supported hypothesis is still 0.900 (compared to 0.927 based on the complete data set), which is considerably larger than the posterior probability of 0.820 which was obtained using the data set after list-wise deletion.

4.6. Testing measures of association

Statistical model and exploratory hypothesis test

Measures of association play a central role in the applied sciences to quantify the degree of association between the variables of interest, possibly while correcting for certain control variables. The Pearson product-moment correlation coefficient is the most commonly used measure of association which expresses the strength of the linear relationship between two continuous variables. **BFpack** allows researchers to test complex hypotheses involving equality and order constraints on dependent overlapping correlations, on dependent non-overlapping correlations, and on independent correlations across independent groups, possibly, while correcting for certain covariates. The methodology builds on the work of Mulder (2016) and Mulder and Gelissen (2019).

It is assumed that the i-th observation of a P dimensional dependent variable follows a P-variate normal distribution, $\mathbf{y}_i \sim N(\mathbf{B}\mathbf{x}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})\mathbf{C}\text{diag}(\boldsymbol{\sigma})$, for observation i, where $\mathbf{x}_i$ denotes a vector of covariates of observation i (of which the first element is generally a 1), $\mathbf{B}$ contains the corresponding unknown coefficients of these covariates, $\boldsymbol{\sigma}$ is a vector of standard deviations, and $\mathbf{C}$ is the correlation matrix where the (p, q)-th element is the association between the p-th and q-th variable. Writing the covariance matrix as a function of a diagonal matrix of standard deviations, $\text{diag}(\boldsymbol{\sigma})$, and the correlation matrix, $\mathbf{C}$, is also referred to as the separation strategy (Barnard, McCulloch, and Meng 2000).

When the interest is testing correlations across J independent groups, the i-th observation in group j is distributed as $\mathbf{y}_{ij} \sim N(\mathbf{B}_j\mathbf{x}_{ij}, \text{diag}(\boldsymbol{\sigma}_j)\mathbf{C}_j\text{diag}(\boldsymbol{\sigma}_j))$, with group specific coefficients, $\mathbf{B}_j$, standard deviations, $\boldsymbol{\sigma}_j$, and correlation matrix, $\mathbf{C}_j$, in group $j = 1, \ldots, J$. In both cases, joint uniform priors are used for the correlation matrices and improper Jeffreys priors are used for the standard deviations and coefficients. Joint uniform priors for the correlations in a correlation matrix have certain attractive properties which are not necessarily shared by alternative proposals (Mulder and Gelissen 2019). Technical details on the implemented Bayes factors for testing measures of association can be found in Section A.4.

First an unconstrained Bayesian correlation analysis needs to be executed using the `cor_test()` function in **BFpack**. This function fits an unconstrained Bayesian model using joint uniform priors for the correlation matrices (Mulder and Gelissen 2019). The resulting object of class '`cor_test`' then needs to be plugged into the `BF()` function to perform Bayes factor tests of constrained hypotheses on the correlations. These Bayes factors are based on uniform priors for the free correlations under the constrained hypotheses. In the exploratory tests exhaustive tests are executed on the correlations, i.e.,

$$H_1 : \rho_{jpq} = 0 \text{ versus } H_2 : \rho_{jpq} < 0 \text{ versus } H_3 : \rho_{jpq} > 0,$$

for $j = 1, \ldots, J$, and $p < q$.

In **BFpack**, the names of the correlations have the following form. In the case of a single group, $J = 1$, and we consider two predictor variables with names, `y1` and `y2`, the correlation between these dependent variables has name `y1_with_y2`. In the case of J independent groups, with $J > 1$, the correlation between `y1` and `y2` in group 1 has name `y1_with_y2_in_g1`. Again, by running the `get_estimates()` function on the output of the `cor_test` object, a vector is printed containing the correlation names. These parameter names can be used for formulated constrained hypotheses using the `hypothesis` argument.

Confirmatory hypothesis test in neuropsychology

Schizophrenia is often conceptualized as a disorder of "dysconnectivity" characterized by disruption in neural circuits that connect different regions of the brain (e.g., Friston and Frith 1995). This data set (originally collected by Ichinose, Han, Polyn, Park, and Tomarken (2019)) can be used to test whether such dysconnection is manifested behaviorally as weaker correlations among measures that we would expect to be highly correlated among non-schizophrenic individuals. 20 patients suffering from schizophrenia (SZ group) and 20 healthy control (HC group) participants were administered six measures of working memory. Ichinose et al. hypothesized that each of the 15 correlations would be smaller in the schizophrenic group relative to the control group, i.e.,

$$\begin{aligned} H_1 &: \rho_{\text{SZ},pq} > \rho_{\text{HC},pq}, \text{ for all } 1 \leq p < q \leq 6 \\ H_2 &: \text{not } H_1. \end{aligned}$$

H_1 specifies that each correlation in the HC group is expected to be larger than the corresponding correlation in the SZ group (i.e., a total of 15 order constraints were imposed). The complement hypothesis H_2 represents any pattern of correlations not consistent with H_1.

As will be shown this application is an interesting case of how a joint order (or one-sided) Bayesian approach can provide a more powerful and more appropriate test relative to alternative methods.

Analyses using ***BFpack***

First the data `memory` is split for the HC group (group 1) and the SZ group (group 2), and an unconstrained Bayesian estimation analysis is performed using the `cor_test()` function from **BFpack**. The group numbers follow from the order that the data matrices are plugged into `cor_test()`. When printing the object, the 2.5%, 50%, and 97.5% quantiles of the posterior distributions of the separate correlations are shown. Subsequently, the hypotheses are tested using `BF()`:

```
R> memoryHC <- subset(memory,Group=="HC")[,-7]
R> memorySZ <- subset(memory,Group=="SZ")[,-7]
R> Hmisc::rcorr(as.matrix(memoryHC))
R> Hmisc::rcorr(as.matrix(memorySZ))
R> set.seed(123)
R> cor6 <- cor_test(memoryHC,memorySZ)
R> print(cor6)
R> constraints6 <- "Del_with_Im_in_g1 > Del_with_Im_in_g2 &
+    Del_with_Wmn_in_g1 > Del_with_Wmn_in_g2 &
+    Del_with_Cat_in_g1 > Del_with_Cat_in_g2 &
+    Del_with_Fas_in_g1 > Del_with_Fas_in_g2 &
+    Del_with_Rat_in_g1 > Del_with_Rat_in_g2 &
+    Im_with_Wmn_in_g1 > Im_with_Wmn_in_g2 &
+    Im_with_Cat_in_g1 > Im_with_Cat_in_g2 &
+    Im_with_Fas_in_g1 > Im_with_Fas_in_g2 &
+    Im_with_Rat_in_g1 > Im_with_Rat_in_g2 &
+    Wmn_with_Cat_in_g1 > Wmn_with_Cat_in_g2 &
+    Wmn_with_Fas_in_g1 > Wmn_with_Fas_in_g2 &
+    Wmn_with_Rat_in_g1 > Wmn_with_Rat_in_g2 &
+    Cat_with_Fas_in_g1 > Cat_with_Fas_in_g2 &
+    Cat_with_Rat_in_g1 > Cat_with_Rat_in_g2 &
+    Fas_with_Rat_in_g1 > Fas_with_Rat_in_g2"
R> BF6 <- BF(cor6, hypothesis = constraints6)
R> summary(BF6)
```

First we present the estimates and the p values of the classical two-sided correlation tests in the HC group (when running `Hmisc::rcorr(as.matrix(memoryHC))`):

```
      Im  Del  Wmn  Cat  Fas  Rat
Im  1.00 0.83 0.65 0.56 0.39 0.54
Del 0.83 1.00 0.50 0.39 0.32 0.47
Wmn 0.65 0.50 1.00 0.77 0.70 0.61
Cat 0.56 0.39 0.77 1.00 0.73 0.77
Fas 0.39 0.32 0.70 0.73 1.00 0.67
Rat 0.54 0.47 0.61 0.77 0.67 1.00

n= 20

P
```

```
    Im     Del    Wmn    Cat    Fas    Rat
Im         0.0000 0.0018 0.0098 0.0911 0.0132
Del 0.0000        0.0249 0.0848 0.1667 0.0384
Wmn 0.0018 0.0249        0.0000 0.0006 0.0042
Cat 0.0098 0.0848 0.0000        0.0002 0.0000
Fas 0.0911 0.1667 0.0006 0.0002        0.0011
Rat 0.0132 0.0384 0.0042 0.0000 0.0011
```

and in the SZ group (when running `Hmisc::rcorr(as.matrix(memorySZ))`):

```
       Im   Del   Wmn   Cat   Fas   Rat
Im   1.00  0.35 -0.07 -0.28 -0.17  0.08
Del  0.35  1.00 -0.22  0.16  0.27  0.09
Wmn -0.07 -0.22  1.00 -0.05  0.01 -0.02
Cat -0.28  0.16 -0.05  1.00  0.22 -0.25
Fas -0.17  0.27  0.01  0.22  1.00 -0.14
Rat  0.08  0.09 -0.02 -0.25 -0.14  1.00

n= 20

P
    Im     Del    Wmn    Cat    Fas    Rat
Im         0.1353 0.7832 0.2313 0.4669 0.7431
Del 0.1353        0.3441 0.4909 0.2520 0.7122
Wmn 0.7832 0.3441        0.8237 0.9674 0.9450
Cat 0.2313 0.4909 0.8237        0.3541 0.2857
Fas 0.4669 0.2520 0.9674 0.3541        0.5452
Rat 0.7431 0.7122 0.9450 0.2857 0.5452
```

The posterior quantiles of the correlations (viewed when running `print(cor6)`) show a similar pattern because joint uniform priors are used. Here we only show the posterior medians to keep the output in the manuscript as concise as possible[5].

```
Unconstrained Bayesian estimates

Group g1:

Posterior median:
       Im   Del   Wmn   Cat   Fas Rat
Im
Del 0.729
Wmn 0.489 0.297
Cat 0.315 0.202 0.562
```

[5]Note that these numerical estimates may vary somewhat across different platforms (e.g., Windows, Mac, Linux, etc.) even though the same seed is used. The reason is that certain functions for sampling values from probability distributions may not result in identical draws across platforms resulting in small variations of numerical estimates.

```
Fas 0.213 0.148 0.499 0.535
Rat 0.338 0.283 0.389 0.568 0.534

Group g2:

Posterior median:
        Im    Del    Wmn    Cat    Fas Rat
Im
Del  0.283
Wmn -0.066 -0.162
Cat -0.234  0.141 -0.042
Fas -0.104  0.213  0.020  0.148
Rat  0.048  0.086 -0.005 -0.205 -0.088
```

Note that medians are closer to 0 than modes for distributions that are skewed towards 0 which also explains the difference between the Bayesian posterior medians and the MLEs. Based on these results several features are notable: (1) Each of the 15 correlations is higher in the HC group (g1) than the SZ group (g2); (2) On average the correlations among the HC group are rather large; and (3) The correlations within the SZ group are close to 0 on average.

