PSYCHOMETRIKA
https://doi.org/10.1007/s11336-018-9615-z

AUTOMATIC BAYES FACTORS FOR TESTING EQUALITY- AND INEQUALITY-CONSTRAINED HYPOTHESES ON VARIANCES

FLORIAN BÖING-MESSING

JHERONIMUS ACADEMY OF DATA SCIENCE

TILBURG UNIVERSITY

JORIS MULDER

TILBURG UNIVERSITY

In comparing characteristics of independent populations, researchers frequently expect a certain structure of the population variances. These expectations can be formulated as hypotheses with equality and/or inequality constraints on the variances. In this article, we consider the Bayes factor for testing such (in)equality-constrained hypotheses on variances. Application of Bayes factors requires specification of a prior under every hypothesis to be tested. However, specifying subjective priors for variances based on prior information is a difficult task. We therefore consider so-called automatic or default Bayes factors. These methods avoid the need for the user to specify priors by using information from the sample data. We present three automatic Bayes factors for testing variances. The first is a Bayes factor with equal priors on all variances, where the priors are specified automatically using a small share of the information in the sample data. The second is the fractional Bayes factor, where a fraction of the likelihood is used for automatic prior specification. The third is an adjustment of the fractional Bayes factor such that the parsimony of inequality-constrained hypotheses is properly taken into account. The Bayes factors are evaluated by investigating different properties such as information consistency and large sample consistency. Based on this evaluation, it is concluded that the adjusted fractional Bayes factor is generally recommendable for testing equality- and inequality-constrained hypotheses on variances.

Key words: default Bayes factor, fractional Bayes factor, heterogeneity, heteroscedasticity, homogeneity of variance, inequality constraint.

1. Introduction

In comparing multiple independent populations, applied researchers commonly focus on the population means, while treating the population variances as nuisance parameters. However, by disregarding the variances one runs the risk of overlooking crucial information in the data about the differences in the populations. In fact, there are often reasons to expect certain relations between the variances of independent populations. For example, males have frequently been found to be more variable than females on a variety of measures, which has been attributed to genetic as well as social factors (e.g., Lehre, Lehre, Laake, & Danbolt, 2009). Arden and Plomin (2006) therefore expected boys to be more heterogeneous in their intelligence than girls. This expectation can be formalized in the inequality-constrained hypothesis $H_1 : \sigma_1^2 < \sigma_2^2$, where σ_1^2 and σ_2^2 denote the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-018-9615-z) contains supplementary material, which is available to authorized users.

This research was partly supported by a Rubicon grant which was awarded to Joris Mulder by The Netherlands Organisation for Scientific Research (NWO).

Correspondence should be made to Florian Böing-Messing, Jheronimus Academy of Data Science, Sint Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands. Email: florian.boeingmessing@gmail.com

© 2018 The Psychometric Society

Published online: 03 May 2018

population variance of girls and boys, respectively. Potential competing hypotheses would be $H_0\colon \sigma_1^2 = \sigma_2^2$ and $H_2\colon \sigma_2^2 < \sigma_1^2$, that is, there is no difference in heterogeneity and girls are more heterogeneous than boys, respectively. In another study, Aunola, Leskinen, Lerkkanen, and Nurmi (2004) expected that the variance of students' mathematics abilities either increases or decreases across grades. These expectations can be translated into the two competing hypotheses $H_1\colon \sigma_1^2 < \cdots < \sigma_J^2$ and $H_2\colon \sigma_J^2 < \cdots < \sigma_1^2$, where σ_j^2 denotes the population variance in grade j and J is the number of grades to be compared. The reasoning behind an increase in variances is that skilled students improve their mathematics abilities over time more than students who are less skilled, which increases interindividual differences. Alternatively, systematic instruction at school might help less skilled students catch up, which would result in a decrease in variances over time.

