The title

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

The title

Introduction

This paper presents a software package called `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. 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 Development Core Team, 2013). Bayes factors enjoy many useful practical and theoretical properties which are not generally shared by classical significance tests. This includes its intuitive interpretation as the relative evidence in the data between two hypotheses, its ability to simultaneously test multiple hypotheses which may contain equality as well as order constraints on the parameters of interest, and its consistent behavior which implies that the true hypothesis will be selected with probability one as the sample size grows. The interested reader is referred to the many important contributions including (but not limited to) Jeffreys (1961); Berger and Delampady (1987); Sellke, Bayarri, and Berger (2001); Wagenmakers (2007); Rouder, Speckman, D. Sun, and Iverson (2009); Masson (2011); Hoijtink (2011); Wagenmakers et al. (2018); Hoijtink, Mulder, Lissa, and Gu (2019), and the references therein. This has resulted in an increasing literature where Bayes factors have been used for testing scientific expectations (Braeken, Mulder, & Wood, 2015; Dogge, Gayet, Custers, Hoijtink, & Aarts, 2019; Flore, Mulder, & Wicherts, 2019; Gronau et al., 2017; Hoijtink & Chow, 2017; Jong, Rigotti, & Mulder, 2017; Mulder & Wagenmakers, 2016; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Van de Schoot et al., 2006; van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019; van Schie, Van Veen, Engelhard, Klugkist, & Van den Hout, 2016; Vrinten et al., 2016; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2017; Well, Kolk, & Klugkist, 2008; Zondervan-Zwijnenburg et al., 2019). The Bayes factors that are implemented in `BFpack` are based on recent developments of Bayesian hypothesis testing of location parameters, such as (adjusted) means, regression coefficients, and other location

parameters (Gu, Hoijtink, Mulder, & Rosseel, 2019; Gu, Mulder, & Hoijtink, 2017; Mulder, 2014b; Mulder et al., 2019), variance components, such as group variances and intraclass correlations (Böing-Messing & Mulder, 2017; Mulder & Fox, 2019), and measures of association, (Mulder, 2016; Mulder & Gelissen, 2019). The package allows users to perform (i) exploratory Bayesian tests of whether a model parameter equals zero, is negative, or is positive, and (ii) confirmatory Bayesian tests where users 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` (Hothorn et al., 2019) and `car` (Fox & Weisberg, 2019), which contain functions for classical significance tests of a single equality constrained null hypothesis via `lmtest::coeftest()` and `car::linearHypothesis()`, for example. 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 for a complete overview), and in the case of a confirmatory test a string that specifies a set of competing hypotheses (example 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 statistics measures which cannot be obtained under a frequentist framework, such as default measures of the relative evidence in the data between competing statistical hypotheses as quantified by the Bayes factor. When testing hypotheses using the Bayes factor, the use of arbitrary or ad hoc priors should generally be avoided (Bartlett, 1957; Berger & Pericchi, 2001; Jeffreys, 1961; Lindley, 1957). 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 (e.g., means or regression coefficients in univariate/multivariate normal linear models), measures of association (e.g., correlations in multivariate normal distributions), and variance components (e.g., group variances, intraclass correlations). For testing unbounded parameters, such as location parameters and variances adjusted fractional Bayes factors (Böing-Messing & Mulder, 2017; Mulder, 2014b; O'Hagan, 1995) have been implemented. These Bayes factors have analytic expressions and are therefore easy to compute. %Under a fractional Bayes methodology the data is split in a minimal fraction, which is used for default prior specification, and a maximal fraction, which is used for hypothesis testing. For testing bounded parameters, such as measures of association and intraclass correlations, proper uniform priors are implemented. When testing intraclass correlations under random intercept models, a novel marginal modeling approach is employed where the random effects are integrated out (Fox, Mulder, & Sinharay, 2017; Mulder & Fox, 2013, 2019). On the one hand, these tests can be used for testing hypotheses on intraclass correlations based on substantive considerations, and on the other hand, the tests can be used as a tool when 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 resulting in an approximate Gaussian posterior 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 hypotheses in general statistical models. Note that even though it is possible to also use the approximate Bayes factor for the testing problems for which exact tailor-made Bayes factors are available in `BFpack`, we recommend to use the exact tailored Bayes factors if they are available as the exact Bayes factors result in exact quantification of the evidence between statistical hypotheses instead of an approximate quantification of the evidence. Table **??** shows for which models an exact Bayes factor is implemented and which 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 et al., 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, & 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., 2018) 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. `BFpack` package also builds on some of the functionality of `bain` in more complex statistical models. Unlike `bain` however, the implementation in `BFpack` builds on existing `R` functions such as `dmvnorm` or `pmvnorm` from the `mvtnorm` package (Genz et al., 2016) instead of calling external `Fortran` subroutines. This result in Bayes factors that essentially have zero Monte Carlo errors. Furthermore it is important to note that the Gaussian nature of the default prior in `bain` may not appropriate when testing bounded parameters, for example, such as measures of association or intraclass correlations, or when the Gaussian approximation of the posterior

would be too crude, such as when testing group variances in the case of small sample sizes. 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`. The Bayes factors implemented in `JASP` rely on other packages such as `BayesFactor` and `bain`. The contribution of `BFpack` is therefore to provide `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. Currently the package can be downloaded and installed from Github by running:

```
devtools::install_github("jomulder/BFpack")
```

It will be made available throught CRAN in the near future. The paper is organized as follows. Section describes the key aspects of the Bayes factor methodology that is implemented in `BFpack`. This section separately describes Bayes factors for location parameters, for measures of association, and for variance components. Section gives a general explanation how the main function `BF` should be used. Section presents 8 different applications of the methodology and software for a variety of testing problems.

```
papaja::apa_table(read.csv("c:/tmp/paper BFpack/table2.csv"), caption =  "Example hypoth
```

Technical background of the default Bayes factors

The general form of the hypotheses that can be tested using `BFpack` consists a set of linear equality constraints and a set of linear order constraints on the vector of model parameters, denoted by θ of size P, i.e.,

$$H_t : \mathbf{R}^E\boldsymbol{\theta} = \mathbf{r}^E \ \& \ \mathbf{R}^O\boldsymbol{\theta} > \mathbf{r}^O, \tag{1}$$

where $[\mathbf{R}^E|\mathbf{r}^E]$ is a $q^E \times P$ augmented matrix specifying the equality constraints and $[\mathbf{R}^O|\mathbf{r}^O]$ is a $q^O \times P$ augmented matrix specifying the order constraints. A hypothesis index is omitted to keep the notation simple. In the case that $\mathbf{R}^O$ is of full row rank (which is most often the case), a parameter transformation can be applied according to

$$\begin{bmatrix} \boldsymbol{\theta}^E \\ \boldsymbol{\theta}^O \\ \boldsymbol{\phi} \end{bmatrix} = \mathbf{T}\boldsymbol{\theta} = \begin{bmatrix} \mathbf{R}^E \\ \mathbf{R}^O \\ \mathbf{D} \end{bmatrix} \boldsymbol{\theta}, \tag{2}$$

where the q^E equality restricted parameters equal $\boldsymbol{\theta}^E = \mathbf{R}^E\boldsymbol{\theta}$, the q^O order-restricted parameters equal $\boldsymbol{\theta}^O = \mathbf{R}^E\boldsymbol{\theta}$, and the $P - q^O - q^E$ nuisance parameters equal $\boldsymbol{\phi} = \mathbf{D}\boldsymbol{\theta}$, where the $(P - q^E - q^O) \times P$ dummy matrix $\mathbf{D}$ is chosen such that the transformation is one-to-one. Subsequently the hypothesis can equivalently be formulated as

$$H_t : \boldsymbol{\theta}^E = \mathbf{r}^E \;\&\; \boldsymbol{\theta}^O > \mathbf{r}^O, \tag{3}$$

where the nuisance parameters $\boldsymbol{\phi}$ are omitted. Note that for most order hypotheses, the matrix $\mathbf{R}^O$ will be of full row rank. For example, $H_t : \theta_1 > \theta_2 > \theta_3$ implies that $[\mathbf{R}^O|\mathbf{r}^O] = \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{array}\right]$. Therefore we will work with the formulation in Equation 3 throughout this paper to keep the notation simple. In the case $\mathbf{R}^O$ is not of full row rank, which is for instance the case for $H_t : (\theta_1, \theta_2) > (\theta_3, \theta_4)$, a similar type of formulation of H_t can be produced as in Equation 3[1]. Next we specify a prior for the free (possibly order constrained) parameters under H_t, denoted by π_t, by truncating an unconstrained prior, π_u, that is specified under an unconstrained alternative model,

$$\pi_t(\boldsymbol{\theta}^O, \boldsymbol{\phi}) = \pi_u(\boldsymbol{\theta}^O, \boldsymbol{\phi}|\boldsymbol{\theta}^E = \mathbf{r}^E) \times \pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E)^{-1} \times P(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E)^{-1} \times I(\boldsymbol{\theta}^O > \mathbf{r}^O), \tag{4}$$