Next we present the results of the exploratory Bayesian tests which are given by

```
Bayesian hypothesis test
Type: exploratory
Object: cor_test
Parameter: correlation coefficients
Method: Bayes factors based on joint uniform priors

Posterior probabilities:
                   Pr(=0) Pr(<0) Pr(>0)
Del_with_Im_in_g1   0.000  0.000  1.000
Wmn_with_Im_in_g1   0.025  0.003  0.972
Cat_with_Im_in_g1   0.142  0.021  0.838
Fas_with_Im_in_g1   0.415  0.122  0.464
Rat_with_Im_in_g1   0.127  0.019  0.854
Wmn_with_Del_in_g1  0.193  0.028  0.779
Cat_with_Del_in_g1  0.418  0.106  0.476
Fas_with_Del_in_g1  0.437  0.123  0.439
Rat_with_Del_in_g1  0.244  0.038  0.718
Cat_with_Wmn_in_g1  0.037  0.006  0.957
Fas_with_Wmn_in_g1  0.041  0.006  0.953
Rat_with_Wmn_in_g1  0.174  0.033  0.793
Fas_with_Cat_in_g1  0.007  0.001  0.992
Rat_with_Cat_in_g1  0.011  0.001  0.987
Rat_with_Fas_in_g1  0.015  0.002  0.983
Del_with_Im_in_g2   0.272  0.056  0.672
Wmn_with_Im_in_g2   0.490  0.314  0.196
```

```
Cat_with_Im_in_g2   0.370  0.537  0.093
Fas_with_Im_in_g2   0.466  0.384  0.151
Rat_with_Im_in_g2   0.490  0.217  0.293
Wmn_with_Del_in_g2  0.434  0.446  0.120
Cat_with_Del_in_g2  0.456  0.148  0.396
Fas_with_Del_in_g2  0.379  0.101  0.520
Rat_with_Del_in_g2  0.486  0.182  0.333
Cat_with_Wmn_in_g2  0.498  0.287  0.215
Fas_with_Wmn_in_g2  0.511  0.219  0.270
Rat_with_Wmn_in_g2  0.499  0.260  0.240
Fas_with_Cat_in_g2  0.440  0.134  0.426
Rat_with_Cat_in_g2  0.388  0.500  0.112
Rat_with_Fas_in_g2  0.475  0.350  0.176
```

In this output the correlations in the (first) HC group and the (second) SZ group end with "`_in_g1`" and "`_in_g2`", respectively. As there are $6 \times 5/2 = 15$ correlations in each of the two groups, in total there are 30 correlations which are tested to be zero, negative, or positive (assuming equal prior probabilities). These posterior probabilities shed some light about whether each correlation is likely to be zero, negative, or positive in light of the observed data.

Given that the overall pattern of estimated correlations across the groups is consistent with the hypotheses in the confirmatory test, simultaneous testing procedures would appear to be a better approach than tests on individual correlations. Indeed, both maximum likelihood and resampling tests convincingly indicated that the covariance and correlation matrices across groups differ ($p < 0.01$). However, there are a number of ways in which two correlation or covariance matrices may differ. Thus, the conventional procedures for comparing matrices do not test the specific hypothesis that, for each of the 15 correlations, the value for the HC group is greater than the value for the SZ group. However hypothesis H_1 can directly be tested against its complement in a straightforward manner using **BFpack**. The results of the confirmatory tests are given below:

```
Bayesian hypothesis test
Type: Confirmatory
Object: cor_test
Parameter: correlation coefficients
Method: Bayes factors based on joint uniform priors for correlations

Posterior probabilities:
   Pr(hypothesis|data)
H1                   1
H2                   0

Evidence matrix (Bayes factors):
   H1       H2
H1  1 5647.244
H2  0    1.000
```

```
Specification table:
   complex= complex> fit=  fit> BF=      BF>       BF PHP
H1        1        0    1 0.146   1 4825.377 4825.377   1
H2        1        1    1 0.854   1    0.854    0.854   0
```

Thus, the Bayes Factor for H_1 against H_2 is approximately 5647 resulting in a posterior probability for H_1 of essentially 1 under the assumption that the two hypotheses are equally likely a priori. Thus the order-constrained analysis indicate decisive support for the researchers' hypothesis.

4.7. Testing intraclass correlations

Statistical model and exploratory hypothesis test

The multilevel or mixed effects model is the gold standard for modeling hierarchically structured data. In the mixed effects model the within-clusters variability is separately modeled from the between-clusters variability. The intraclass correlation plays a central role as a measure of the relative degree of clustering in the data where an intraclass correlation close to 1 (0) indicates a very high (low) degree of clustering in the data. Despite the widespread usage of mixed effects models in the (applied) statistical literature, there are few statistical tests for testing variance components; exceptions include Westfall and Gönen (1996); Gancia-Donato and Sun (2007); Saville and Herring (2009); Thalmann, Niklaus, and Oberauer (2017).

Recently, a Bayesian testing framework was proposed on intraclass correlations (and random intercept variances) was proposed by Mulder and Fox (2019) under a marginal modeling framework (Fox *et al.* 2017; Mulder and Fox 2013). In the marginal model the random effects are integrated out and the intraclass correlations have become covariance parameters in a structured covariance matrix. As a consequence the intraclass correlations can attain negative values. A negative intraclass correlation implies that there is a smaller degree of clustering than under random group assignment. As the intraclass correlations are bounded proper uniform priors can be specified under the constrained hypotheses. For example under the unconstrained model, uniform priors are specified for the intraclass correlations in the interval $(-\frac{1}{p-1}, 1)$, where p is the cluster size. This prior is equivalent to a shifted-F prior on the between-cluster variances. Improper Jeffreys priors are used for the nuisance parameters $\boldsymbol{\beta}$ and ϕ^2. This methodology is implemented in **BFpack**.

First a random intercept model with, possibly, different random intercept variances (yielding different intraclass correlations) for different cluster types needs to be fit using the `lmer()` function from the **lme4** package (Bates, Mächler, Bolker, and Walker 2015). Next the fitted model is plugged into `BF()` to compute Bayes factors and posterior probabilities for the constrained hypotheses on the intraclass correlations. By default, exhaustive exploratory tests are executed of whether each intraclass correlation is zero, negative, or positive, i.e.,

$$H_1 : \rho_c = 0 \text{ versus } H_2 : \rho_c < 0 \text{ versus } H_3 : \rho_c > 0,$$

for the intraclass correlation in cluster type $c = 1, \ldots, C$. Technical details of the methodology can be found in Section A.5.

Confirmatory hypothesis test in educational testing

Data from the Trends in International Mathematics and Science Study (TIMSS; http://www.iea.nl/timss) were used to examine differences in intraclass correlations of four countries – The Netherlands (NL), Hungary (HR), Germany (DE), and Denmark (DK) – with respect to the mathematics achievements of fourth graders (e.g., the first plausible value was used as a measure of mathematics achievement). The sample design of the TIMSS data set is known to describe three levels with students nested within classrooms/schools, and classrooms/schools nested within countries (e.g., one classroom is sampled per school). In this example, the TIMSS 2011 assessment was considered.

The intraclass correlation was defined as the correlation among measured mathematics achievements of grade-4 students attending the same school. This intraclass correlation was assumed to be homogeneous across schools in the same country, but was allowed to be different across countries. For the four countries, differences in intraclass correlations were tested using the Bayes factor. The size of the intraclass correlation can be of specific interest, since sampling becomes less efficient when the intraclass correlation increases. Countries with low intraclass correlations have fewer restrictions on the sample design, where countries with high intraclass correlations require more efficient sample designs, larger sample sizes, or both. Knowledge about the size of the heterogeneity provide useful information to optimize the development of a suitable sample design and to minimize the effects of high intraclass correlations.

The TIMSS data sample in **BFpack** consists of four countries, where data was retrieved from The Netherlands $(93, 112)$, Hungary $(139, 106)$, Germany $(179, 170)$, and Denmark $(166, 153)$ with the sampled number of schools in brackets for 2011 and 2015, respectively. Differences in intraclass correlations were tested conditional on several student variables (e.g., gender, student sampling weight variable). The following hypotheses on intraclass correlations were considered in the analyses. Country-ordered intraclass correlations were considered by hypothesis H_1, equal (invariant) intra-class correlations were represented by hypothesis H_2, and their complement was specified as hypothesis H_3:

$$\begin{aligned} H_1 &: \rho_{NL} < \rho_{HR} < \rho_{DE} < \rho_{DK} \\ H_2 &: \rho_{NL} = \rho_{HR} = \rho_{DE} = \rho_{DK} \\ H_3 &: \text{neither } H_1, \text{ nor } H_2. \end{aligned}$$

The ordering in the intraclass correlations was hypothesized by considering the reported standard errors of the country-mean scores. From the variance inflation factor, $1 + (p-1)\rho$, with p the number of students in each school (balanced design), followed that the variance of the mean increases for increasing values of the intraclass correlation coefficient. As a result, the ordering in estimated standard errors of the average mathematics achievements of fourth graders of the cycles from 2003 to 2015 was used to hypothesize the order in intraclass correlations. From a more substantive perspective, it is expected that schools in the Netherlands do not differ much with respect to their performances (low intraclass correlation) in contrast to Denmark, where school performances may differ considerably (high intraclass correlation).