While the examples in the previous paragraph deal with existing groups, constrained hypotheses on variances are conceivable in experimental studies as well. For example, one may expect variances in treatment groups to be larger than the variance in a control group because subjects may react differently to a certain treatment (e.g., Grissom, 2000). This suggests testing a hypothesis of the form $H_1\colon \sigma_1^2 < \sigma_2^2 = \sigma_3^2$, where σ_1^2 denotes the variance in the control group and σ_2^2 and σ_3^2 denote the variance in treatment groups 1 and 2, respectively. We could test H_1 against $H_2\colon \sigma_1^2 < \left(\sigma_2^2, \sigma_3^2\right)$ to determine whether there is evidence in favor of equal treatment group variances in addition to the assumption that both variances are greater than the variance in the control group. The comma symbol in H_2 indicates that there is no constraint on the relation between σ_2^2 and σ_3^2. Another potential competing hypothesis would be the null hypothesis $H_0\colon \sigma_1^2 = \sigma_2^2 = \sigma_3^2$. In case there is just one treatment that is administered in two different intensities, one may expect that an intense treatment results in a larger variance than a mild treatment. This suggests testing the order-constrained hypothesis $H_3\colon \sigma_1^2 < \sigma_2^2 < \sigma_3^2$, where group 2 undergoes a mild treatment and group 3 an intense treatment. In the above examples, testing constrained hypotheses on the group variances in addition to testing group means gives a more complete picture of the relationships between the groups.

In this article, we consider the problem of testing $T \geq 2$ hypotheses on the variances of $J \geq 2$ independent populations. The hypotheses are of the form

$$H_t\colon \boldsymbol{R}_t^E \boldsymbol{\sigma}^2 = \mathbf{0} \wedge \boldsymbol{R}_t^I \boldsymbol{\sigma}^2 > \mathbf{0}, \quad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{\sigma}^2 = \left[\sigma_1^2 \cdots \sigma_J^2\right]^T$ is a J-dimensional vector containing the population variances. Let q_t^E and q_t^I denote the number of equality and inequality constraints on the variances in $\boldsymbol{\sigma}^2$ under H_t, respectively. Then, $\boldsymbol{R}_t^E$ ($\boldsymbol{R}_t^I$) is a $q_t^E \times J$ ($q_t^I \times J$) matrix containing the coefficients for the equality (inequality) constraints on the variances under H_t and $\mathbf{0} = \left[0 \cdots 0\right]^T$ is a q_t^E-dimensional (q_t^I-dimensional) vector of zeroes. We consider tests where each row of $\boldsymbol{R}_t^E$ and $\boldsymbol{R}_t^I$ is a permutation of $\{-1, 1, 0, \ldots, 0\}$. Thus, we test constraints with equal coefficients for the variances (e.g., $\sigma_1^2 < \sigma_2^2$), but not complex mathematical constraints such as $2\sigma_1^2 < \sigma_2^2$, $\sigma_1^2 + \sigma_2^2 < \sigma_3^2$, or $\sigma_1^2/\sigma_2^2 < \sigma_3^2/\sigma_4^2$. Note that the formulation in Eq. (1) includes the classical null and alternative hypothesis as special cases.

The multiple hypothesis test in Eq. (1) is much more general than the standard test of a null hypothesis where all variances are equal against an alternative where the variances are unrestricted. Besides Böing-Messing, van Assen, Hofman, Hoijtink, and Mulder (2017), the testing framework in Eq. (1) has not yet been considered in the literature for variance components. This is quite surprising given the central role of variance components in the statistical sciences (see also Carroll, 2003). The test is particularly useful in a confirmatory setting when one has expectations about possible patterns of the population variances. For example, when it is expected that the heterogeneity across patients increases when the intensity of the treatment increases, say,

$H_1: \sigma_1^2 < \sigma_2^2 < \sigma_3^2$, and the hypothesis is against the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, testing these hypotheses directly against each other is the preferred method over an omnibus test with additional post hoc testing. When the omnibus test results in a rejection of the null hypothesis, post hoc testing can result in low power (due to Type I error corrections, e.g., Bonferroni) or conflicting conclusions, for example, $\sigma_1^2 = \sigma_2^2$ is not rejected, $\sigma_1^2 = \sigma_3^2$ is not rejected, but $\sigma_2^2 = \sigma_3^2$ is rejected. Directly testing the order-constrained hypothesis against the null hypothesis avoids such issues and gives a direct answer to the research question.