[1] If $\mathbf{R}^O$ is not of full row rank, then Equation 3 would become $H_t : \boldsymbol{\theta}^E = \mathbf{r}^E \;\&\; \tilde{\mathbf{R}}^O\boldsymbol{\theta}^O > \tilde{\mathbf{r}}^O$, where $\tilde{\mathbf{R}}^O = \mathbf{R}^O\tilde{\mathbf{D}}^{-1}$, where the $(P - q^E) \times P$ matrix $\tilde{\mathbf{D}}$ consists of the unique rows of $\mathbf{I}_P - \mathbf{R}^{E^\top}(\mathbf{R}^E\mathbf{R}^{E^\top})^{-1}\mathbf{R}^E$, and $\tilde{\mathbf{r}}^O = \mathbf{r}^O - \mathbf{R}^O[\mathbf{R}^E]^{-1}\mathbf{r}^E$, where (generalized) Moore-Penrose inverses are used for the non square matrices.

where $I(\cdot)$ denotes the indicator function. Using this pair of priors under the constrained hypothesis H_t and the unconstrained alternative hypothesis, we can write the Bayes factor of H_t against H_u as

$$B_{tu} = \frac{\pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y})}{\pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E)} \times \frac{P_u(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{P_u(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E)}, \tag{5}$$

where the first factor is ratio of posterior and prior densities of $\boldsymbol{\theta}$ evaluated at a constant vector $\mathbf{R}^E$, which can be viewed as a multivariate Savage-Dickey density ratio (Dickey, 1971; Mulder, Hoijtink, & Klugkist, 2010; Wetzels, Grasman, & Wagenmakers, 2010), and the second factor is a ratio of conditional posterior and prior probabilities that the order constraints hold conditional on the equality constraints. We shall refer to Equation 5 as the extended Savage-Dickey density ratio. Different variations have been reported in the literature of this simple expression of the Bayes factor including Klugkist, Laudy, and Hoijtink (2005), Pericchi, Liu, and Torres (2008), Mulder et al. (2010), Gu et al. (2017), among others. The expression can simply be computed when the marginal and conditional posterior and priors belong to known probability distributions (examples will be given later), and thus direct computation of the marginal likelihood, which can be a challenging problem, can be avoided. The four different statistical measures in Equation 5 have the following intuitive interpretations:

- 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 as 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 as 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 as 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 $\boldsymbol{\theta}^E = \mathbf{r}^E$ or 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 as 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).

It is important to note that by conditioning on $\boldsymbol{\theta}^E = \mathbf{r}^E$ in Equation 4, we make specific assumptions about the prior of the free parameters under H_t in relation to the unconstrained prior (Marin & Robert, 2010; Verdinelli & Wasserman, 1995), and therefore the expression should be used with some care (for an interesting discussion on this topic, see Consonni & Veronese, 2008). Below we provide examples of Bayes factors that can and Bayes factors that cannot be expressed as an extended Savage-Dickey density ratio.

Testing location parameters

Many common testing problems in statistical science involve testing of location parameters that determine the `location' or`shift' of the distribution of the data. Examples of location parameters are means, regression coefficients, or factor loadings. These parameters are unbounded for which flat improper priors are specified under an objective Bayesian estimation framework, i.e., $\pi_u(\boldsymbol{\theta}) \propto 1$. Fractional Bayes methodology is an effective framework for testing location parameters. Informative (subjective) prior specification is avoided by splitting the data in a minimal fraction that is used for updating a noninformative improper prior to a proper default prior and a maximal fraction that is used

for hypothesis testing (De Santis & Spezzaferri, 1999; O'Hagan, 1995). Despite the various useful properties of fractional Bayes factors (e.g., consistency, coherence when testing multiple hypotheses, invariance to transformations of the data, O'Hagan, 1997), an adjustment was needed in order for the fractional Bayes factor to function as an Occam's razor when testing order hypotheses (Mulder, 2014b; Mulder & Olsson-Collentine, 2019). This is achieved by shifting the default prior to the boundary of the constrained space[2]. In the simple case when testing $\theta < 0$ versus $\theta > 0$, the default prior would be centered at 0 (instead of around the MLE) so that the prior probabilities of $\theta < 0$ and $\theta > 0$ under the unconstrained model are equal to 0.5, which suggests that a negative effect is equally likely as a positive effect. Centering the unconstrained prior to the boundary also resulted in desirable testing behavior of order hypotheses when using intrinsic Bayes factors (Mulder, 2014a; Mulder et al., 2010) and when using the BIC (Mulder & Raftery, n.d.) Interestingly when testing location parameters with flat improper priors, the adjusted fractional Bayes factor (and the fractional Bayes factor as well) of H_t against H_u can be expressed as an extended Savage-Dickey density ratio as in Equation 3, i.e.,

$$B^F_{tu} = \frac{\pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y})}{\pi^*_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y}^b)} \times \frac{P_u(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{P^*_u(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y}^b)}, \tag{6}$$

where the distributions conditional on $\mathbf{Y}^b$ in the denominator denote the unconstrained default priors that contain a minimal fraction b of the complete data $\mathbf{Y}$, and the asterisk (*) denotes the default prior adjustment. When the data contains information from different groups and the sample sizes highly varies across groups, it is generally recommended to use group specific fractions to properly control the amount of prior information from each group (De Santis & Spezzaferri, 2001; H. Hoijtink et al., 2018a).

Univariate/multivariate normal linear models. Recently, (Mulder et al., 2019) derived the adjusted fractional Bayes factor for testing hypotheses under the multivariate normal linear model with multiple groups. Under this model the unconstrained posterior of

[2] When testing a constrained hypothesis of the form of Equation 3, the default prior is centered on the boundary which implies that the prior is centered around $\boldsymbol{\theta}_0$ with $\mathbf{R}^E\boldsymbol{\theta}_0 = \mathbf{r}^E$ and $\mathbf{R}^O\boldsymbol{\theta}_0 = \mathbf{r}^O$.

the matrix of location parameters follows a matrix Student t distribution, and the unconstrained default prior has a matrix Cauchy distribution, i.e.,

$$B^F_{tu} = \frac{\mathcal{T}(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y})}{\mathcal{C}(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y}^{\mathbf{b}})} \times \frac{\mathcal{T}(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{\mathcal{C}(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y}^{\mathbf{b}})}, \tag{7}$$

where the $\mathcal{T}_u$ and $\mathcal{C}_u$ denote the unconstrained matrix Student t and matrix Cauchy distribution, respectively, and $\mathbf{b}$ denotes a vector of minimal fractions that are group specific. Under these matrix-variate distributions, the posterior and prior densities at $\mathbf{r}^E$, and the conditional posterior and prior probabilities that the order constraints hold do not have analytic expressions. In `BFpack`, these quantities are computed using Monte Carlo integration. We use the fact that draws from a matrix Student t and matrix Cauchy distribution can be obtained by first sampling a covariance matrix from an inverse Wishart distribution, and subsequently drawing the matrix of location parameters from its respective matrix Gaussian distribution conditional on the drawn covariance matrix (Box & Tiao, 1973). Therefore, the posterior density evaluated at $\boldsymbol{\theta}^E = \mathbf{r}^E$ can be obtained by repeatedly drawing covariance matrices from its marginal posterior, and subsequently, computing the posterior density as the arithmetic average of the Gaussian densities evaluated at $\boldsymbol{\theta}^E = \mathbf{r}^E$, which have analytic expressions. The Gaussian densities are computed using the `dmvnorm` function from the `mvtnorm` package (Genz et al., 2016). Such a procedure is also implemented to obtain the prior density, and the conditional prior and posterior probabilities. The Gaussian probabilities are obtained using the `pmvnorm` function from the `mvtnorm` package. In case the constraints are formulated only on the effects belonging to the same dependent variable, or only on the effects belonging to the same independent (predictor) variable, the marginal and conditional distributions for the unconstrained parameters follow multivariate Student t distributions. The respective measures of relative complexity and fit then have analytic expressions which are efficiently computed using `dmvt` and `pmvt` (`mvtnorm` package) in `BFpack`. Finally note that fractional Bayes factors between the constrained hypotheses using the coherence property of the Bayes factor, i.e., $B_{12} = B_{1u}/B_{2u}$. This Bayes factor test is executed when the data are fitted using the `R` functions `t_test`, `lm`, `aov`, and `manova`. Note

that the usual t test function in `R`, `t.test`, cannot be used because the output (of class `htest`) does not contain the observed sample means and sample variances of the two groups which are needed for the computation of the Bayes factors. For this reason, the equivalent function `t_test` was used (from the `bain` package) which contains the sample means and variances in addition to the standard output of `t.test`.