Analyses using ***BFpack***

A linear mixed effects model was used to obtain (restricted) maximum likelihood estimates of the fixed effects of the student variables and the country means, the four random effects

corresponding to the clustering of students in schools in each country, and the measurement error variance, given the 2011 assessment data.

```
R> library("lme4")
R> timssICC_subset <- subset(timssICC, groupNL11 == 1 | groupHR11 == 1 |
+    groupDE11 == 1 | groupDK11 == 1)
R> lmer7 <- lmer(math ~ -1 + gender + weight + lln +
+    groupNL11 + (0 + groupNL11 | schoolID) +
+    groupHR11 + (0 + groupHR11 | schoolID) +
+    groupDE11 + (0 + groupDE11 | schoolID) +
+    groupDK11 + (0 + groupDK11 | schoolID),
+    data=timssICC_subset)
R> print(lmer7)
```

where the `schoolID` factor variable assigns a unique code to each school, and each country-specific group variable (e.g., `groupNL11`) equals one when it concerns a school in that country and zero otherwise. As the interest is mainly on the random effects variances, we only print (via `print(lmer7)`) these here (to keep the current presentation of the results as concise as possible):

```
Random effects:
 Groups     Name      Variance Std.Dev.
 schoolID   groupNL11  356.2   18.87
 schoolID.1 groupHR11  477.8   21.86
 schoolID.2 groupDE11  633.0   25.16
 schoolID.3 groupDK11  831.3   28.83
 Residual             3429.6   58.56
Number of obs: 8655, groups:  schoolID, 577
```

As can be seen the estimated random effects variances show the expected trend. However to quantify the evidence against competing hypotheses we need to executed a formal statistical test. Therefore, the `lmer` output object (Bates *et al.* 2015) is plugged into the `BF()` function for computing Bayes factors between the hypotheses of interest:

```
R> set.seed(123)
R> constraints7 <- "groupNL11 < groupHR11 < groupDE11 < groupDK11;
+    groupNL11 = groupHR11 = groupDE11 = groupDK11"
R> BF7 <- BF(lmer7, hypothesis = constraints7)
R> summary(BF7)
```

The exploratory tests provide posterior probabilities of whether each intraclass correlation equals zero, negative, or positive. Evidence in favor of a negative intraclass correlation indicates that a multilevel model may not be appropriate for modeling these data (Mulder and Fox 2019). The results are given by

```
Bayesian hypothesis test
Type: Exploratory
Object: lmerMod
```

```
Parameter: intraclass correlations
Method: Bayes factors based on uniform priors

          icc=0 icc<0 icc>0
groupNL11     0     0     1
groupHR11     0     0     1
groupDE11     0     0     1
groupDK11     0     0     1
```

As can be seen the exploratory results indicate that a multilevel model is a appropriate for these data.

In the confirmatory test, the posterior probabilities of the specified hypotheses shows how our beliefs are updated in light of the observed data regarding the hypotheses that were formulated on the variation of school performance across countries. The results are given by

```
Bayesian hypothesis test
Type: Confirmatory
Object: lmerMod
Parameter: intraclass correlations
Method: Bayes factor based on uniform priors

Posterior probabilities:
   Pr(hypothesis|data)
H1               0.568
H2               0.418
H3               0.014

Evidence matrix (Bayes factors):
      H1    H2     H3
H1 1.000 1.359 40.501
H2 0.736 1.000 29.812
H3 0.025 0.034  1.000

Specification table:
   complex= complex> fit=  fit>    BF=    BF>     BF   PHP
H1       NA    0.043   NA 0.644  1.000 15.044 15.044 0.568
H2       NA    1.000   NA 1.000 11.073  1.000 11.073 0.418
H3       NA    0.957   NA 0.356  1.000  0.371  0.371 0.014

Hypotheses:
H1: groupNL11<groupHR11<groupDE11<groupDK11
H2: groupNL11=groupHR11=groupDE11=groupDK11
H3: complement
```

The posterior probabilities of the three hypotheses in the confirmatory test reveal that the order hypothesis H_1 and the equality hypothesis H_2 are approximately equally plausible given the observed data (with $P(H_1 \mid \mathbf{Y}) = 0.568$ and $P(H_2 \mid \mathbf{Y}) = 0.418$), with a slight preference

Statistic	NL	HR	DE	DK
REML	0.094	0.122	0.156	0.195
Mean	0.099	0.126	0.159	0.198
Median	0.098	0.124	0.158	0.198
2.5%	0.061	0.091	0.123	0.157
97.5%	0.146	0.168	0.201	0.245

Table 4: TIMSS 2011: Intraclass correlation estimates for the Netherlands (NL), Hungary (HR), Germany (DE), and Denmark (DK).

for H_1, and that the complement seems unlikely (with $\Pr(H_3\mathbf{Y}) = 0.014$). These results indicate that the degree of clustering is either increasing or stable between countries. More data are needed in order to draw clearer evidence towards the order hypothesis or the equality hypothesis. Similar as when testing group variances, the Bayes factor for testing the equality constraints cannot be expressed as a Savage-Dickey density ratio (Appendix A.5).

Statistical elaborations: Comparison of unconstrained estimates

We end with presenting the unconstrained posterior means, medians, and interval estimates of the ICCs which can be obtained by running

```
R> BF7$estimates
```

The results are represented in Table 4. The REML intraclass correlation estimates are also given for each country, which followed directly from the random effect estimates of the `lmer` output. It can be seen that the posterior mean and REML estimates are essentially equal for these data. The Bayesian analysis however also provides useful interval estimates with a clear Bayesian interpretation using uniform priors. For further reading on the properties of these estimates we refer the interested reader to Mulder and Fox (2019) and Nielsen, Smink, and Fox (2020).

4.8. Relational event models

Statistical model and exploratory hypothesis test

The relational event model (REM) was introduced to analyze sequences of time-stamped relational events between actors in a social network (Butts 2008; DuBois, Butts, McFarland, and Smyth 2013). The REM can be used to understand what mechanisms drive interaction dynamics in a temporal social network (Mulder and Leenders 2019). It builds on the survival (or event history) model with time-varying covariates where the dependent variable is the event rate between all possible dyads of senders and receivers in the network. For the technical details about the methodology we refer the reader to the above references.

The **relevent** package can be used for fitting REMs in R (Butts 2021). As a fitted REM object of class '`rem.dyad`' is currently not supported by **BFpack** (see Table 1), adjusted fractional Bayes factors based on Gaussian approximations (Section A.3) can be computed between constrained hypotheses using the default function of `BF()`. First the REM is fitted using the `rem.dyad()` function. Next the (named) vector with the maximum likelihood estimates

(MLEs), the error covariances matrix, and the sample size are extracted from the `rem.dyad` object, which are plugged in the `BF()` function (Section 3). This calls the default `BF()` function which performs exhaustive exploratory tests on the separate parameters, i.e.,

$$H_1 : \beta_q = 0 \text{ versus } H_2 : \beta_q < 0 \text{ versus } H_3 : \beta_q > 0,$$

for $q = 1, \ldots, Q$. Constrained hypotheses can be specified using the names of the parameter estimates.

Confirmatory hypothesis test in communication networks

As was illustrated by Mulder and Leenders (2019) interaction behavior can be positively driven by past activity between actors and common attributes of actors (also known as homophily). To illustrate this we consider a simulated event sequence consisting of 226 relational events (communication messages) in a small network of 25 actors (generated using the methodology in DuBois *et al.* 2013) belonging to different cultures, and having different locations where they are based. The event rate of actor pair (s, r) at time t, denoted by $\lambda(s, r, t)$, is then modeled using a log linear model,

$$\begin{aligned} \log \lambda(s, r, t) \quad = \quad & \beta_0 + \beta_{\text{inertia}}\, x_{\text{inertia}}(s, r, t) + \beta_{\text{culture}}\, x_{\text{culture}}(s, r) \\ & + \beta_{\text{location}}\, x_{\text{location}}(s, r) \end{aligned}$$

where β_0 is the intercept capturing the baseline of the event rate, β_{inertia} is the inertia effect (i.e., the general tendency for actors to keep sending messages to actors who they sent messages to in the past), $x_{\text{inertia}}(s, r, t)$ is the fraction of past events sent by s that were received by r until time t, $x_{\text{culture}}(s, r)$ and $x_{\text{location}}(s, r)$ are dichotomous variables whether actors s and r have the same culture (1=yes, 0=no) and whether actors s and r are based at the same location (1=yes, 0=no), respectively, and β_{culture} and β_{location} are the corresponding effects.

The following competing hypotheses will be considered:

$$\begin{aligned} H_1 &: \beta_{\text{culture}} = \beta_{\text{location}} > 0 \\ H_2 &: \beta_{\text{culture}} > \beta_{\text{location}} > 0 \\ H_3 &: \beta_{\text{location}} > \beta_{\text{culture}} > 0 \\ H_4 &: \text{neither } H_1, H_2, \text{ nor } H_3. \end{aligned}$$

Hypothesis H_1 assumes that having the same culture and being based at the same location have an equal positive effect on the event rate. Hypothesis H_2 assumes that having the same culture has a larger effect than begin based at the same location, and both effects are positive. Hypothesis H_3 assumes that being based at the same location has a larger effect than having the same culture, and both effects are positive. The complement hypothesis H_4 assumes that neither the constraints under H_1 nor the constraints under H_2 or H_3 hold.

Analyses using ***BFpack***

To test these hypotheses first the unconstrained REM is fit using the `rem.dyad()` function using the **relevent** (Butts 2021):

```
R> library("relevent")
R> CovEventEff <- array(NA, dim = c(3, nrow(actors), nrow(actors)))
```

```
R> CovEventEff[1,,] <- 1
R> CovEventEff[2,,] <- as.matrix(same_culture)
R> CovEventEff[3,,] <- as.matrix(same_location)
R> dimnames(CovEventEff)[[1]] <- c("baseline", "culture", "location")
R> set.seed(9227)
R> remdyad8 <- rem.dyad(edgelist = relevents, n = nrow(actors), effects =
+    c("FrPSndSnd", "CovEvent"), covar = list(CovEvent = CovEventEff),
+    hessian = TRUE, fit.method = "BPM")
R> summary(remdyad8)
```

The MLEs and p values are then given by

```
Relational Event Model (Ordinal Likelihood)

              Estimate     Std.Err Z value Pr(>|z|)
FrPSndSnd   0.60034728  0.48674016  1.2334   0.2174
CovEvent.1  0.00078988 89.32141774  0.0000   1.0000
CovEvent.2  1.21939161  0.13587178  8.9746   <2e-16 ***
CovEvent.3 -0.01028619  0.25387330 -0.0405   0.9677
```

Next the estimates, the error covariance matrix, and the sample size are extracted from the fitted object and plugged in the `BF()` function, together with the constrained hypotheses:

```
R> names(remdyad8$coef) <- c("inertia","baseline","culture", "location")
R> constraints8 <- "culture = location > 0; culture > location > 0;
+    location > culture > 0"
R> estimates8 <- remdyad8$coef
R> covmatrix8 <- remdyad8$cov
R> samplesize8 <- remdyad8$m
R> BF8 <- BF(estimates8, Sigma = covmatrix8, n = samplesize8,
+    hypothesis = constraints8)
R> summary(BF8)
```

In the first line new names are given to the estimated values with a clearer interpretation. These names are then used for formulating constrained hypotheses in the `hypothesis` argument. The estimates, the corresponding error covariance matrix, and the sample size are then extracted from the fitted `rem.dyad` object `remfit`. Subsequently, these are plugged into the `BF()` function which then calls `BF.default`.