In this article, we consider the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995) for the testing problem formulated in Eq. (1). The Bayes factor is a Bayesian hypothesis testing criterion that is becoming increasingly popular. It has a number of advantages over alternative approaches to hypothesis testing like null hypothesis significance testing by means of p values and hypothesis testing by means of information criteria like the AIC (Akaike, 1973) and the BIC (Schwarz, 1978): First, unlike p values, Bayes factors are able to quantify the evidence in the data in favor of a hypothesis (including null hypotheses) relative to another hypothesis (Berger & Sellke 1987; Wagenmakers, 2007). Second, using Bayes factors it is straightforward to simultaneously test multiple (non-)nested hypotheses (Berger & Mortera, 1999). This property is not shared by p values either. Third, Bayes factors are consistent in the sense that they converge to the true hypothesis as the sample size increases. This also holds for a true null hypothesis. The AIC, on the other hand, is not consistent (e.g., O'Hagan, 1995), and p values are only consistent if the null hypothesis is false. Fourth, Bayes factors function as an Occam's razor by automatically taking the parsimony of (in)equality-constrained hypotheses as in Eq. (1) into account. By contrast, p values have no inherent mode of taking the parsimony of a hypothesis into account. The AIC and the BIC are able to incorporate the parsimony introduced by equality constraints, but not inequality constraints (Mulder et al., 2009). Consequently, they do not provide a solution to the testing problem in Eq. (1).

Application of Bayes factors requires the specification of a prior distribution under every hypothesis to be tested. Often, however, prior information about the parameters is not available or a researcher would like to refrain from adding prior knowledge. But even when prior information is available, it is a difficult and time-consuming task to translate this into mathematical functions such as prior distributions (e.g., Berger, 2006). Researchers therefore developed so-called automatic or default Bayes factors. These methods enable the computation of Bayes factors without having the user specify proper subjective priors. Automatic Bayes factors have been developed for various testing problems frequently encountered in practice. Klugkist, Laudy, and Hoijtink (2005) developed an automatic Bayes factor for testing inequality-constrained hypotheses on means in ANOVA models. Mulder, Hoijtink, and Klugkist (2009) presented Bayes factors for testing means in repeated measures situations. Mulder, Hoijtink, and de Leeuw (2012) developed Bayes factors for testing (in)equality constraints on means and regression coefficients in multivariate normal linear regression models. Mulder and Fox (2013) considered Bayes factor tests of multiple intraclass correlations. Mulder (2016) applied the Bayes factor to the problem of testing order-constrained hypotheses on correlations. Recently, Böing-Messing and Mulder (2016) and Böing-Messing et al. (2017) considered the Bayes factor for testing constrained hypotheses on variances.

In this article, we will present three different automatic Bayes factors for testing (in)equality-constrained hypotheses on variances as in Eq. (1): a balanced Bayes factor, a generalized fractional Bayes factor, and an adjusted fractional Bayes factor. The first two methods are novel developments that have not yet been considered for the testing problem in Eq. (1). The adjusted fractional Bayes factor was proposed by Böing-Messing et al. (2017). The main idea of the three methods is to use a small share of the information in the sample data to automatically specify proper priors. Subsequently, the remaining share is used for hypothesis testing. This methodology avoids the need for the user to specify proper subjective priors based on prior information. As will be shown,

the three Bayes factors use different methods for splitting the information for automatic prior specification and hypothesis testing.

The Bayes factors will be evaluated based on six important criteria. First, it will be checked whether the automatic prior used for the computation of the Bayes factors contains the information of a minimal experiment. This is important to avoid problems such as Bartlett's paradox (Bartlett, 1957). Second, it will be investigated whether the Bayes factors are scale invariant. This is crucial because the hypothesis test should not depend on the scale of the data. Third, it will be checked whether the Bayes factors function as an Occam's razor when testing inequality-constrained hypotheses. This is not always the case with automatic Bayes factors (e.g., Mulder, Hoijtink, & Klugkist, 2010). Fourth, it will be examined whether the Bayes factors are information consistent. Information consistency implies that in the case of overwhelming evidence toward a particular hypothesis (based on the unrestricted estimates), the Bayes factor toward this hypothesis goes to infinity (Liang, Paulo, Molina, Clyde, & Berger, 2008). Fifth, large sample consistency will be investigated. In particular, through numerical simulation we will show how fast the evidence toward the true hypothesis accumulates. Sixth, we examine the Bayes factors' robustness to non-normality of the data (since the Bayes factors use the normal distribution to model the data). The contributions of this article are (i) the presentation of two new automatic Bayes factors for the testing problem in Eq. (1), (ii) the evaluation of all Bayes factors based on the six criteria above, and (iii) the application of the methodology to three different motivating examples that highlight the importance of testing (in)equality-constrained hypotheses on variances in practice.

The remainder of this article is structured as follows. In the next section, we introduce the three motivating examples. Following this, we give a brief introduction to the Bayes factor for testing variances. We then give a detailed discussion of the six criteria for evaluating the Bayes factors. Next, we develop the three automatic Bayes factors for testing (in)equality-constrained hypotheses on variances. After that, we evaluate the Bayes factors by checking whether they satisfy the six criteria. Subsequently, we illustrate the practical utility of the Bayes factors by applying them to actual data from the three motivating examples. We conclude the article with a discussion of our approach to testing variances.