General statistical models. Under more complex statistical models where the four quantities in Equation 6 do not have analytic expressions or when they cannot be computed efficiently via Monte Carlo estimation, an approximation of the adjusted fractional Bayes factor can be used (Gu et al., 2019, 2017). This approximation relies on large sample theory where the unconstrained posterior and default prior are approximated with Gaussian distributions. As such this approximate default Bayes factor can be viewed as a Bayesian counterpart of the classical Wald test. First the nuisance parameters are integrated out to yield the marginal posterior for $(\boldsymbol{\theta}^E, \boldsymbol{\theta}^O)$. Following large sample theory (Ch. 4 Gelman et al., 2014), this posterior can then be approximated with a multivariate Gaussian distribution using the MLE and error covariance matrix. The approximated Gaussian posteriors can then be used to get estimates of the posterior quantities in the numerators in Equation 6. The corresponding default prior for $(\boldsymbol{\theta}^E, \boldsymbol{\theta}^O)$ is obtained by raising the posterior to a minimal fraction b, which results in a multivariate Gaussian distribution where the error covariance matrix is multiplied with the reciprocal of the minimal fraction, and the mean is shifted towards the boundary of the constrained space. This default Bayes factor can then be written as

$$B^F_{tu} = \frac{\mathcal{N}(\boldsymbol{\theta}^E = \mathbf{r}^E | \mathbf{Y})}{\mathcal{N}(\boldsymbol{\theta}^E = \mathbf{r}^E | \mathbf{Y}^b)} \times \frac{\mathcal{N}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{\mathcal{N}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y}^b)}, \tag{8}$$

where $\mathcal{N}$ denotes an unconstrained multivariate (or matrix-variate) normal distribution. In `BFpack` the posterior and prior densities, and posterior and prior probabilities are directly computed using `dmvnorm` and `pmvnorm` functions from the `mvtnorm` package, respectively. In the case the matrix of order constraints is not of full row rank, `pmvnorm` cannot be used for computing the needed probabilities. In this special case the `bain` function is called from the

`bain` package. Hence, in order to compute the approximated Bayes factor in Equation 8 only the MLEs, the error covariance matrix, and the sample size are needed. The minimal fraction is then set equal to the number of parameters that are tested divided by the sample size (Gu et al., 2019). These three elements can simply be extracted from fitted model objects obtained using other packages in `R`. Currently `BFpack` supports objects of class `glm`, `coxph`, `rem`, `rem.dyad`, `polr`, `survreg`, and `zeroinfl`. When executing `BF()` on an object of these classes, the function `BF_Gaussian` is called which extracts the estimates, the error covariance matrix, and the sample size from the fitted model object to compute Equation 8 for the hypotheses of interest. Thus, this Bayes factor can easily be used for testing hypotheses under other model classes by calling the function `BF_Gaussian`.

Testing measures of association

Correlation coefficients and other measures of association play a central role in applied research to quantify the strength of the linear relationship between two variables, possibly controlling for other variables. Measures of association abide two conditions. First they are bounded between -1 and 1, and second they lie in a correlation matrix which must be positive definite. The second condition implies that a correlations need to satisfy a complex set of constraints (e.g., Rousseeuw & Molenberghs, 1994). The volume of this subspace for increasing dimensions of the correlation matrix was explored by Joe (2006). As measures of association are bounded, fractional Bayes methodology is not needed as the noninformative joint uniform prior for the correlations in the correlation matrix is already proper, and thus a regular default Bayes factor can be computed. This was also recommended by Jeffreys (1935). This joint uniform prior assumes that any configuration of correlations that results in a positive definite correlation matrix is equally likely a priori. Equivalently, proper uniform priors can be formulated for the measures of association under the constrained hypotheses under investigation. It is easy to show that this proper uniform prior under H_t can be written as a truncation of the unconstrained joint uniform prior as in Equation 4, and

therefore, the Bayes factor of constrained hypothesis against an unconstrained alternative can be expressed as an extended Savage-Dickey density ratio in Equation 5, where the unconstrained prior in the denominator is the joint uniform prior and the unconstrained posterior is proportional to the likelihood and this uniform prior (Mulder & Gelissen, 2019). Furthermore as was shown by Mulder (2016) the unconstrained posterior for the measures of association can be well approximated with a multivariate normal distribution after a Fisher transformation of the parameters. This can be explained by the fact that the sample correlation and the population correlation have a similar role in the likelihood (Johnson & Kotz, 1970), and therefore approximate normality is achieved for the posterior when using a noninformative prior such as the employed joint uniform prior. The Bayes factor on measures of association that is implement in `BFpack` can therefore be written as[3]

$$B_{tu} = \frac{\mathcal{N}(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y})}{\mathcal{U}(\boldsymbol{\theta}^E = \mathbf{r}^E)} \times \frac{\mathcal{N}(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{\mathcal{U}(\boldsymbol{\theta}^O > \mathbf{r}^O|\boldsymbol{\theta}^E = \mathbf{r}^E)}, \tag{9}$$

To obtain the prior measures for relative complexity, numerical estimates can be obtained by approximating the joint prior using unconstrained draws, from which the prior density and probability can simply be computed using the number of draws satisfying the constraints. In `BFpack` this is done by calling `Fortran 90` subroutines from `R`. %The multivariate posterior density for the measures of association can be well approximated using a multivariate normal distribution on the Fisher transformed space. This approximation will be sufficiently accurate due to the unconstrained joint uniform prior that is used for the measures of association and the noninformative priors for the nuisance parameters. This Bayes factor test can be executed when the fitted model is a multivariate linear regression model (so that the fitted object is of class `mlm`). %and the argument `prior=correlation` is set (more detail is provided in Sections **??** and . Furthermore, an approximation of the Bayes factor is

[3] Note there is a slight abuse of notation in Equation 9 as both the numerator and denominator for $\boldsymbol{\theta}$ have to lie on the same scale to avoid the Borel-Kolmogorov paradox [@Wetzels2010]. In the computation in 'BFpack', the numerator and denominator are either both computed under the Fisher transformed space or under the untransformed space depending on the test.

obtained when the fitted model object is obtained using the `R` function `hetcor` (from the `polycor` package; Fox, 2016). The mean vector and covariance matrix of the approximately multivariate normal posterior for the measures of association are obtained by extracting the estimates and standard errors from the `hetcor` object.

Testing variance components

Testing group variances. Testing the heterogeneity of group variances plays a central role in psychological science and related fields. A default Bayes factor for testing equality and order hypotheses was developed by (Böing-Messing & Mulder, 2017) using adjusted fractional Bayes methodology. As variance parameters belong to the family of scale parameters, a scale adjustment is needed to obtain a default Bayes factor that functions as an Occam's razor for order hypotheses on variances (instead of a location shift as for location parameters, see Böing-Messing & Mulder, 2016, 2018). Because the noninformative independence Jeffreys prior for group variances across competing equality constrained hypotheses does not satisfy Equation 4, the fractional Bayes factor for the equality part (i.e., the first factor in Equation 3) cannot be expressed as a Savage-Dickey density ratio but the ratio of (conditional) probabilities is present. The Bayes factor for the group variance test can be written as follows

$$B^F_{tu} = B^F_{t'u} \times \frac{\mathcal{IG}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{\mathcal{IG}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y}^b)}, \tag{10}$$

where $B^F_{t'u}$ denotes the fractional Bayes factor of hypothesis $H_{t'} : \boldsymbol{\theta}^E = \mathbf{r}^E$ (i.e., hypothesis H_t where the order constraints are omitted, see also Pericchi et al., 2008) against H_u, and $\mathcal{IG}$ denotes an unconstrained inverse gamma distribution. We refer the interested reader to (Böing-Messing & Mulder, 2017) for the mathematical expressions and derivations. This Bayes factor test can be executed when the fitted model is obtained from the `R` function `bartlett_test`, designed for `BFpack`. This test is equivalent to the usual `bartlett.test` but the output object (of class `BF_bartlett`) also contains sample variances and sample sizes which are needed for computing the Bayes factors in Equation 10.

Testing between-cluster variances and intraclass correlations in mixed effects models. 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). The complicating factor is that testing whether the between-cluster variance equals zero is a boundary problem. In `BFpack` 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 (Fox et al., 2017; Mulder & Fox, 2013, 2019). Under the marginal model the intraclass correlations become covariance parameters which may attain negative values. This crucial step allows us to test the appropriateness of a random effects model using the posterior probability that an intraclass correlation is positive. The implemented Bayes factors make use of stretched uniform priors 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. 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. The Bayes factor can be written as

$$B_{tu} = B_{t'u} \times \frac{\text{shifted-}\mathcal{F}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E, \mathbf{Y})}{\text{shifted-}\mathcal{F}(\boldsymbol{\theta}^O > \mathbf{r}^O | \boldsymbol{\theta}^E = \mathbf{r}^E)}, \tag{11}$$

where shifted-$\mathcal{F}$ refers to the fact that the conditional draws for the between cluster variances are drawn from shifted-F priors in the Gibbs sampler. The marginal likelihood is estimated using importance sampling; see Mulder and Fox (2019) for the mathematical details. These Bayes factors can be used for testing the degree of clustering in the data (e.g.,

testing whether clustering is present among students from different schools), or for testing whether the degree of clustering varies across different cluster categories (e.g., testing the degree of clustering among students from private schools against the degree of clustering among students from public schools). To execute these tests, an object from the `lmer` function with random intercepts (which may be category specific) is needed. Currently `BFpack` only supports intraclass correlation testing in the case of equally sized clusters.