For the exploratory analysis the following output is then obtained:

```
Bayesian hypothesis test
Type: Exploratory
Object: numeric
Parameter: general parameters
Method: adjusted fractional Bayes factors using Gaussian approximations

Posterior probabilities:
```

```
          Pr(=0) Pr(<0) Pr(>0)
inertia    0.778  0.024  0.197
baseline   0.883  0.059  0.059
culture    0.000  0.000  1.000
location   0.882  0.061  0.057
```

The results clearly show that working at the same culture has a positive effect given the observed data. For the other parameters the null hypothesis of zero effect is most plausible.

Next we discuss the results of the confirmatory test which is given by

```
Bayesian hypothesis test
Type: confirmatory
Object: numeric
Parameter: general parameter
Method: adjusted fractional Bayes factors using Gaussian approximations

Posterior probabilities:
   Pr(hypothesis|data)
H1               0.000
H2               0.894
H3               0.000
H4               0.106

Evidence matrix (Bayes factors):
         H1    H2          H3    H4
H1    1.000 0.000     46.092 0.001
H2 6070.727 1.000 279808.969 8.412
H3    0.022 0.000      1.000 0.000
H4  721.686 0.119  33263.615 1.000

Specification table:
   complex= complex> fit=  fit> BF=   BF>    BF   PHP
H1    0.136    0.500    0 1.000   0 2.000 0.001 0.000
H2    1.000    0.082    1 0.484   1 5.921 5.921 0.894
H3    1.000    0.185    1 0.000   1 0.000 0.000 0.000
H4    1.000    0.733    1 0.516   1 0.704 0.704 0.106

Hypotheses:
H1: culture=location>0
H2: culture>location>0
H3: location>culture>0
H4: complement
```

The Bayes factors and posterior probabilities reveal there is most evidence for H_2 (with a posterior probability of 0.894), followed by the complement hypothesis H_4 (with a posterior probability of 0.106), and finally hypotheses H_1 and H_3 received a posterior probability of zero. This suggests that there is most support for the hypothesis which assumes that belonging to the same culture has a larger effect on interaction rates than being based at the same location

and that both effects are positive. There is still a probability of 0.106 that the complement may be true after observing the data. This can be explained from the very small negative estimate of the `location` parameter of −0.0103, having a very large standard error of 0.2539. More data would be needed in order to draw more decisive conclusions.

5. Concluding remarks

The R package **BFpack** was designed to allow substantive researchers to perform Bayes factor tests via commonly used statistical functions in R, such as `lm()`, `aov()`, or `glm()`. By specifying a simple character string that captures the hypotheses of interest, users can make use of the flexibility of Bayes factors to simultaneously test multiple hypotheses which may involve equality as well as order (or one-sided) constraints on the parameters of interest. This will allow users to move beyond traditional null hypothesis (significance) testing.

Specific choices were made regarding the Bayes factors and priors that are implemented in **BFpack**. When testing parameters in an unbounded space, adjusted fractional Bayes factors (using minimal fractions) were implemented and when testing parameters in a bounded space, Bayes factors based on proper uniform priors were considered. These Bayes factors are well-developed for testing hypotheses with equality as well as order constraints on the parameters of interest (Appendix A). Furthermore, as was shown in Section 4.5.5, due to the extended Savage-Dickey density ratio, these Bayes factors can be computed in an efficient manner when observations are missing at random. Other Bayes factors and priors could also be considered of course. For testing parameters under a regression model, intrinsic priors (Casella and Moreno 2006; Consonni and Paroli 2017), (hyper) g priors (Bayarri and Garcia-Donato 2007; Liang, Paulo, Molina, Clyde, and Berger 2008; Mulder, Berger, Pena, and Bayarri 2020a), or non-local priors (Johnson and Rossell 2010) could be specified. All these Bayes factors, including the ones implemented in **BFpack**, all abide the notion of minimal prior information (via different routes), and they are all consistent for the proposed testing problems, implying that the evidence for the true hypothesis goes to infinity as the information in the data grows. Thus, the quantification of statistical evidence based on these different Bayes factors may only vary somewhat for relatively small samples, which seems reasonable in the case of limited information. In the future it may be interesting to also implement other Bayes factors in the package.

BFpack is (currently) mainly intended for simple exploratory tests and more complex confirmatory tests where a limited set of hypotheses are formulated with equality and/or order constraints on the parameters together with their prior probabilities based on one's scientific expectations. The default settings will therefore not be suitable when testing a huge number of hypotheses (or models) such as in variable selection problems where the goal is to search for the best model among all 2^k possible regression statistical models of k possible predictor variables, where k is large. In such problems it is crucial to explicitly deal with multiplicity. In a Bayesian framework this can naturally be done by specifying beta priors for the inclusion probabilities (see, Scott and Berger (2010), and the references therein; or the **BAS** package (Clyde *et al.* 2018)). Therefore, when researchers do not have clear expectations about the relationships between the parameters of interest, and when the goal is to search for the best model across all possible constrained models (with any possible combination of equality and order constraints), the prior model probabilities should be appropriately specified to correct for multiplicity. This may be an interesting direction for further research.

The aim is to extend **BFpack** further. In the near future, Bayes factors will be implemented for meta-analyses (Van Aert and Mulder accepted) via the **metafor** package (Viechtbauer 2010), for testing network autocorrelations (Dittrich, Leenders, and Mulder 2017, 2019, 2020), for testing other types of measures of association (e.g., polychoric correlations), and for testing variance components in more general random effects models.

Computational details

For the analyses in Section 4.4 and Section 4.6, the analyses rely on Fortran or C++ compilers which may result in slightly different results under different operating systems when using the same seed. The analyses in Section 4.8 rely on computations with the **relevent** package and results may differ under different operating systems. Note that the output in this paper was generated with R 4.1.1, **relevent** R package 1.0.4, **gfortran** GNU Fortran (GCC) 8.2.0, and Apple **clang** version 12.0.5. Exact results may depend on the operating system, version of R, compiler, and compiler version. However, the results in other setups will be very similar and lead to the same conclusions qualitatively.

Acknowledgments

The authors are grateful to the feedback of the editor and two anonymous reviewers which resulted in important improvements on the presentation of the paper. Joris Mulder is supported by a Vidi grant awarded by the Netherlands Organization of Scientific Research (NWO). Eric-Jan Wagenmakers is supported by a Vici grant awarded by NWO. Regarding the applications, Application 1 was provided by Xin Gu, Application 2 by Herbert Hoijtink, Application 3 by Florian Böing-Messing, Application 4 by Andrew Tomarken, Application 5 by Anton Olsson-Collentine, Application 6 by Andrew Tomarken, Application 7 by Jean-Paul Fox, and Application 8 by Marlyne Meijerink.

References

Barnard J, McCulloch R, Meng XL (2000). "Modelling Covariance Matrices in Terms of StandardDeviations and Correlations with Applications to Shrinkage." *Statistica Sinica*, **10**(4), 1281–1311. ISSN 10170405, 19968507. doi:10.2307/24306780.

Bartlett M (1957). "A Comment on D. V. Lindley's Statistical Paradox." *Biometrika*, **44**, 533–534. doi:10.1093/biomet/44.3-4.533.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Bayarri MJ, Garcia-Donato G (2007). "Extending Conventional Priors for Testing General Hypotheses in Linear Models." *Biometrika*, **94**(1), 135–152. ISSN 0006-3444. doi:10.1093/biomet/asm014.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, Boeck PD,

Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Ho TH, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE (2018). "Redefine Statistical Significance." *Nature Human Behaviour*, **2**, 6–10. doi:10.1038/s41562-017-0189-z.

Berger JO (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis*, **1**, 385–402. doi:10.1214/06-ba115.

Berger JO, Brown LD, Wolpert RL (1994). "A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing." *The Annals of Statistics*, **2522**, 1787–1807. doi:10.1214/aos/1176325757.

Berger JO, Delampady M (1987). "Testing Precise Hypotheses." *Statistical Science*, **2**, 317–335. doi:10.1214/ss/1177013238.

Berger JO, Mortera J (1995). "Discussion to Fractional Bayes Factors for Model Comparison (by O'Hagan)." *Journal of the Royal Statistical Society B*, **56**, 130. doi:10.1111/j.2517-6161.1995.tb02017.x.

Berger JO, Pericchi L (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison." In P Lahiri (ed.), *Model Selection*, pp. 135–207. Institute of Mathematical Statistics, Hayward.

Böing-Messing F, Mulder J (2018). "Automatic Bayes Factors for Testing Equality- And Inequality-Constrained Hypotheses on Variances." *Psychometrika*, **83**(3), 586–617. doi:10.1007/s11336-018-9615-z.

Böing-Messing F, Van Assen MALM, Hofman A, Hoijtink H, Mulder J (2017a). "Bayesian Evaluation of Constrained Hypotheses on Variances of Multiple Independent Groups." *Psychological Methods*, **22**, 262–287. doi:10.1037/met0000116.

Böing-Messing F, Van Assen MALM, Hofman AD, Hoijtink H, Mulder J (2017b). "Bayesian Evaluation of Constrained Hypotheses on Variances of Multiple Independent Groups." *Psychological Methods*, **22**, 262–287. doi:10.1037/met0000116.

Box GEP, Tiao GC (1973). *Bayesian Inference in Statistical Snalysis*. Addison-Wesley, Reading. doi:10.1002/9781118033197.

Braeken J, Mulder J, Wood S (2015). "Relative Effects at Work: Bayes Factors for Order Hypotheses." *Journal of Management*, **41**(2), 544–573. doi:10.1177/0149206314525206.