2. Motivating Examples

In this section, we introduce three examples we use to highlight the practical relevance of testing (in)equality-constrained hypotheses on variances of independent populations. We selected examples where previous research or theoretical considerations clearly suggested certain (in)equality-constrained hypotheses on the variances. At a later stage, we will apply the Bayes factors to be developed in this article to actual data from the three example studies in order to demonstrate the Bayes factors' usefulness for analyzing real data in practice.

The first example we consider is a hypothetical study with four treatment groups from Weerahandi (1995). The author reports an increasing pattern of sample variances across the four groups. In practice such an increasing pattern of sample variances could emerge, for example, if the groups receive a new drug in an increasing dosage. Patients may respond quite differently to a new drug, especially if the dosage is high. As a result, the variance is larger in groups receiving higher dosages. Alternatively, an increasing pattern of variances might be observed for independent groups that receive the same treatment for time periods of increasing length (e.g., Aunola et al., 2004). Here, it is expected that subjects respond more heterogeneously when they received the treatment for a longer time. Such expectations can be formulated as an inequality-constrained hypothesis of the form $H_1 : \sigma_1^2 < \cdots < \sigma_4^2$. To determine the evidence in the data in favor of H_1 we need one or more competing hypotheses H_1 can be tested against. Two competing hypotheses that are often important in practice are the null hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_4^2$ stating equality of

variances and the complement $H_2: \neg(H_0 \vee H_1)$, which comprises all possible hypotheses except H_0 and H_1.

The second example we consider is a study by Silverstein, Como, Palumbo, West, and Osborn (1995), who compared attentional performances of 17 Tourette's and 17 ADHD patients with those of a group of 17 controls. Participants were shown 120 strings of 12 letters. Each string contained either a T or an F at a random position; the remaining 11 letters were random letters other than T and F. Each string was presented for 55 ms. After each presentation, participants had to indicate as quickly as possible whether the string contained a T or an F. After completion of the 120 strings, the accuracy of the respondents was computed as the percentage of correct answers. Now, psychological research has frequently found ADHD patients to be more heterogeneous in their attentional performances than unaffected controls (see, e.g., Kofler et al., 2013; Russell et al., 2006). The heterogeneity of attentional performances of Tourette's patients as compared to unaffected controls is less well documented. Given this information, we will test the following hypotheses on the variances of the accuracies to investigate whether there is evidence that Tourette's patients (group 2) are as heterogeneous in their attentional performances as either unaffected controls (group 1) or ADHD patients (group 3): $H_1: \sigma_1^2 = \sigma_2^2 < \sigma_3^2$ and $H_2: \sigma_1^2 < \sigma_2^2 = \sigma_3^2$. We will compare H_1 and H_2 to the competing hypotheses $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and $H_3: \neg(H_0 \vee H_1 \vee H_2)$.

Contrary to the studies above, our third example is a study involving a design with two factors. Lucas (2003) investigated the influence of group leaders on subordinate group members. The author was interested in whether a leader's influence depends on the leader's gender and the way the leader was appointed. The author considered two types of appointment: Either the leader was chosen at random or based on ability. Lucas conducted a 2×2 factorial experiment with 30 participants in each condition. Influence of the group leader was measured as the number of times (in 10 trials) that a participant changed his/her opinion to match the group leader's opinion. Our interest is in the variability of the counts in the four groups. Research on gender differences suggests that the variability is greater for male leaders than for female leaders (e.g., Lehre et al., 2009). Due to a lack of theoretical underpinning, we assume that there is no effect of appointment type. These expectations correspond to the hypothesis $H_1: \sigma_2^2 = \sigma_4^2 < \sigma_1^2 = \sigma_3^2$, where $\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\}$ are the variances of the groups whose leader is {(male, random), (female, random), (male, based on ability), (female, based on ability)}. We will test H_1 against the competing hypotheses $H_0: \sigma_1^2 = \cdots = \sigma_4^2$ and $H_2: \neg(H_0 \vee H_1)$. This example illustrates that (in)equality-constrained hypotheses on variances can be formulated not only within one factor, but also across multiple factors.