Bayes factor computation for data with missing observations

Bayesian (and non-Bayesian) hypothesis testing in the case the data contains missing observations has not received a lot of attention in the literature. This is quite surprising as missing data are ubiquitous in statistical practice. If the data contain missing observations, listwise deletion is generally not recommended as this results in a loss of information and possible bias (Rubin, 1987, 1996). Multiple imputation is generally the recommended method in which many imputed data sets are randomly created using an imputation model. The analyses are then performed over all the imputed data sets, and averaged in a proper manner (Little & Rubin, 2002). In the case of model uncertainty, properly handling missing data may become increasingly complex as different imputation models need to be used for computing the marginal likelihoods under the different hypotheses. H. Hoijtink et al. (2018b) however showed that the computation can be considerably simplified for specific Bayes factors and testing problems. This is also the case for Bayes factors that can be expressed as the extended Savage-Dickey density ratio in Equation 5. The reason is that the four key quantities (i.e., the measures of relative fit and relative complexity for the equality and order constraints) are all computed under the same unconstrained model. Therefore we only need to get an unbiased estimate of the unconstrained posterior (and possibly unconstrained default prior in the case of a data-based prior), and use this to estimate the four key quantities. If we write a complete data matrix $\mathbf{Y}$ as a data matrix which only contains the observations, $\mathbf{Y}^o$, and a data matrix which only contain the missings as $\mathbf{Y}^m$, the

relative fit of the equality constraints can be computed as

$$\begin{aligned} \pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y}^o) &= \int \pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y}^o, \mathbf{Y}^m)\pi_u(\mathbf{Y}^m|\mathbf{Y}^o)d\mathbf{Y}^m \\ &\approx M^{-1}\sum_{m=1}^{M} \pi_u(\boldsymbol{\theta}^E = \mathbf{r}^E|\mathbf{Y}^o, \mathbf{Y}^{(m)}), \end{aligned}$$

where $\mathbf{Y}^{(m)}$ is the m-th set of imputed missing observations given the observed data matrix $\mathbf{Y}^o$, for $m = 1, \ldots, M$. Similar expressions can be obtained for the other three measures. Section illustrates how to compute Bayes factors and posterior probabilities via the output from `BFpack` in the presence of missing data using the imputation software of the `mice` package (van Buuren et al., 2019).

Bayes factor testing using the package

The Bayes factors described in the previous section can be executed by calling the function `BF`. The function has the following arguments:

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

- 'hypothesis', a string that specifies the hypotheses with equality and/or order constraints on the parameters of interest.

  - The parameter names are based on the names of the estimated effects. Thus, if the coefficients in a fitted 'lm' object have the names 'weight', 'height', and 'length', then the constraints in the 'hypothesis' argument should be formulated on these names.

  - 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 effect of 'weight' is larger than the effect of 'height'

and that the effects of ‘height‘ is positive, and the second hypothesis assumes that the two effects 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 both the effects of ‘weight‘, ‘height‘, and ‘length‘ are positive. This could equivalently have been written as ‘hypothesis = "weight > 0 “&“ height > 0 ‘&‘ length > 0"‘.

- In the case the subspaces under the hypotheses do not cover the complete parameter space, a complement hypothesis is automatically added. 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 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 the effect of ‘weight‘ is larger than the effect of ‘height‘, and the effect of ‘height‘ is larger than the effect of ‘length‘) is nested in the second hypothesis (which assumes that the effects of ‘weight‘ is largest but no constraints are specified between the effects of ‘height‘ and ‘length‘). The reason is that the Bayes factor for the simpler hypothesis against the more complex hypothesis would 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 @Mulder2010]. If however a researcher has theoretical reasons to formulate nested order hypotheses these can be formulated

and tested using the ‘BF‘ function of course.

  - The default setting is ‘hypothesis = NULL‘, which only gives the output for exploratory tests of whether each parameter is zero, negative, or positive when assuming equal prior probabilities, e.g., ‘hypothesis = "weight = 0; weight < 0; weight > 0‘, for the effect of ‘weight‘. This exploratory tests is also executed when a confirmatory test is of interest via the ‘hypothesis‘ argument.

  - When testing hypotheses on variance components (Section), only simple constraints are allowed where a parameter is equal to, greater than, or smaller than another parameter. When testing intraclass correlations, the intraclass correlation can also be compared to 0 under a hypothesis.

- ‘prior‘, a numeric vector of prior probabilities of the hypotheses. The default setting is ‘prior = NULL‘ which specifies equal prior probabilities.

- ‘parameter‘, a character string specifying the parameter type of interest. Currently this argument is only used for an object of class ‘mlm‘, where ‘parameter = regression‘ (default) performs a Bayes factor test on the regression coefficients and ‘parameter = correlation‘ performs a Bayes factor test on the correlations under the multivariate normal model.

The output is an object of class `BF`. When printing an object of class `BF` via the `print()` function, the posterior probabilities for the hypotheses under evaluation are provided, or, in the case `hypothesis = NULL`, the posterior probabilities are given for exploratory tests of whether each parameter is zero, negative, or positive. The `summary()` function shows the results for the exploratory tests, and if hypotheses are specified in the `hypothesis` argument, the results of the confirmatory tests consisting of the posterior probabilities of the hypotheses of interest, the evidence matrix which shows the Bayes factor between each pair of two hypotheses, a specification table which shows all the measures of

relative fit and complexity for the equality and/or order constraints of the hypotheses, and an overview of the hypotheses that are tested.

Applications

This section presents a variety of testing problems that can be executed using `BFpack`.

Application 1: Bayesian t testing in medical research

The example for a one-sample t test was discussed in [**?**,]p. 196]Howell:2012, and originally presented in Rosa, Rosa, Sarner, and Barrett (1998). An experiment was conducted to investigate whether practitioners of the therapeutic touch (a widely used nursing practice) can effectively identify which of their hands is below the experimenter's under blinded condition. Twenty-eight practitioners were involved and tested 10 times in the experiment. Researchers expected an average of 5 correct answers from each practitioner as it is the number by chance if they do not outperform others. In this example, the data are the number of correct answers from 0 to 10 of $n = 28$ practitioners. The null and alternative hypotheses are $H_1 : \mu = 5$ and $H_2 : \mu > 5$ where μ is the mean of the data. If $H_1 : \mu = 5$ is true, it means that practitioners give correct answers by chance, whereas if $H_2 : \mu > 5$, this implies that practitioners do better than expected by random chance. The `BF` function automatically adds the complement hypothesis, $H_3 : \mu < 5$, which would imply that practitioners do worse than expected by chance. As there is virtually no prior belief that H_3 may be true, and we (for this example) assume that the hypotheses of interest, H_1 and H_2, are equally likely a priori we set the prior probabilities for H_1, H_2, and H_3 in the confirmatory test to 0.5, 0.5, and 0, respectively, using the `prior` argument. Hypotheses $H_1 : \mu = 5$ versus $H_2 : \mu > 5$ are tested used the frequentist t test function `t\_test` from the `R` package `bain` and Bayesian t test function `BF` in the `R` package `BFpack`.

```
devtools :: install_github("jomulder/BFpack")

install.packages("bain")
```

```r
library(BFpack)
library(bain)
ttest1 <- t_test(therapeutic, alternative = "greater", mu = 5)
print(ttest1)
BF1 <- BF(ttest1, hypothesis = "mu = 5; mu > 5", prior = c(.5,.5,0))
summary(BF1)
```

The first six lines install and load the R package `BFpack`. In the 8th line, `t\_test` function renders classical right one-sided t test and stores the result in object `ttest1`, which contains t value, degree of freedom, and p value, as well as 95% confidence interval: \ The results of the exploratory tests show that the posterior probabilities of the precise null ($\mu = 5$), a negative effect ($\mu < 5$), and a positive effect ($\mu > 5$) are 0.345, 0.634, and 0.021, respectively, while assuming equal prior probabilities for the three hypotheses, i.e., $P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$. The results from the exploratory test show that the presence of a negative is most plausible given the observed data but the evidence is relatively small as there is still a probability of 0.345 that the precise null is true, and a small probability of 0.021 that there is a positive population effect. The exploratory test however ignores the researchers prior expectations that the first two hypotheses were assumed to be equally likely while there was no reason to believe that the third hypothesis could be true, i.e., $P(H_1) = P(H_2) = \frac{1}{2}$ and $P(H_3) = 0$. Taking these prior probabilities into account, the confirmatory test shows that there is clearly most evidence that a therapeutic touch does not exist (H_1) with a posterior probability of 0.943, followed by the hypothesis that a therapeutic touch exists (H_2) with a posterior probability of 0.057. Furthermore the `Evidence matrix` shows that the Bayes factor for H_1 against H_2 equals 16.473, which is equal to the ratio of the (non rounded) posterior probabilities of the respective hypotheses as equal prior probabilities were specified. Finally the `Specification table` shows that the measures of relative complexity and relative fit for the constrained hypotheses. The relative

fit of the one-sided hypotheses ($H_2 : \mu > 5$ and $H_3 : \mu < 5$) equal 0.5 (column `comp_O`), which can be explained by the fact that the implied one-sided subspaces cover half of the unconstrained space. Furthermore the posterior probability mass in the region $\mu > 5$ and $\mu < 5$ under the unconstrained model equal 0.032 and 0.968 (column `fit_O`), respectively, which quantify the relative fit of the one-sided hypotheses. The unconstrained default prior and posterior density at $\mu = 5$ equal 0.195 and 0.205 (column `comp_E` and `fit_E`), which quantify the relative complexity and fit of the precise hypothesis, respectively.