Butts CT (2008). "A Relational Event Framework for Social Action." *Sociological Methodology*, **38**(1), 155–200. doi:10.1111/j.1467-9531.2008.00203.x.

Butts CT (2021). ***relevent***: *Relational Event Models*. R package version 1.1, URL https://CRAN.R-project.org/package=relevent.

Casella G, Moreno E (2006). "Objective Bayesian Variable Selection." *Journal of American Statistical Association*, **101**, 157–167. doi:10.1198/016214505000000646.

Clyde M, Ghosh J, Littman M (2018). "Bayesian Adaptive Sampling for Variable Selection and Model Averaging." *Behavior Research Methods*, **20**, 80–101. doi:10.1198/jcgs.2010.09049.

Conigliani C, O'Hagan A (2000). "Sensitivity of the Fractional Bayes Factor to Prior Distributions." *Canadian Journal of Statistics*, **28**(2), 343–352. doi:10.2307/3315983.

Consonni G, Paroli R (2017). "Objective Bayesian Comparison of Constrained Analysis of Variance Models." *Psychometrika*. doi:10.1007/s11336-016-9516-y.

Dablander F, Van den Bergh D, Ly A, Wagenmakers EJ (2020). "Default Bayes Factors for Testing the (In)Equality of Several Population Variances." arXiv:2003.06278 [stat.ME], URL https://arxiv.org/abs/2003.06278.

De Jong J, Rigotti T, Mulder J (2017). "One After the Other: Effects of Sequence Patterns of Breached and Overfulfilled Obligations." *European Journal of Work and Organizational Psychology*, **26**(3), 337–355. doi:10.1080/1359432x.2017.1287074.

De Santis F, Spezzaferri F (2001). "Consistent Fractional Bayes Factors for Nested Normal Linear Models." *Journal of Statistical Planning and Inference*, **97**(2), 305–321. doi:10.1016/s0378-3758(00)00240-8.

Dickey J (1971). "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters." *The Annals of Statistics*, **42**(1), 204–223. doi:10.1214/aoms/1177693507.

Dittrich D, Leenders RTJA, Mulder J (2017). "Bayesian Estimation of the Network Autocorrelation Model." *Social Networks*, **48**, 213–236. doi:10.1016/j.socnet.2016.09.002.

Dittrich D, Leenders RTJA, Mulder J (2019). "Network Autocorrelation Modeling: A Bayes Factor Approach for Testing (Multiple) Precise and Interval Hypotheses." *Sociological Methods & Research*, **48**, 642–676. doi:10.1177/0049124117729712.

Dittrich D, Leenders RTJA, Mulder J (2020). "Network Autocorrelation Modeling: Bayesian Techniques for Estimating and Testing Multiple Network Autocorrelations." *Sociological Methodology*, **50**, 168–214. doi:10.1177/0081175020913899.

DuBois C, Butts CT, McFarland D, Smyth P (2013). "Hierarchical Models for Relational Event Sequences." *Journal of Mathematical Psychology*, **57**(6), 297–309. doi:10.1016/j.jmp.2013.04.001.

Fox J, Weisberg S (2021). ***car**: Companion to Applied Regression*. R package version 3.0-12, URL https://CRAN.R-project.org/package=car.

Fox JP, Mulder J, Sinharay S (2017). "Bayes Factor Covariance Testing in Item Response Models." *Psychometrika*, **82**(4), 979–1006. doi:10.1007/s11336-017-9577-6.

Friston KJ, Frith CD (1995). "Schizophrenia: A Disconnection Syndrome?" *Clinical Neuroscience*, **3**, 89–97. doi:10.1016/0920-9964(95)95309-w.

Gancia-Donato G, Sun D (2007). "Objective Priors for Hypothesis Testing in One-Way Random Effects Models." *Canadian Journal of Statistics*, **35**, 302–320. doi:10.1002/cjs.5550350207.

Garcia-Donato G, Forte A (2018). "Bayesian Testing, Variable Selection and Model Averaging in Linear Models Using R with **BayesVarSel**." *The R Journal*, **10**(1), 329. doi:10.32614/rj-2018-021.

Garcia-Donato G, Forte A, Vergara-Hernandez C (2020). ***BayesVarSel**: Bayes Factors, Model Choice and Variable Selection in Linear Models*. R package version 2.0.1, URL https://CRAN.R-project.org/package=BayesVarSel.

Gelfand AE, Smith AFM (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the Americann Statistical Association*, **85**, 398–409. doi:10.1080/01621459.1990.10476213.

Genz A, Bretz F, Miwa T, Mi X, Hothorn T (2021). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 1.1-3, URL https://CRAN.R-project.org/package=mvtnorm.

Gu X, Hoijtink H, Mulder J, Rosseel Y (2019). "**bain**: A Program for the Evaluation of Inequality Constrained Hypotheses Using Bayes Factors in Structural Equation Models." *Journal of Statistical Computation and Simulation*, **89**(8), 1526–1553. doi:10.1080/00949655.2019.1590574.

Gu X, Hoijtink H, Mulder J, Van Lissa CJ, Jones J, Waller N, The R Core Team (2021). ***bain**: Bayes Factors for Informative Hypotheses*. R package version 0.2.6, URL https://CRAN.R-project.org/package=bain.

Gu X, Mulder J, Hoijtink H (2017). "Approximated Adjusted Fractional Bayes Factors: A General Method for Testing Informative Hypotheses." *British Journal of Mathematical and Statistical Psychology*, **71**, 229–261. doi:10.1111/bmsp.12110.

Hoijtink H (2011). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Chapman & Hall/CRC, New York.

Hoijtink H, Gu X, Mulder J (2019a). "Bayesian Evaluation of Informative Hypotheses for Multiple Populations." *British Journal of Mathematical and Statistical Psychology*, **72**, 219–243. doi:10.1111/bmsp.12145.

Hoijtink H, Gu X, Mulder J, Rosseel Y (2019b). "Computing Bayes Factors From Data with Missing Values." *Psychological Methods*, **24**(2), 253–268. doi:10.1037/met0000187.

Hoijtink H, Mulder J, Van Lissa C, Gu X (2019c). "A Tutorial on Testing Hypotheses Using the Bayes Factor." *Psychological Methods*, **24**(5), 539–556. doi:10.1037/met0000201.

Hubbard R, Bayarri MJ (2003). "Confusion over Measures of Evidence (P's) versus Errors (Alpha's) in Classical Statistical Testing." *The American Statistician*, **57**, 171–182. doi:10.1198/0003130031856.

Ichinose MC, Han G, Polyn S, Park S, Tomarken AJ (2019). "Verbal Memory Performance Discordance in Schizophrenia: A Reflection of Cognitive Dysconnectivity?" Data collected by Sohee Park's laboratory at Vanderbilt University that assesses working memory performance among schizophrenic patients and a normal control group.

Irony TZ, de Pereira CAB, Tiwari RC (2000). "Analysis of Opinion Swing: Comparison of Two Correlated Proportions." *The American Statistician*, **54**(1), 57–62. doi:10.1080/00031305.2000.10474510.

Janiszewski C, Uy D (2008). "Precision of the Anchor Influences the Amount of Adjustment." *Psychological Science*, **19**(2), 121–127. doi:10.1111/j.1467-9280.2008.02057.x.

Jeffreys H (1961). *Theory of Probability*. 3rd edition. Oxford University Press, New York.

Johnson VE, Rossell D (2010). "On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests." *Journal of the Royal Statistical Society B*, **72**, 143–170. doi:10.1111/j.1467-9868.2009.00730.x.

Kass RE, Raftery AE (1995). "Bayes Factors." *Journal of American Statistical Association*, **90**, 773–795. doi:10.1080/01621459.1995.10476572.

Klugkist I, Laudy O, Hoijtink H (2005). "Inequality Constrained Analysis of Variance: A Bayesian Approach." *Psychological Methods*, **10**(4), 477–493. doi:10.1037/1082-989x.10.4.477.

Kofler MJ, Rapport MD, Sarver DE, Raiker JS, Orban SA, Friedman LM, Kolomeyer EG (2013). "Reaction Time Variability in ADHD: A Meta-Analytic Review of 319 Studies." *Clinical Psychology Review*, **33**(6), 795–811. doi:10.1016/j.cpr.2013.06.001.

Koop G, Potter SM (1999). "Bayes Factors and Nonlinearity: Evidence from Economic Time Series." *Journal of Econometrics*, **88**(2), 251 – 281. ISSN 0304-4076. doi:10.1016/s0304-4076(98)00031-1.

Krypotos AM, Klugkist I, Engelhard IM (2017). "Bayesian Hypothesis Testing for Human Threat Conditioning Research: An Introduction and the **condir** R Package." *European Journal of Psychotraumatology*, **8**(sup1). doi:10.1080/20008198.2017.1314782.

Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of American Statistical Association*, **103**(481), 410–423. doi:10.1198/016214507000001337.

Lindley DV (1957). "A Statistical Paradox." *Biometrika*, **44**(1), 187–192. doi:10.1093/biomet/44.1-2.187.

Little RJ, Rubin DB (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau QF, Šmíra M, Epskamp S, Matzke D, Wild A, Knight P, Rouder JN, Morey RD, Wagenmakers EJ (2019). "**JASP**: Graphical Statistical Software for Common Statistical Designs." *Journal of Statistical Software*, **88**(2), 1–17. doi:10.18637/jss.v088.i02.

McCullagh P, Nelder J (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall, London.

McGuigin RW, Newton AT, Tamber-Rosenau B, Tomarken AJ, Gauthier I (2020). "Thickness of Deep Layers in the Fusiform Face Area Predicts Face Recognition." *Journal of Cognitive Neuroscience*, **32**, 1316–1329. doi:10.1162/jocn_a_01551.

McGuigin RW, Van Gulick AE, Gauthier I (2016). "Cortical Thickness in Fusiform Face Area Predicts Face and Object Recognition Performance." *Journal of Cognitive Neuroscience*, **28**(2), 282–294. doi:10.1162/jocn_a_00891.

Morey RD, Rouder JN, Jamil T, Urbanek S, Forner K, Ly A (2018). "**BayesFactor**: Computation of Bayes Factors for Common Designs." R package version 0.9.12-4.2, URL https://CRAN.R-project.org/package=BayesFactor.

Mulder J (2014). "Prior Adjusted Default Bayes Factors for Testing (In)Equality Constrained Hypotheses." *Computational Statistics & Data Analysis*, **71**, 448–463. doi:10.1016/j.csda.2013.07.017.