3. The Bayes Factor for Testing Variances

In this article, we assume that the data $\boldsymbol{x}_j = \left[x_{1j} \cdots x_{n_j j}\right]^T$ come from a normal population with mean μ_j and variance σ_j^2:

$$x_{ij} \overset{\text{i.i.d.}}{\sim} N\left(\mu_j, \sigma_j^2\right), \quad i = 1, \ldots, n_j, \;\; j = 1, \ldots, J. \tag{2}$$

Furthermore, we denote $\boldsymbol{x} = \left[\boldsymbol{x}_1^T \cdots \boldsymbol{x}_J^T\right]^T$.

Before we give the expression for the marginal likelihood under an (in)equality-constrained hypothesis H_t, the key ingredient of the Bayes factor, we need to introduce some additional notation. Under a hypothesis H_t with q_t^E equality constraints and q_t^I inequality constraints on J population variances, there are $K_t = J - q_t^E$ unique variances which we denote by $\boldsymbol{\sigma}_t^2 = \left[\sigma_1^2 \cdots \sigma_{K_t}^2\right]^T$ (note that we omitted the hypothesis index t on the individual variances to simplify the notation). Furthermore, let J_k be the number of populations sharing the unique variance σ_k^2 and

let μ_{k_j} denote the mean of the jth population sharing the unique variance σ_k^2, for $j = 1, \ldots, J_k$ and $k = 1, \ldots, K_t$. Similarly, let $\boldsymbol{x}_{k_j} = \left[x_{1jk} \cdots x_{n_{k_j}jk}\right]^T$ be the vector of n_{k_j} observations from the jth population sharing the unique variance σ_k^2. If there are no equality constraints under H_t, we omit the subscript j and write μ_k, $\boldsymbol{x}_k$, and n_k instead of μ_{k_1}, $\boldsymbol{x}_{k_1}$, and n_{k_1} to simplify the notation. In a similar manner, under the null hypothesis where there is just 1 unique variance, we omit the subscript k and write σ^2, μ_j, $\boldsymbol{x}_j$, and n_j instead of σ_1^2, μ_{1_j}, $\boldsymbol{x}_{1_j}$, and n_{1_j}. Finally, we denote the admissible parameter space of the unique variances under H_t by Ω_t and the vector of the unconstrained population means by $\boldsymbol{\mu}$.

We illustrate the above notation by means of the hypothesis $H_1 : \sigma_1^2 = \sigma_2^2 < \sigma_3^2$ on the variances of $J = 3$ populations. Under H_1 there is $q_1^E = 1$ equality constraint and $q_1^I = 1$ inequality constraint, resulting in $K_1 = 3 - 1 = 2$ unique variances denoted by $\boldsymbol{\sigma}_1^2 = \left[\sigma_1^2\ \sigma_2^2\right]^T$. Population 1 and 2, which have equal variances under H_1, share the unique variance σ_1^2 and population 3 has the unique variance σ_2^2. Consequently, the number of populations sharing the unique variances σ_1^2 and σ_2^2 is given by $J_1 = 2$ and $J_2 = 1$, respectively. Furthermore, $\boldsymbol{\mu} = \left[\mu_{1_1}\ \mu_{1_2}\ \mu_{2_1}\right]^T$ is the vector of the means of populations 1, 2, and 3. Similarly, $\boldsymbol{x} = \left[\boldsymbol{x}_{1_1}^T\ \boldsymbol{x}_{1_2}^T\ \boldsymbol{x}_{2_1}^T\right]^T$ is the vector of the data from populations 1, 2, and 3 with sample sizes of n_{1_1}, n_{1_2}, and n_{2_1}, respectively. Finally, the admissible parameter space of the unique variances under H_1 is given by $\Omega_1 = \left\{\boldsymbol{\sigma}_1^2 : \sigma_1^2 < \sigma_2^2\right\}$.

The Bayes factor is defined as the ratio of the marginal likelihoods under two competing hypotheses H_t and $H_{t'}$:

$$B_{tt'} = \frac{m_t(\boldsymbol{x})}{m_{t'}(\boldsymbol{x})}, \tag{3}$$

where $m_t(\boldsymbol{x})$ denotes the marginal likelihood under an (in)equality-constrained hypothesis H_t formulated according to Eq. (1), which is given by

$$m_t(\boldsymbol{x}) = \int_{\Omega_t} \int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \pi_t\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2. \tag{4}$$

The expression $f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)$ is the likelihood under H_t, which is given by

$$f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) = \prod_{k=1}^{K_t} \prod_{j=1}^{J_k} f\left(\boldsymbol{x}_{k_j}|\mu_{k_j}, \sigma_k^2\right) \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right) = \prod_{k=1}^{K_t} \prod_{j=1}^{J_k} \prod_{i=1}^{n_{k_j}} N\left(x_{ijk}|\mu_{k_j}, \sigma_k^2\right) \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right), \tag{5}$$

where $\mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right)$ is the indicator function which is 1 if $\boldsymbol{\sigma}_t^2 \in \Omega_t$ and 0 otherwise. The second component of the marginal likelihood in Eq. (4) is $\pi_t\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)$, the prior distribution of the model parameters under H_t. The prior contains the information about the model parameters before observing the data.