Application 2: 2-way ANOVA to investigate numerical judgement

Janiszewski and Uy (2008) executed several experiments to investigate the numerical judgments of participants. In one of the experiments (referred to as `4a') the outcome variable was the amount by which the price for a television estimated by a participant differed from an anchor price (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=roundedorprecise, respectively); and (2) whether the participants received a suggestion that the estimated price is close to the anchor value or whether they did not receive this suggestion (motivation=loworhigh, respectively). An example of a question, withanchor=roundedandmotivation=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 apreciseanchor price is obtained. By changingslightly lower' to `lower' a question with ahighmotivation is obtained. This 2×2 ANOVA design can be tested usingBFpack' as follows

```
aov1 <- aov(price ~ anchor * motivation, data = tvprices)
BF(aov1)
```

```
## Call:
## BF.lm(x = aov1)
##
## Bayesian hypothesis test
## Type: Exploratory
## Object: aov
## Parameter: group means
## Method: generalized adjusted fractional Bayes factor
##
## 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
##
## Main effects:
##            Pr(null) Pr(alt)
## anchor            0       1
## motivation        0       1
##
## Interaction effects:
##                   Pr(null) Pr(alt)
## anchor:motivation    0.251   0.749
```

For an object of class `aov`, `BFpack` also provides the Bayes factors for the existence of the main effects and interactions effects in the exploratory tests The results show clear evidence that there there is a main effect for the `anchor` factor and a main effect for the

`motivation` factor (with posterior probabilities of approximately 1). Furthermore, there is some evidence that there interaction effect between the two factors is present (with a posterior probability of 0.749). More data need to be collected in order to draw a more decisive conclusion regarding the existence of an interaction.

Application 3: Testing group variances 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): $H_1 \colon \sigma^2_C = \sigma^2_{TS} < \sigma^2_{ADHD}$ and $H_2 \colon \sigma^2_C < \sigma^2_{TS} = \sigma^2_{ADHD}$. We will test these hypotheses against the null hypothesis stating equality of variances, $H_0 \colon \sigma^2_C = \sigma^2_{TS} = \sigma^2_{ADHD}$, as well as the complement of the three aforementioned hypotheses given by $H_3 \colon \neg\,(H_0 \vee H_1 \vee H_2)$. We include the complement to safeguard against the data supporting neither of (H_0, H_1, H_2).

Silverstein et al. (1995) reported the following sample variances of the accuracies in the three groups: $s^2_C = 15.52$, $s^2_{TS} = 20.07$, and $s^2_{ADHD} = 38.81$. The data are contained in a

dataset called `attention`. In **BFpack**, we can conduct the multiple hypothesis test and weigh the evidence in favor of the four hypotheses as follows:

```
## Loaded from  bartlett_14.RData
```

```
bartlett <- bartlett_test(x = attention$accuracy, g = attention$group)
hypothesis <- c("Controls = TS < ADHD; Controls < TS = ADHD; Controls = TS = ADHD")
set.seed(358)
BF_var <- BF(bartlett, hypothesis)
```

Note that we use equal prior probabilities of the hypotheses by omitting the `prior` argument in the call to the `BF` function. The exploratory posterior probabilities for homogeneity of group variances can be obtained by running `summary(BF_var)` which yields resulting in evidence for equality of group variances. Note that the p value in the classical Bartlett test for these data equals 0.1638 which implies that the null hypothesis of homogeneity of variances cannot be rejected using common significance levels, such as 0.05 or 0.01. Note however that the this p value cannot be used as a measure for the evidence in the data in favor of homogeneity of group variances. This can be done using the proposed Bayes factor test which shows that the probability that the variances are equal is approximately 0.803. The confirmatory test provides a more detailed analysis about the most plausible relationship between the hypotheses (also obtained using the `summary()` call): Thus, H_1 receives strongest support from the 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. To conclude, the results indicate that TS population are as heterogeneous in their attentional performances as the healthy control population in this specific task, but further research would be required to obtain more conclusive evidence.

Application 4: Multivariate linear regression 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 et al., n.d.). 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.

Analysis of the complete data. In this example, two alternative hypotheses are tested. 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. (H_1) is the most parsimonious extension of these findings. It specifies that the magnitude and direction of the association between object recognition and layer thickness is not moderated by layer. To elaborate, consider a multivariate multiple regression model model with cortical thickness measures for the superficial, middle, and deep layers as three repeated (dependent) measures for each participant %[and layer (a factor with three levels) DELETE????] , 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. In other words, this hypothesis specifies that:

$$\begin{aligned} H_1 \quad : \quad & \beta_{Deep\_on\_Face} = \beta_{Middle\_on\_Face} = \beta_{Superficial\_on\_Face} < 0 < \beta_{Deep\_on\_Vehicle} \\ & = \beta_{Middle\_on\_Vehicle} = \beta_{Superficial\_on\_Vehicle}. \end{aligned}$$

That is, 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. This evidence leads to the following hypothesis:

$$H_2 \quad : \quad \beta_{Deep\_on\_Face} < \beta_{Middle\_on\_Face} = \beta_{Superficial\_on\_Face} < 0 < \beta_{Deep\_on\_Vehicle} = \beta_{Middle\_on\_Vehicle} = \beta_{Superficial\_on\_Vehicle}$$

That is, 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. One could attempt to test and compare these two hypotheses using linear mixed effects models software (e.g., the `gls` function in the `lme` package in `R`) 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`. Unfortunately, while these approaches can test some components of each hypothesis, they are not well suited to test the directional component of H_1, which specifies that all coefficients involving faces are smaller than 0 and that all coefficients involving vehicles are larger than 0. This hypothesis can, however, be tested using `BFpack` in the following way:

```
fmri.lm <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle, data = fmri)
constraints.fmri <- "Face_on_Deep = Face_on_Superficial = Face_on_Middle < 0 < Vehicle_o


set.seed(123)
BF_fmri <- BF(fmri.lm, hypothesis = constraints.fmri)
summary(BF_fmri)
```

This results in the following posterior probabilities and evidence matrix: In this analysis, hypothesis H_3 is the complement hypothesis. The evidence matrix reveals 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 evidence and the Bayesian probability of drawing the wrong conclusion after observing the data would be relatively small, namely, 0.025.

Analysis with missing observations. Here we illustrate how a Bayes factor test can be executed in the case of missing observations in the fMRI data set that are missing at random. A slightly simpler hypothesis test is considered to reduce the computation time[4]

```
constraints.fmri2 <- "Face_on_Deep = Face_on_Superficial = Face_on_Middle < 0; Face_on_D
```

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

```
fmri.lm2 <- lm(cbind(Superficial,Middle,Deep) ~ Face + Vehicle, data = fmri)
BF.fmri2 <- BF(fmri.lm2, hypothesis = constraints.fmri2)
```

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 $B_{21} = 18.443$. Now 10 missing observations (out of 65 separate observations in total) are randomly created that are missing at random:

```
fmri_missing <- fmri
set.seed(123)
for(i in 1:10){
```

[4] The hypotheses from Section 4.4 *Analysis of complete data* has constraints on the effects across different predictor variables and different dependent variables, therefore requiring Monte Carlo estimation to obtain the Bayes factors. On the other hand, when the constraints are formulated on the effects of the same predictor on different dependent variables, an analytic expression is available for the Bayes factors.

```
    fmri_missing[sample(1:nrow(fmri), 1), sample(1:ncol(fmri), 1)] <- NA
}
```

This results in 6 rows out (of the 13 rows in total) that contain at least one missing observation. Therefore listwise deletion would leave us with only 7 observations, which is almost half of the original data set. Even though listwise deletion is generally not recommended (Rubin, 1987, 1996), for this illustration we compute the Bayes factors and posterior probabilities based on this 7 by 5 data set.

```
fmri_listdel <- fmri_missing[!is.na(apply(fmri_missing, 1, sum)),]
fmri.lm2_listdel <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle, data = fmri_li
BF.fmri2 <- BF(fmri.lm2_listdel, hypothesis = constraints.fmri2)
print(BF.fmri2)
```