Mulder J (2016). "Bayes Factors for Testing Order-Constrained Hypotheses on Correlations." *Journal of Mathematical Psychology*, **72**, 104–115. doi:10.1016/j.jmp.2014.09.004.

Mulder J, Berger JO, Pena V, Bayarri MJ (2020a). "On the Prevalence of Information Inconsistency in Normal Linear Models." *TEST*, **30**, 103–132. doi:10.1007/s11749-020-00704-4.

Mulder J, Fox JP (2013). "Bayesian Tests on Components of the Compound Symmetry Covariance Matrix." *Statistics and Computing*, **23**, 109–122. doi:10.1007/s11222-011-9295-3.

Mulder J, Fox JP (2019). "Bayes Factor Testing of Multiple Intraclass Correlations." *Bayesian Analysis*, **14**(2), 521–552. doi:10.1214/18-ba1115.

Mulder J, Gelissen J (2019). "Bayes Factor Testing of Equality and Order Constraints on Measures of Association in Social Research." arXiv:1807.05819 [stat.ME], URL https://arxiv.org/abs/1807.05819.

Mulder J, Gu X (2021). "Bayesian Testing of Scientific Expectations Under Multivariate Normal Linear Models." *Multivariate Behavioral Research*. doi:10.1080/00273171.2021.1904809. Forthcoming.

Mulder J, Hoijtink H, De Leeuw C (2012). "**BIEMS**: A Fortran 90 Program for Calculating Bayes Factors for Inequality and Equality Constrained Models." *Journal of Statistical Software*, **46**(2), 1–39. doi:10.18637/jss.v046.i02.

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. doi:10.1016/j.jspi.2009.09.022.

Mulder J, Leenders RTJA (2019). "Modeling the Evolution of Interaction Behavior in Social Networks: A Dynamic Relational Event Approach for Real-Time Analysis." *Chaos,*

Solitons, and Fractals: An Interdisciplinary Journal of Nonlinear Science, **119**, 73–85. doi:10.1016/j.chaos.2018.11.027.

Mulder J, Olsson-Collentine A (2019). "Simple Bayesian Testing of Scientific Expectations in Linear Regression Models." *Behavioral Research Methods*, **51**(3), 1117–1130. doi:10.3758/s13428-018-01196-9.

Mulder J, Van Lissa C, Williams DR, Gu X, Olsson-Collentine A, Böing-Messing F, Fox JP (2021). ***BFpack**: Flexible Bayes Factor Testing of Scientific Expectations*. R package version 1.0.0, URL https://CRAN.R-project.org/package=BFpack.

Mulder J, Wagenmakers EJ (2016). "Editors' Introduction to the Special Issue "Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments"." *Journal of Mathematical Psychology*, **72**, 1–5. doi:10.1016/j.jmp.2016.01.002.

Mulder J, Wagenmakers EJ, Marsman M (2020b). "A Generalization of the Savage-Dickey Density Ratio for Testing Equality and Order Constrained Hypotheses." *The American Statistician*, pp. 1–8. doi:10.1080/00031305.2020.1799861.

Nielsen NM, Smink WAC, Fox JP (2020). "Small and Negative Cluster Correlations." *Behaviormetrika*, **48**, 51–77. doi:10.1007/s41237-020-00130-8.

Nuijten MB, Wetzels R, Matzke D, Dolan CV, Wagenmakers EJ (2014). "A Default Bayesian Hypothesis Test for Mediation." *Behavior Research Methods*, **47**, 85–97. doi:10.3758/s13428-014-0470-2.

O'Hagan A (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society B*, **57**(1), 99–138. doi:10.1111/j.2517-6161.1995.tb02017.x.

Pericchi LR, Liu G, Torres D (2008). "Objective Bayes Factors for Informative Hypotheses: "Completing" the Informative Hypothesis and "Splitting" the Bayes Factors." In H Hoijtink, I Klugkist, PA Boelen (eds.), *Bayesian Evaluation of Informative Hypotheses*, pp. 131–154. Springer-Verlag.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). ***nlme**: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-153, URL https://CRAN.R-project.org/package=nlme.

Porter MD (2016). "A Statistical Approach to Crime Linkage." *The American Statistician*, **70**(2), 152–165. doi:10.1080/00031305.2015.1123185.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ripley BD (2021). ***MASS**: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-54, URL https://CRAN.R-project.org/package=MASS.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009). "Bayesian t Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin & Review*, **16**(2), 225–237. doi:10.3758/pbr.16.2.225.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons. doi:10.1002/9780470316696.

Rubin DB (1996). "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association*, **91**(434), 473–489. doi:10.2307/2291635.

Russell VA, Oades RD, Tannock R, Killeen PR, Auerbach JG, Johansen EB, Sagvolden T (2006). "Response Variability in Attention-Deficit/Hyperactivity Disorder: A Neuronal and Glial Energetics Hypothesis." *Behavioral and Brain Functions*, **2**(1), 1–25. doi:10.1186/1744-9081-2-30.

Saville BR, Herring AH (2009). "Testing Random Effects in the Linear Mixed Model Using Approximate Bayes Factors." *Biometrics*, **65**(2), 369–376. doi:10.1111/j.1541-0420.2008.01107.x.

Scott J, Berger JO (2010). "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem." *The Annals of Statistics*, **38**, 2587–2619. doi:10.1214/10-aos792.

Sellke T, Bayarri MJ, Berger JO (2001). "Calibration of p Values for Testing Precise Null Hypotheses." *The American Statistician*, **55**(1), 62–71. doi:10.1198/000313001300339950.

Silverstein SM, Como PG, Palumbo DR, West LL, Osborn LM (1995). "Multiple Sources of Attentional Dysfunction in Adults with Tourette's Syndrome: Comparison with Attention Deficit-Hyperactivity Disorder." *Neuropsychology*, **9**(2), 157–164. doi:10.1037/0894-4105.9.2.157.

Thalmann M, Niklaus M, Oberauer K (2017). "Estimating Bayes Factors for Linear Models and Random Slopes and Continuous Predictors." doi:10.31234/osf.io/4xqvr. PsyArXiv.

Van Aert R, Mulder J (accepted). "Bayesian Hypothesis Testing and Estimation Under the Marginalized Random-Effects Meta-Analysis Model." *Psychonomic Bulletin & Review*. doi:10.31234/osf.io/ktcq4.

Van Buuren S, Groothuis-Oudshoorn CGM (2011). "**mice**: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67. doi:10.18637/jss.v045.i03.

Van Buuren S, Groothuis-Oudshoorn K (2021). ***mice**: Multivariate Imputation by Chained Equations*. R package version 3.13.0, URL https://CRAN.R-project.org/package=mice.

Van Ravenzwaaij D, Monden R, Tendeiro JN, Ioannidis JP (2019). "Bayes Factors for Superiority, Non-Inferiority, and Equivalence Designs." *BMC Medical Research Methodology*, **19**. doi:10.1186/s12874-019-0699-7.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag.

Viechtbauer W (2010). "Conducting Meta-Analyses in R with the **metafor** Package." *Journal of Statistical Software*, **36**(3), 1–48. doi:10.18637/jss.v036.i03.

Wagenmakers EJ (2007). "A Practical Solution to the Pervasive Problem of p Values." *Psychonomic Bulletin and Review*, **14**(5), 779–804. doi:10.3758/bf03194105.

Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, Selker R, Gronau QF, Smira M, Epskamp S, Matzke D, Rouder JN, Morey RD (2018). "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications." *Psychonomic Bulletin & Review*, **25**(1), 35–57. doi:10.3758/s13423-017-1343-3.

Well SV, Kolk AM, Klugkist I (2008). "Effects of Sex, Gender Role Identification, and Gender Relevance of Two Types of Stressors on Cardiovascular and Subjective Responses: Sex and Gender Match/Mismatch Effects." *Behavior Modification*, **32**(4), 427–449. doi:10.1177/0145445507309030.

Westfall P, Gönen M (1996). "Asymptotic Properties of ANOVA Bayes Factors." *Communications in Statistics – Theory and Methods*, **25**, 3101–3123. doi:10.1080/03610929608831888.

Wetzels R, Grasman RPPP, Wagenmakers EJ (2010). "An Encompassing Prior Generalization of the Savage-Dickey Density Ratio Test." *Computational Statistics & Data Analysis*, **38**, 666–690. doi:10.1016/j.csda.2010.03.016.

Wilson JP, Rule NO (2015). "Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes." *Psychological Science*, **26**, 1325–1331. doi:10.1177/0956797615590992.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. URL https://CRAN.R-project.org/doc/Rnews/.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27**(8), 1–25. doi:10.18637/jss.v027.i08.

A. Technical and computational details

A.1. Adjusted fractional Bayes factors under multivariate normal model

Under a multivariate normal linear model for J groups with K predictor variables, the i-th observations of the p-th outcome variable is defined by

$$y_{ip} = \sum_{j=1}^{J} d_{ij}\mu_{jp} + \sum_{k=1}^{K} x_{ik}\beta_{kp} + \epsilon_{ip},$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iP})' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $d_{ij} = 1$ if the i-th observation belongs to group j, and zero elsewhere, μ_{jp} is the (adjusted) mean of the p-th dependent variable for group j, x_{ik} is the i-th observation of the k-the predictor variable, and β_{kp} is the effect of the k-th predictor variable on the p-th outcome variable. The assumed distribution of the i-th observation can compactly be written as $\mathbf{y}_i \sim N(\boldsymbol{\Theta}^\top \tilde{\mathbf{x}}_i, \boldsymbol{\Sigma})$, where $\tilde{\mathbf{x}}_i^\top = (\mathbf{d}_i^\top, \mathbf{x}_i^\top)$ and $\boldsymbol{\Theta}^\top = [\mathbf{M}^\top\ \mathbf{B}^\top]$, where $\mathbf{M}$ is a $J \times P$ matrix where the (j,p)-th element is the (adjusted) group mean μ_{jp} and $\mathbf{B}$ is a $K \times P$ matrix where the (k,p)-th element is β_{kp}. Tests that fall under this model are (multivariate) t tests, univariate/multivariate regression, (M)AN(C)OVA, among others.