This results in posterior probabilities of 0.115, 0.812, and 0.073 for the two constrained hypotheses and the complement hypothesis, respectively. As expected the evidence between the hypotheses is less pronounced. Furthermore the evidence for H_2 against H_1 decreased to $B_{21} = 7.061$. Now we generate 500 imputed data sets using `mice` from the `mice` package (van Buuren et al., 2019), and use the `BF()` function 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`[5]

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

[5] 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.

```
relmeas_all <- matrix(unlist(lapply(1:M, function(m){
    fmri.lm_m <- lm(cbind(Superficial, Middle, Deep) ~ Face + Vehicle,
    data = mice::complete(mice_fmri, m))
    BF.fmri2_m <- BF(fmri.lm_m, hypothesis = constraints.fmri2)
    c(BF.fmri2_m$BFtable_confirmatory[, 1:4])
  })),ncol = M)
relmeas <- matrix(apply(relmeas_all, 1, mean),nrow = 3)
row.names(relmeas) <- c("H1", "H2", "H3")
colnames(relmeas) <- c("comp_E", "comp_O", "fit_E", "fit_O")
BF_tu_confirmatory <- relmeas[,3] * relmeas[,4] / (relmeas[,1] *
    relmeas[,2])
PHP <- BF_tu_confirmatory / sum(BF_tu_confirmatory)
BF_21_confirmatory <- BF_tu_confirmatory[2] / BF_tu_confirmatory[1]
```

This results in posterior probabilities of 0.077, 0.887, and 0.036 for the two constrained hypotheses and the complement hypothesis, respectively. As can be seen the evidence between the hypotheses is still clearer towards H_2 in comparison to the analysis after listwise deletion, but (as expected) still less pronounced in comparison to the complete data set. Furthermore, the evidence for H_2 against H_1 now equals $B_{21} = 11.552$. This illustration shows that less evidence gets lost when performing the Bayesian hypothesis test based on multiple imputed data sets than when performing the test based on the data after listwise deletion.

Application 5: Logistic regression 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/). 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 sterotypically`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 postive 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 & 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

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

Before fitting the logistic regression we reverse-coded the trustworthiness scale and standardized it to be able to compare the magnitude the three effects. We can then test these hypotheses using `BFpack` and the fitted `glm` object from `R`. Note that the fitted object also contains covariates. The full logistic regression model was first fitted, and then the above hypotheses were tested on the fitted `glm` object:

```
fit <- glm(sent ~ ztrust + zfWHR + zAfro + glasses + attract +
    maturity + tattoos, family = binomial(), data = wilson)
set.seed(123)
```

```
BF_glm <- BF(fit, hypothesis = "ztrust > (zfWHR, zAfro) > 0;
    ztrust > 0 & zfWHR = zAfro = 0")
summary(BF_glm)
```

```
## Call:
## BF.glm(x = fit, hypothesis = "ztrust > (zfWHR, zAfro) > 0;\n    ztrust > 0 & zfWHR =
##
## Bayesian hypothesis test
## Type: Exploratory
## Object: glm
## Parameter: General
## Method: Bayes factor 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
##
## Bayesian hypothesis test
## Type: Confirmatory
## Object: glm
```

```
## Parameter: General
## Method: Bayes factor using Gaussian approximations
##
## Posterior probabilities:
##    Pr(hypothesis|data)
## H1               0.078
## H2               0.006
## H3               0.916
##
## Evidence matrix:
##        H1      H2    H3
## H1  1.000  12.304 0.085
## H2  0.081   1.000 0.007
## H3 11.755 144.630 1.000
##
## Specification table:
##    comp_E comp_O fit_E fit_O  BF_E  BF_O    BF   PHP
## H1  1.000  0.036     1 0.003 1.000 0.088 0.088 0.078
## H2  0.035  0.500     0 1.000 0.004 2.000 0.007 0.006
## H3  1.000  0.964     1 0.997 1.000 1.034 1.034 0.916
##
## Hypotheses:
## H1: ztrust>(zfWHR,zAfro)>0
## H2: ztrust>0&zfWHR=zAfro=0
## H3: complement
```

In the output we see little support for the first two hypotheses; the complement receives most support: The evidence matrix shows that the complement hypothesis is around

11.589 times as likely as the second best hypothesis: Based on these results we see that the complement receives most evidence. The fact that none of the two anticipated hypotheses were supported by the data indicates that the theory is 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 ($\hat{\beta}_{afro} = -0.18071$). This unexpected result is discussed further by Wilson and Rule (2015) in their Supplementary Materials (https://journals.sagepub.com/doi/suppl/10.1177/0956797615590992).

Application 6: Testing measures of association 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 \& Firth, 1995). This data set (originally collected by Ichinose, Han, Polyn, Park and Tomarken (2019; summarized in Tomarken \& Mulder, in preparation) 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. This data set is an interesting case of how an order-constrained Bayesian approach can provide a more powerful and more appropriate test relative to alternative methods. Table \ref{tablecorr} presents the Pearson correlations for the two groups. Several features are notable: (1) Each of the 15 correlations is higher in the HC group than the SZ group; (2) On average the correlations among the HC group are rather high (on average 0.59); and, (3) The average correlation within the SZ group is

essentially 0. Despite this clear pattern, there were significant differences between the HC and SZ groups on only 2 of 15 correlations when the false discovery rate was used to control for multiple testing.

```
\begin{table}[ht] \centering
\newcommand{\mysubscript}[1]{\raisebox{-0.34ex}{\scriptsize#1}}
\renewcommand\thetable{3} \begin{tabular}{rrrrrrr}   \hline  & Im & Del &
Wmn & Cat & Fas & Rat \\    \hline   Im &  & 0.35 & -0.07 & -0.28 & -0.17 &
0.08 \\    Del & 0.83 & & -0.22 & 0.16 & 0.27 & 0.09 \\    Wmn & 0.65 & 0.50
&  & -0.05 & 0.01 & -0.02 \\    Cat & 0.56 & 0.39 & 0.77 & & 0.22 & -0.25 \\
Fas & 0.39 & 0.32 & 0.70 & 0.73  & &-0.14 \\    Rat & 0.54 & 0.47 & 0.61 &
0.77 & 0.67 & \\     \hline \end{tabular}\label{tablecorr}
\caption{Correlations for the SZ (above diagonal) and HC (below diagonal)
groups.} \end{table} Given that the overall pattern of group differences is
consistent with hypotheses, 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. This hypothesis can, however, be tested in a straightforward
manner usingBFpack‘. $H_1$ 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). H_A represents any pattern of correlations other than those that were consistent with H_1. The R syntax is as follows:

```
lm6 <- lm(cbind(Im, Del, Wmn, Cat, Fas, Rat) ~ -1 + Group,
    data = schiz)
set.seed(123)
BF6_cor <- BF(lm6,parameter = "correlation", hypothesis =
    "Del_with_Im_in_GroupHC > Del_with_Im_in_GroupSZ &
    Del_with_Wmn_in_GroupHC > Del_with_Wmn_in_GroupSZ &
    Del_with_Cat_in_GroupHC > Del_with_Cat_in_GroupSZ &
    Del_with_Fas_in_GroupHC > Del_with_Fas_in_GroupSZ &
    Del_with_Rat_in_GroupHC > Del_with_Rat_in_GroupSZ &
    Im_with_Wmn_in_GroupHC > Im_with_Wmn_in_GroupSZ &
    Im_with_Cat_in_GroupHC > Im_with_Cat_in_GroupSZ &
    Im_with_Fas_in_GroupHC > Im_with_Fas_in_GroupSZ &
    Im_with_Rat_in_GroupHC > Im_with_Rat_in_GroupSZ &
    Wmn_with_Cat_in_GroupHC > Wmn_with_Cat_in_GroupSZ &
    Wmn_with_Fas_in_GroupHC > Wmn_with_Fas_in_GroupSZ &
    Wmn_with_Rat_in_GroupHC > Wmn_with_Rat_in_GroupSZ &
    Cat_with_Fas_in_GroupHC > Cat_with_Fas_in_GroupSZ &
    Cat_with_Rat_in_GroupHC > Cat_with_Rat_in_GroupSZ &
    Fas_with_Rat_in_GroupHC > Fas_with_Rat_in_GroupSZ")
summary(BF6_cor)
```

Based on the summary, which can be obtained by running the Bayes Factor for H_1 against H_A was approximately 7888.696 and the posterior probability for H_1 was effectively 1. Thus the order-constrained analysis indicate decisive support for the researchers' hypothesis.

Application 7: Testing intraclass correlations 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), Croatia (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 homogenous 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 $(93, 112)$, Croatia $(139, 106)$, 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 followed, $1 + (p - 1)\rho$, with p the number of students in each school (balanced design), it follows 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 mathematic achievements of fourth graders of the cycles from 2003 to 2015 was used to hypothesis 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). 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.

```
library(lme4)
timssICC_subset <- timssICC[(timssICC$groupNL11 == 1) +
    (timssICC$groupHR11 == 1) + (timssICC$groupDE11 == 1) +
    (timssICC$groupDK11 == 1) > 0,]
outlme1 <- 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)
```

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. The `lmer` output object(Bates et al., 2019) was used as input in the BF function for the Bayes factor computation, where hypothesis H_1 and H_2 were added as arguments in the function call;

```
set.seed(123)
BFicc <- BF(outlme1, hypothesis =
    "groupNL11 < groupHR11 < groupDE11 < groupDK11;
    groupNL11 = groupHR11 = groupDE11 = groupDK11")
```

The output object contains the posterior mean and median estimates of the ICCs (obtained via `BFicc$estimates`), which are represented in Table **??**. 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 quite close, and the REML estimates are also located between the 2.5% and 97.5% percentile estimates.

By running `summary(BFicc)` we get the results of the exploratory and confirmatory tests. 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 & Fox, 2019). As can be seen the exploratory results indicate that a multilevel model is a appropriate for these data: Furthermore 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 posterior probabilities of the three hypotheses in the confirmatory test reveal

898 that there is approximately equal plausibility for H_2 and H_3 to be true (with posterior
899 probabilities of 0.509 and 0.471, respectively), and the complement hypothesis is unlike to be
900 true (with a posterior probability of 0.020). It can be concluded that the data gave most
901 support to an ordering of the intraclass correlations, where the Netherlands have the smallest
902 intraclass correlation and Denmark the highest. The evidence however is practically equal to
903 the evidence for the equality hypothesis. Efficient sampling strategies are needed in countries
904 with positive intraclass correlations, where countries with higher intraclass correlations will
905 benefit more from efficient stratification strategies.

906 **Application 8: Relational event models in communication networks**

In the current application, a simulated sequence of e-mail messages is analyzed based on the study of Mulder and Leenders (2019). This sample consists of 247 relational events in a network of 25 actors. It was investigated which mechanisms drive employees of a consultancy firm to send emails about innovation to each other, and to what degree. Homophily is often an important driver of relational events, which implies that individuals with similar attributes have an increased rate of interaction. Three attributes are considered in this context. Here the rate with which sender s sends receiver r an e-mail message about innovation at time t is modeled as a loglinear function of (1) whether sender s and receiver r work in the same building (1 = same building; 0 = different buildings); (2) whether sender s and receiver r work in the same division (1 = same division; 0 = different divisions); and (3) whether sender s and receiver r have the same hierarchical position (1 = same hierarchical position; 0 = different hierarchical positions). Endogenous drivers (e.g., inertia, reciprocity, or transitivity) are omitted for illustrative purposes but can be added in a straightforward manner [e.g., using Butts (2015)}. Based on Mulder and Leenders (2019), the following five hypotheses are formulated to investigate the order of strength of the effects of different

sources of similarity on e-mail interaction rates in the data:

$$H_0 : \beta_{\text{same division}} = \beta_{\text{same hierarchy}} = \beta_{\text{same building}},$$

$$H_1 : \beta_{\text{same division}} > \beta_{\text{same hierarchy}} = \beta_{\text{same building}},$$

$$H_2 : \beta_{\text{same division}} > \beta_{\text{same hierarchy}} > \beta_{\text{same building}},$$

$$H_3 : \beta_{\text{same division}} > \beta_{\text{same building}} > \beta_{\text{same hierarchy}},$$

$$H_c : \text{none of the above}$$

Here hypothesis H_0 assumes that the effects of working in the same division, having similar hierarchical position and working in the same building on the e-mail interaction rate are equal. Hypothesis H_1 until H_3 represent different expectations about the ordering of the strength of these effects. Finally, hypothesis H_c is the complement hypothesis, covering all other possible orderings of effects, thereby representing the possibility that something else is going on that was not specified To estimate this relational event model, the `R` package `relevent` (Butts, 2015) was used by running the following lines of code:

```
library(relevent)
CovEventEff <- array(NA, dim = c(3, nrow(actors), nrow(actors)))
CovEventEff[1,,] <- as.matrix(same_division)
CovEventEff[2,,] <- as.matrix(same_hierarchy)
CovEventEff[3,,] <- as.matrix(same_building)
set.seed(9227)
fit <- relevent::rem.dyad(edgelist = relevents, n = nrow(actors),
    effects = "CovEvent", ordinal = FALSE,
    covar = list(CovEvent = CovEventEff), hessian = TRUE,
    fit.method = "BPM")
```

Bayes factors and posterior model probabilities for the evaluation of these informative hypotheses can be obtained with `BFpack` by running the following lines of code:

```
names(fit$coef) <- c("division", "hierarchy", "building")
hyp <- "division = hierarchy = building;
    division > hierarchy = building;
    division > hierarchy > building;
    division > building > hierarchy"
set.seed(8389)
BF_rem <- BF(x = fit, hypothesis = hyp)
summary(BF_rem)
```

Note that names of the estimated coefficients (`fit$coef`) need to be explicitly given to simplify the formulation of the hypotheses on these parameters. The following posterior probabilities are then obtained for the hypotheses:

```
Posterior probabilities:
   Pr(hypothesis|data)
H1               0.000
H2               0.715
H3               0.063
H4               0.222
H5               0.000


Evidence matrix:
          H1    H2     H3    H4        H5
H1     1.000 0.000  0.000 0.000     2.530
H2 29280.089 1.000 11.400 3.214 74073.254
H3  2568.463 0.088  1.000 0.282  6497.739
H4  9111.318 0.311  3.547 1.000 23049.963
H5     0.395 0.000  0.000 0.000     1.000
```

```
H5: complement


Specification table:

   comp_E comp_O fit_E fit_O  BF_E  BF_O     BF   PHP
H1  0.020  1.000 0.000 1.000 0.000 1.000  0.000 0.000
H2  0.103  0.500 0.727 1.000 7.082 2.000 14.165 0.715
H3  1.000  0.088 1.000 0.109 1.000 1.243  1.243 0.063
H4  1.000  0.202 1.000 0.890 1.000 4.408  4.408 0.222
H5  1.000  0.710 1.000 0.000 1.000 0.000  0.000 0.000