Under the adjusted fractional Bayes factor, which is employed for Bayesian hypothesis testing under the multivariate normal linear model, the marginal likelihood for the constrained hypothesis H_t is defined by

$$p_t(\mathbf{Y}) = \frac{\int_{\boldsymbol{\Sigma}} \int_{\boldsymbol{\Theta}_t} \prod_{i=1}^{N} p(\mathbf{y}_i \mid \mathbf{d}_i, \mathrm{b}fx_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) |\boldsymbol{\Sigma}|^{-\frac{P+1}{2}} d\boldsymbol{\Theta} d\boldsymbol{\Sigma}}{\int_{\boldsymbol{\Sigma}} \int_{\boldsymbol{\Theta}_t^*} \prod_{i=1}^{N} p(\mathbf{y}_i \mid \mathbf{d}_i, \mathbf{x}_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma})^{b_i} |\boldsymbol{\Sigma}|^{-\frac{P+1}{2}} d\boldsymbol{\Theta} d\boldsymbol{\Sigma}}, \tag{6}$$

where $\boldsymbol{\Theta}$ contains the (adjusted) group means and regression effects, on which constraints are formulated under H_t, the constrained parameter space under H_t is defined by $\boldsymbol{\Theta}_t = \{\boldsymbol{\Theta} \mid \mathbf{R}^e\boldsymbol{\theta} = \mathbf{r}^e\ \&\ \mathbf{R}^o\boldsymbol{\theta} > \mathbf{r}^o\}$, where $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$, b_i denotes a group specific minimal fraction which is equal to $\frac{(P+K)/J}{N_j}$ if the i-th observation belongs to the j-th group, and the adjusted parameter space is defined by $\boldsymbol{\Theta}_t^* = \{\boldsymbol{\Theta} \mid \mathbf{R}^e(\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{r}^e\ \&\ \mathbf{R}^o(\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > \mathbf{r}^o\}$, where $\boldsymbol{\theta}_0 = [\mathbf{R}^{e'}\mathbf{R}^{o'}]^{'-1}[\mathbf{r}^{e'}\mathbf{r}^{o'}]'$, using the generalized Moore-Penrose inverse. Unlike the standard definition of a marginal likelihood in Equation 2, no proper prior needs to be specified based on external (subjective) considerations. Instead, the marginal likelihood in Equation 6, implicitly uses a fraction, denoted by "b_i" for the i-th observations, for constructing a fractional prior, while the remaining fraction of the information in the data is used for hypothesis testing. Minimal fractions are used so that the remaining fraction that is used for hypothesis testing is maximal (Berger and Mortera 1995; Conigliani and O'Hagan 2000). The generalization of the original fractional Bayes factor (O'Hagan 1995) to group specific fractions ensures that the implied fractional prior contains minimal information in the case of unbalanced data with unequal group sizes and it avoids inconsistent selection behavior (De Santis and Spezzaferri 2001; Hoijtink, Gu, and Mulder 2019a). Furthermore, the adjusted parameter space results in a shift of the unconstrained fractional prior to the boundary of the constrained space (Mulder 2014). This ensures that the default Bayes factor that captures the relative complexity of an order or one-sided hypothesis as the relative size of the constrained parameter space, e.g., 0.5 for a univariate one-sided test of $H_1 : \mu > 0$, which covers half of the parameter space, as a result of an unconstrained fraction prior that is centered at the test value 0.

Consequently the default Bayes factor of a constrained hypothesis against an unconstrained hypothesis can then be written as a Savage-Dickey density ratio in Equation 5 of the form

$$B_{tu} = \frac{\pi_u(\boldsymbol{\theta}^e = \mathbf{r}^e \mid \mathbf{Y})}{\pi_u(\boldsymbol{\theta}^e = \mathbf{r}^e \mid \mathbf{Y}, \mathbf{b})} \times \frac{P_u(\boldsymbol{\theta}^o > \mathbf{r}^o \mid \boldsymbol{\theta}^e = \mathbf{r}^e, \mathbf{Y})}{P_u(\boldsymbol{\theta}^o > \mathbf{r}^o \mid \boldsymbol{\theta}^e = \mathbf{r}^e \mid \mathbf{Y}, \mathbf{b})}, \tag{7}$$

where $\boldsymbol{\theta}^e = \mathbf{R}^e\boldsymbol{\theta}$ and $\boldsymbol{\theta}^o = \mathbf{R}^o\boldsymbol{\theta}$, and the unconstrained posterior for $\boldsymbol{\Theta}$, where $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$, in the numerators follows a $K \times P$ matrix t distribution and the unconstrained fractional prior in the denominators follows a $K \times P$ matrix Cauchy distribution given by

$$\begin{aligned} \pi_u(\boldsymbol{\Theta} \mid \mathbf{Y}) &= T_{K\times P}(\hat{\boldsymbol{\Theta}}, (\mathbf{X}^\top\mathbf{X})^{-1}, \mathbf{S}, N - K - P + 1) \\ \pi_u(\boldsymbol{\Theta} \mid \mathbf{Y}, \mathbf{b}) &= C_{K\times P}(\boldsymbol{\Theta}_0, (\mathbf{X}_\mathbf{b}^\top\mathbf{X}_\mathbf{b})^{-1}, \mathbf{S}_\mathbf{b}), \end{aligned}$$

where the first element in the matrix t distribution is a location parameter, the second and third element are scale matrices, and the fourth element is the degrees of freedom, and the first element in the matrix Cauchy distribution is a location parameter, the second and third element are scale matrices. Furthermore, the OLS estimate is given by $\hat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, the sums of square matrix equals $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}})$, the sums of square matrix in the fractional prior equals $\mathbf{S}_\mathbf{b} = (\mathbf{Y}_\mathbf{b} - \mathbf{X}_\mathbf{b}\hat{\boldsymbol{\Theta}}_\mathbf{b})^\top(\mathbf{Y}_\mathbf{b} - \mathbf{X}_\mathbf{b}\hat{\boldsymbol{\Theta}}_\mathbf{b})$, with $\hat{\boldsymbol{\Theta}}_\mathbf{b} = (\mathbf{X}_\mathbf{b}^\top\mathbf{X}_\mathbf{b})^{-1}\mathbf{X}_\mathbf{b}^\top\mathbf{Y}_\mathbf{b}$, where $\mathbf{Y}_\mathbf{b}$ and $\mathbf{X}_\mathbf{b}$ are the stacked matrices of $\mathbf{y}_{i,b_i}^\top$ and $\mathbf{x}_{i,b_i}^\top$, with $\mathbf{y}_{i,b_i} = \sqrt{b_i}\mathbf{y}_i$ and $\mathbf{x}_{i,b_i} = \sqrt{b_i}\mathbf{x}_i$, and the i-th fraction is equal to the $\frac{K+P}{JN_j}$ if the i-th observation belongs to group j, where N_j is the sample size of group j. The fact that the fractional prior has a matrix Cauchy distribution (which is equivalent to a matrix Student t distribution with 1 degree of freedom, implying minimal information) is a direct consequence of the group specific minimal fractions. The fact that the unconstrained prior is located at the value that is tested, as $\mathbf{R}^e\boldsymbol{\theta}_0 = \mathbf{r}^e$ and $\mathbf{R}^o\boldsymbol{\theta}_0 = \mathbf{r}^o$ hold, where $\boldsymbol{\theta}_0 = \text{vec}(\boldsymbol{\Theta}_0)$, is a direct consequence of the prior adjusted parameter space. This implies that small deviations from the test value are more likely a priori than large deviations and that negative deviations are equally likely a priori as positive deviations from the test value, similar as in Jeffreys (1961) recommendations for prior specifications. Note that other commonly used priors are also centered at the test value, such as intrinsic priors (Casella and Moreno 2006; Consonni and Paroli 2017), (hyper) g priors (Bayarri and Garcia-Donato 2007; Liang *et al.* 2008; Mulder *et al.* 2020a), or non-local priors (Johnson and Rossell 2010). For technical details on the derivation this default Bayes factor we refer the interested reader to Mulder and Olsson-Collentine (2019) for the univariate regression model, and to Mulder and Gu (2021) for the general multivariate normal model with multiple groups. Furthermore, for univariate models the probability densities in the first factor in Equation 7 can be computed using the `dmvt()` function in the **mvtnorm** package (Genz *et al.* 2021), and the probabilities in the second factor can be computed using the `pmvt()` function in the **mvtnorm** package. For multivariate models we use a Monte Carlo estimate using the fact that a matrix-variate Student t distribution can be written as an inverse Wishart mixture of matrix-variate normal distributions (Box and Tiao 1973; Mulder and Gu 2021).

A.2. Adjusted fractional Bayes factors for testing group variances

The generalized adjusted fractional Bayes methodology is also used for testing equality/order constraints on group variances (Böing-Messing *et al.* 2017b; Böing-Messing and Mulder 2018). As variances are scale parameters, the adjustment is done by re-scaling the fractional prior instead of the shifting the fractional prior as was done when testing location parameters

under the multivariate normal linear model. The fractional priors of the group variances then follow inverse gamma distributions based on minimal fractions of $\frac{2}{n_j}$ in group j, as each group contains two unknown parameters (a group mean and a group variance), where n_j is the sample size of group j. This Bayes factor cannot be written as an extended Savage-Dickey density ratio for testing equality constraints. As the final expression of the marginal likelihoods are quite extensive we refer the interested reader to Böing-Messing *et al.* (2017b, Appendix B) for the technical details. Interestingly, Dablander, Van den Bergh, Ly, and Wagenmakers (2020) showed that this default Bayes factor is virtually identical to an actual Bayes factor based on a minimally informative beta prior on the group variance and the sum of the group variances.

A.3. Adjusted fractional Bayes factors using Gaussian approximations

When testing coefficients under non-normal models, such as generalized linear models (Section 4.4) or relational event models (Section 4.8), an approximation of this default Bayes factor can be used (Gu *et al.* 2017), which is also implemented in **BFpack** as `BF.default()`. It relies on a large sample Gaussian approximation of the unconstrained posterior of the key parameters in the numerator in Equation 7, i.e., $\pi_u(\boldsymbol{\theta} \mid \mathbf{Y}) \approx N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where the mean and covariance matrix can be computed using available R packages (e.g., using `glm()` for generalized linear models). Similarly, the adjusted fractional prior based on a common fraction b can be written as $\pi_u(\boldsymbol{\theta} \mid \mathbf{Y}, b) \approx N(\boldsymbol{\theta}_0, b^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where the prior mean satisfies $\mathbf{R}^e\boldsymbol{\theta}_0 = \mathbf{r}^e$ and $\mathbf{R}^o\boldsymbol{\theta}_0 = \mathbf{r}^o$, and the prior covariance matrix is a re-scaled version of the approximated posterior covariance matrix based on a minimal fraction, which is equal to number of key parameters divided by the sample size, i.e., $b = \frac{\dim(\boldsymbol{\theta})}{N}$. The probabilities densities in the first factor can then be computed using `dmvnorm()` and the probabilities can be computed using `pmvnorm()` using the **mvtnorm** package (Genz *et al.* 2021).

A.4. Testing measures of association

The dependent variables in group j are assumed to follow a multivariate normal model, $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j$ is the mean vector under group j (which are nuisance parameters), $\boldsymbol{\Sigma}_j$ is the covariance matrix, such that $\boldsymbol{\Sigma}_j = \text{diag}(\boldsymbol{\sigma}_j)\mathbf{C}_j\text{diag}(\boldsymbol{\sigma}_j)$, where $\boldsymbol{\sigma}_j$ is a vector containing the standard deviations of the dependent variables in group j, and $\mathbf{C}_j$ is the correlation matrix, where its (p, q)-th element is the correlation between variable p and q in group j, denoted by ρ_{jpq}. Proper uniform priors are specified for the free correlations under a constrained hypotheses in the restricted space of positive definite correlation matrices, and improper independent Jeffreys priors are used for the nuisance parameters. This implies a proper joint uniform prior under the unconstrained model, which is given by

$$\pi(\mathbf{C}_j) = V_j^{-1} \times I(\mathbf{C}_j > 0)$$

where $\mathbf{C}_j > 0$ implies that $\mathbf{C}_j$ is positive definite, V_j is the volume of the parameter space of positive definite matrices, $V_j = \int_{\mathbf{C}_j > 0} 1 d\mathbf{C}_j$, and $I(\cdot)$ is the indicator function.

It can be shown that the Bayes factor for a constrained hypothesis on measures of association against an unconstrained alternative can be written as Equation 5. In order to compute the four quantities in the Savage-Dickey density ratio, it is useful to apply a Fisher transformation to the correlations, as the resulting unconstrained posterior then follows an approximate multivariate normal distribution (Mulder 2016; Mulder and Gelissen 2019), i.e., $\pi_u(\boldsymbol{\rho}^F \mid \mathbf{Y}) \approx$

$N(\boldsymbol{\mu}_{\boldsymbol{\rho}}^F, \boldsymbol{\Sigma}_{\boldsymbol{\rho}}^F)$, where the Fisher transformation of a single correlation is defined by $\rho_{jpq}^F = \frac{1}{2}\log\left(\frac{1+\rho_{jpq}}{1-\rho_{jpq}}\right)$, where ρ_{jpq} and ρ_{jpq}^F denote the correlation in group j between variables p and q and its corresponding Fisher transformed equivalence, respectively. Therefore for the posterior parts in the numerators in Equation 5 we can make use of `dmvnorm()` and `pmvnorm()` from the **mvtnorm** package (Genz *et al.* 2021). For the prior parts a numerical (non-parametric) estimates are used. Technical details on the computation can be found in Mulder and Gelissen (2019, Section 4 and 5).

A.5. Testing intraclass correlations

A Bayes factor testing procedure is implemented for intraclass correlations (and random intercept variances) under a marginal modeling framework where the random effects are integrated out (Mulder and Fox 2019; Fox *et al.* 2017; Mulder and Fox 2013), i.e.,

$$\left\{\begin{array}{lcl} y_{cij} & = & \mathbf{x}_{cij}^\top\boldsymbol{\beta} + \delta_{ci} + \epsilon_{cij} \\ \delta_{ci} & \sim & N(0, \tau_c^2) \\ \epsilon_{cij} & \sim & N(0, \sigma^2) \end{array}\right. \Rightarrow \left\{\begin{array}{lcl} \mathbf{y}_{cj} & \sim & N\left(\mathbf{X}_{cj}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{cj}\right) \\ \boldsymbol{\Sigma}_{cj} & = & \phi^2(1-\rho_1)\left(\mathbf{I}_{p_{cj}} + \frac{\rho_c}{1-\rho_c}\mathbf{J}_{p_{cj}}\right), \end{array}\right.$$

where y_{cij} is the outcome variable of the i-th observation in cluster j of cluster type c, $\mathbf{x}_{cij}$ contain its predictor variables, $\boldsymbol{\beta}$ are the fixed effects, τ_c^2 is the between-cluster variance in cluster type c, σ^2 is the within-cluster variance, $\phi^2 = \tau_1^2 + \sigma^2$ is the total variance in cluster 1, $\rho_c = \frac{\tau_c^2}{\tau_c^2+\sigma^2}$ is the intraclass correlation in cluster type c, and p_{cj} is the number of observations in cluster j of cluster type c. Furthermore $\mathbf{I}_{p_{cj}}$ is a $p_{cj} \times p_{cj}$ identity matrix and $\mathbf{J}_{p_{cj}}$ is a $p_{cj} \times p_{cj}$ matrix of ones.

As explained above the intraclass correlations become covariance parameters under the integrated model which may attain negative values while keeping the covariance matrix positive definite: The covariance matrix of the observations in cluster j of type c, $\boldsymbol{\Sigma}_{cj}$, is positive definite if $\rho_c \in (-\frac{1}{p_{cj}-1}, 1)$ (where the negative lower bound depends on the cluster size p_{cj}). In **BFpack** proper uniform priors are assumed for the intraclass correlations, i.e., under cluster type c a uniform prior is set for ρ_c in the interval $(-\frac{1}{\max_j\{p_{cj}\}-1}, 1)$, where $\max_j\{p_{cj}\}$ is the largest cluster size of type c. Specifying this prior (and the more general stretched beta prior) is equivalent to a shifted F prior on the between-cluster variance (Mulder and Fox 2019, Lemma 1). Improper independent Jeffreys priors are used for the nuisance parameters $\boldsymbol{\beta}$ and ϕ^2.

By allowing the intraclass correlations to be negative we can test the appropriateness of a random effects model using the posterior probability that an intraclass correlation is positive. By default **BFpack** computes posterior probabilities for the hypotheses assuming a zero, a negative, or a positive intraclass correlation. Similar as when testing group variances, the equality part of the Bayes factor of a constrained hypothesis on the intraclass correlations against an unconstrained alternative cannot be expressed as a Savage-Dickey density ratio. For the technical derivations of the marginal likelihood we refer the interested reader to Mulder and Fox (2019, Lemma 3)

B. Computing Bayes factors with missing observations

Handling incomplete data matrices due to missing data is a ubiquitous problem in statistical practice. The natural Bayesian solution is to treat the missing observations as unknown parameters and sample them from their posterior predictive distribution in the MCMC sampler to properly take the induced uncertainty caused by the missing observations into account in the estimation of the model parameters. In the current situation of model uncertainty, we would thus need to compute the marginal likelihoods using the respective posterior predictive distributions under all constrained models under investigation. This would be a great computational burden when considering many different hypotheses with different combinations of equality and order constraints on the parameters of interest.

The computation can be greatly simplified when the Bayes factor can be written as an extended Savage-Dickey density ratio in Equation 5. As a simple example, let us consider a precise test of $H_1 : \theta = 0$ versus $H_2 : \theta \in \mathbb{R}$ (with nuisance parameters $\boldsymbol{\phi}$) using the Savage-Dickey density ratio,

$$B_{12} = \frac{\pi_u(\theta = 0 \mid \mathbf{Y}_{obs})}{\pi_u(\theta = 0)}, \tag{8}$$

where $\mathbf{Y}_{obs}$ denotes the matrix with the observed data. Using standard probability calculus it follows automatically that

$$\pi_u(\theta = 0 \mid \mathbf{Y}_{obs}) = \int \pi_u(\theta = 0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{miss}) p_u(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs}) d\mathbf{Y}_{miss},$$

where $p_u(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs})$ is the posterior predictive distribution under the unconstrained model. This basic identity plays a central role in multiple imputation (Rubin 1996, p. 476). Following the theory on multiple imputation, we could obtain a statistically valid estimate of the unconstrained posterior density at $\theta = 0$ based on the *observed* data as the arithmetic average based on the complete-data posterior of θ under the unconstrained model, i.e.,

$$\pi_u(\theta = 0 \mid \mathbf{Y}_{obs}) \approx M^{-1} \sum_{m=1}^{M} \pi_u(\theta = 0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{miss}^{(m)}),$$

where $\mathbf{Y}_{miss}^{(m)}$ denotes the m-th draw from the unconstrained posterior predictive distribution of the missing observations.

To obtain complete-data matrices from the unconstrained (marginal) posterior predictive distribution we use the R package **mice** (Van Buuren and Groothuis-Oudshoorn 2021). The MICE algorithm (Van Buuren and Groothuis-Oudshoorn 2011) is a Gibbs sampler, which is a Bayesian simulation technique that samples sequentially from the conditional posterior distributions to obtain draws from the joint distribution (Gelfand and Smith 1990). The required posterior estimate can then be obtained using the above Monte Carlo estimate based on the complete-data matrices acquired by **mice**. Hence to obtain the Bayes factor in Equation 8 we only need a sample of complete data matrices under the unconstrained model, and not require a sample under the equality constrained model. This holds in general when considering many different constrained hypotheses with equality and order constraints where the Bayes factor of each constrained hypothesis against an unconstrained alternative can be written as an extended Savage-Dickey density ratio where all the statistical quantities need to be estimated under the same unconstrained model. Therefore we only need to get one sample

of the complete-data matrices from the unconstrained posterior predictive distribution using **mice** to obtain the statistical quantities to compute the Bayes factors for all constrained hypotheses against the unconstrained alternative. This methodology can also be used for computing the relative complexity measures based on the fractional prior in the generalized adjusted fractional Bayes factor (Mulder and Gu 2021) and its approximation (Hoijtink *et al.* 2019b) when missing observations are present.

Affiliation:

Joris Mulder
Department of Methodology and Statistics
Tilburg University
Warrandelaan 1
5000 LE, Tilburg, The Netherlands
E-mail: j.mulder3@tilburguniversity.edu
URL: https://jorismulder.com/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2021, Volume 100, Issue 18 *Submitted:* 2019-06-28
doi:10.18637/jss.v100.i18 *Accepted:* 2021-04-13