Hypotheses:
H1: division=hierarchy=building
H2: division>hierarchy=building
H3: division>hierarchy>building
H4: division>building>hierarchy
H5: complement
```

The Bayes factors and posterior probabilities reveal there is most evidence for H_2 (with a posterior probability of 0.715), followed by H_4 (with a posterior probability of 0.222), followed by H_3 (with a posterior probability of 0.063). The results show that we can rule out hypothesis H_1 and the complement hypothesis. This implies that there is clear support that working in the same division has the largest effect on information sharing, followed by hierarchical similarity and working in the same building. More data need to be collected in order to draw clearer conclusions about which of these latter two effects is largest.

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`, `hetcor`, or `glm`. Furthermore by specifying a simple 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 constraints on the parameters of interest. This will allow users to move beyond traditional null hypothesis (significance) testing. In the near future the package will be extended by also including more complex statistical models such as structural equation models and generalized linear mixed models.

Acknowledgments

The first author is supported by a Vidi grant from the Netherlands Organization of Scientific Research (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 Ollson Collentine, Application 6 by Andrew Tomarken, Application 7 by Jean-Paul Fox, and Application 8 by Marlyne Meijerink.

Bartlett, M. (1957). A Comment on D. V. Lindley's Statistical Paradox. *Biometrika*, *44*, 533–534.

Bates, D., Maechler, M., Boler, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Fox, J. (2019). *lme4: Linear mixed-effects models using "eigen" and s4.* Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Berger, J. O. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, *1*, 385–402. doi:10.1214/06-BA115

Berger, J. O., & Delampady, M. (1987). Testing Precise Hypotheses. *Statistical*

Science, *2*, 317–335. doi:10.1214/ss/1177013238

Berger, J. O., & Pericchi, L. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison (with Discussion). In P. Lahiri (Ed.), *Model selection* (pp. 135–207). Hayward, CA: Institute of Mathematical Statistics.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical snalysis*. Reading, MA: Addison-Wesley.

Böing-Messing, F., & Mulder, J. (2016). Automatic bayes factors for testing variances of two independent normal distributions. *Journal of Mathematical Psychology*, *72*, 158–170. doi:10.1016/j.jmp.2015.08.001

Böing-Messing, F., & Mulder, J. (2017). Bayesian evaluation of constrained hypotheses on variances of multiple independent groups. *Psychological Methods*, *22*(2), 262–287. doi:10.1037/met0000116

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

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, C. T. (2015). *relevent: Relational event models*. Retrieved from https://cran.r-project.org/web/packages/relevent/index.html

Consonni, G., & Veronese, P. (2008). Compatibility of prior distribution across linear models. *Statistical Science*, *23*(3), 332–353. doi:10.1214/08-STS258

De Santis, F., & Spezzaferri, F. (1999). Methods for default and robust Bayesian model comparison: The fractional Bayes factor approach. *International Statistical Review*,

67(3), 267–286. doi:10.1111/j.1751-5823.1999.tb00449.x

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.

Dogge, M., Gayet, S., Custers, R., Hoijtink, H., & Aarts, H. (2019). Perception of action-outcomes is shaped by life-long and contextual expectations. *Scientific Reports*, *9*, 1–9.

Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of dutch high school students: A registered report. *Comprehensive Results in Social Psychology*. doi:10.1080/23743603.2018.1559647

Fox, J. (2016). *polycor: Polychoric and polyserial correlations*. Retrieved from https://cran.r-project.org/web/packages/polycor/index.html

Fox, J.-P., 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

Fox, J., & Weisberg, S. (2019). *car: Companion to applied regression*. Retrieved from https://cran.r-project.org/web/packages/car/index.html

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third.). Boca Raton: Taylor & Francis Group.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., ... Hothorn, T. (2016). *mvtnorm: Multivariate normal and t distributions*. Retrieved from https://cran.r-project.org/web/packages/mvtnorm/index.html

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138.

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*. doi:10.1080/00949655.2019.1590574

Gu, X., Hoijtink, H., Mulder, J., van Lissa, C. J., Jones, J., Waller, N., & The R Core Team. (2018). *bain: Bayes factors for informative hypotheses*. Retrieved from https://cran.r-project.org/web/packages/bain/index.html

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.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. New York: Chapman & Hall/CRC.

Hoijtink, H., & Chow, S.-M. (2017). Bayesian hypothesis testing: Editorial to the special issue on Bayesian data analysis. *Psychological Methods*, *22*(2), 211–216. doi:10.1037/met0000143

Hoijtink, H., Gu, X., & Mulder, J. (2018a). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical*

Psychology. doi:doi.org/10.1111/bmsp.12145

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2018b). Computing Bayes factors from data with missing values. *Psychological Methods*, *24*(2), 253–268. doi:10.1037/met0000187

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

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., & Mitchell, D. (2019). *lmtest: Testing linear regression models*. Retrieved from https://cran.r-project.org/web/packages/lmtest/index.html

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. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*(2), 203–222. doi:10.1017/S030500410001330X

Jeffreys, H. (1961). *Theory of probability-3rd ed*. New York: Oxford University Press.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189. doi:10.1016/j.jmva.2005.05.010

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions, vol. 2*. Boston, MA: Houghton Mifflin Co.

Jong, J. de, 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

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, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (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

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1), 187–192.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... Wagenmakers, E.-J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*(2). doi:10.18637/jss.v088.i02

Marin, J.-M., & Robert, C. (2010). On resolving the Savage-Dickey paradox. *Electronic Journal of Statistics*, *4*, 643–654. doi:10.1214/10-EJS564

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. doi:10.3758/s13428-010-0049-5

McGuigin, R. W., T., A., Newton, Tamber-Rosenau, B., Tomarken, A. J., & Gauthier, I. (n.d.). Thickness of deep layers in the fusiform face area predicts face recognition.

McGuigin, R. W., Van Gulick, A. E., & 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, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of bayes factors for common designs. Retrieved from

https://cran.r-project.org/web/packages/BayesFactor/index.html

Mulder, J. (2014a). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Statistical and Mathematical Psychology*, *67*(1), 153–71. doi:10.1111/bmsp.12013

Mulder, J. (2014b). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and 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., & Fox, J.-P. (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, J.-P. (2019). Bayes factor testing of multiple intraclass correlations. *Bayesian Analysis*.

Mulder, J., & Gelissen, J. (2019). Bayes factor testing of equality and order constraints on measures of association in social research. Retrieved from https://arxiv.org/abs/1807.05819

Mulder, J., Hoijtink, H., & Gu, X. (2019). Default Bayesian model selection of constrained multivariate normal linear models. Retrieved from https://arxiv.org/abs/1904.00679

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., Hoijtink, H., & Leeuw, C. de. (2012). BIEMS: A fortran 90 program for calculating Bayes factors for inequality and equality constrained model. *Journal of Statistical Software*, *46*. doi:10.18637/jss.v046.i02

Mulder, J., & Leenders, R. A. (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., & Raftery, A. E. (n.d.). BIC extensions for order-constrained model selection. *Sociological Methods & Research*.

Mulder, J., & Wagenmakers, E.-J. (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

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society B*, *57*(1), 99–138. doi:10.1111/j.2517-6161.1995.tb02017.x

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, *6*(1), 101–118. doi:10.1007/BF02564428

Pericchi, L. R., 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, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131–154). New York: Springer-Verlag.

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosa, L., Rosa, E., Sarner, L., & Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association*, *279*(13), 1005–1010. doi:10.1001/jama.279.13.1005

Rouder, J. N., Speckman, P. L., D. Sun, R. D. M., & 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

Rousseeuw, P. J., & Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, *48*(4), 276–279. doi:10.2307/2684832

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi:10.2307/2291635

Russell, V. A., Oades, R. D., Tannock, R., Killeen, P. R., Auerbach, J. G., Johansen, E. B., & 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, B. R., & Herring, A. H. (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

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi:10.1037/met0000061

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71. doi:10.1198/000313001300339950

Silverstein, S. M., Como, P. G., Palumbo, D. R., West, L. L., & Osborn, L. M. (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. Retrieved from https://psyarxiv.com/4xqvr/

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., ... Gray, B. (2019). *mice: Multivariate imputation by chained equations*. Retrieved from https://cran.r-project.org/web/packages/mice/index.html

Van de Schoot, R., Hoijtink, H., Mulder, J., Aken, M. V., de Castro, B. O., Meeus, W., & Romeijn, J.-W. (2006). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology*, *47*, 203–212. doi:10.1037/a0020957

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*. doi:10.1186/s12874-019-0699-7

van Schie, K., Van Veen, S., Engelhard, I., Klugkist, I., & Van den Hout, M. (2016). Blurring emotional memories using eye movements: Individual differences and speed of eye movements. *European Journal of Psychotraumatology*, *7*.

Verdinelli, I., & Wasserman, L. (1995). Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of American Statistical Association*, *90*(430), 614–618. doi:10.2307/2291073

Vrinten, C., Gu, X., S. Weinreich, Schipper, M., Wessels, J., Ferrari, M., ... Verschuuren, J. (2016). An n-of-one rct for intravenous immunoglobulin g for inflammation in hereditary neuropathy with liability to pressure palsy (hnpp). *Journal of Neurology, Neurosurgery and Psychiatry*, *87*(7), 790–791. doi:10.1136/jnnp-2014-309427

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (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

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2017). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790

Well, S. V., Kolk, A. M., & 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*, 427–449.

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, R. P. P. P., & Wagenmakers, E. -J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis*, *38*, 666–690. doi:10.1.1.149.885

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*, 1325–1331. doi:10.1177/0956797615590992

Zondervan-Zwijnenburg, M. A., Veldkamp, S. A., Neumann, A., Barzeva, S. A., Nelemans, S. A., van Beijsterveldt, C. E., ... Boomsma, A. J. O. D. I. (2019). Parental age and offspring childhood mental health: A multi-cohort, population-based investigation. *Child Development*. doi:10.1111/cdev.13267

Table 1

X.R..function	package	test	
‘t_test‘	‘bain‘	Student t test	mean (1-sample test) mean diffe
‘var_test‘	‘BFpack‘	heterogeneity of variances	group variances
‘aov‘	‘stats‘	AN(C)OVA	group means
‘manova‘	‘stats‘	MAN(C)OVA	group means
‘lm‘	‘stats‘	linear regression multivariate regression	regression coefficients regression
‘lmer‘	‘lme4‘	random intercept model	group specific intraclass correlat
‘hetcor‘	‘polycor‘	correlation analysis	measures of association
‘glm‘	‘stats‘	generalized linear model	regression coefficients
‘coxph‘, ‘survreg‘	‘survival‘	survival analysis	regression coefficients
‘rem‘, ‘rem.dyad‘	‘relevent‘	relational event model	regression coefficients
‘polr‘	‘MASS‘	ordinal regression	regression coefficients
‘zeroinfl‘	‘pscl‘	zero-inflated regression models	regression coefficients

Note. R functions, packages, descriptions of tests, parameter of interest, and example name of the paramete
outcome variable, ’x1’ is the label of a numeric predictor variable, and ’g1’ is the label of a level of a grou

Table 2

Example hypothesis tests that can be executed using 'BFpack'.

X	Example.hypothes
Exploratory testing	$H_0:\theta=0$ vs $H_1:\theta < 0$ vs $H_2:\theta > 0$.
Interval testing	$H_0:\|\theta\|\le \epsilon$ vs $H_1:\|\theta\| > \epsilon$, for given $\epsilo
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
Equality and order testing	$H_1:\theta_{12}<\theta_{13}=\theta_{14}$ versus $H_2:$ "not $H_1