The marginal likelihood m_t quantifies how well an (in)equality-constrained hypothesis H_t with prior π_t was able to predict the observed data (Jeffreys, 1961). Consequently, the Bayes factor $B_{tt'}$ quantifies how much better H_t is able to predict the data as compared to $H_{t'}$. The Bayes factor $B_{tt'}$ can thus be interpreted as a measure of relative evidence in the data in favor of H_t relative to $H_{t'}$. Because of this intuitive interpretation, Bayes factors are becoming increasingly popular for testing scientific theories in psychological research (see the special issue on this topic in Mulder & Wagenmakers, 2016).

In addition, one may compute the posterior probabilities of the hypotheses under investigation using the marginal likelihoods and the prior probabilities of the hypotheses $P(H_1), \ldots, P(H_T)$. The prior probabilities quantify the likelihood of the hypotheses before observing any data. A

widely accepted default choice is to set equal prior probabilities $P(H_1) = \cdots = P(H_T) = 1/T$ (e.g., Berger & Mortera, 1999; Hoijtink, 2011; Mulder, Hoijtink, & de Leeuw, 2012). After observing the data, the posterior probabilities of the hypotheses are obtained by updating the prior probabilities with the marginal likelihoods according to

$$P(H_t|\boldsymbol{x}) = \frac{m_t(\boldsymbol{x})P(H_t)}{\sum_{t'=1}^{T} m_{t'}(\boldsymbol{x})P(H_{t'})}, \tag{6}$$

for $t = 1, \ldots, T$. The resulting posterior probabilities $P(H_1|\boldsymbol{x}), \ldots, P(H_T|\boldsymbol{x})$ quantify the plausibility of the hypotheses after observing the data.

4. Desirable Properties of the Bayes Factor When Testing (In)equality-Constrained Hypotheses on Variances

For the testing problem of multiple (in)equality-constrained hypotheses on variances, we deem the following properties to be of vital importance for the Bayes factor:

1. *Minimal information prior:* The choice of the prior is a key aspect when quantifying the relative evidence between hypotheses using the Bayes factor. Generally, it is recommended that the prior should be neither too informative nor too vague. When the prior is too informative, it might dominate the data. On the other hand, when the prior is specified arbitrarily vague, the evidence toward an equality-constrained null hypothesis can become arbitrarily large regardless of the observed data. This can be explained by the fact that unrealistically large effects are anticipated under the unconstrained alternative due to the extremely vague prior. This is also known as Bartlett's paradox (Bartlett, 1957). For this reason, a prior containing the information of a minimal experiment (discussed later) is generally recommended.
2. *Scale invariance:* It is crucial that a Bayes factor is invariant to the scale of the data. Thus, when the outcome variable is rescaled from, say, a 0–10 scale to a 0–100 scale, the heterogeneity of the group measurements does not change in a relative manner, and therefore, the relative evidence between hypotheses on variances, as quantified by the Bayes factor, should also remain unchanged.
3. *Occam's razor when testing inequality-constrained hypotheses:* Bayes factors naturally balance between fit and complexity as an Occam's razor when testing hypotheses with equality constraints (Jefferys & Berger, 1992). When testing hypotheses with inequality constraints, on the other hand, this is not always the case (e.g., Mulder, 2014a). To evaluate this property when testing inequality-constrained hypotheses on variances, we will consider a test of the order-constrained hypothesis $H_1 : \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ against the larger (more complex) unconstrained hypothesis $H_u : \sigma_1^2, \sigma_2^2, \sigma_3^2$. The Occam's razor property implies that in the case of overwhelming evidence for the order constraints, hypothesis H_1 should be preferred over the larger alternative hypothesis H_u.
4. *Information consistency:* A Bayes factor for an unconstrained hypothesis against the null hypothesis is called information consistent if it goes to infinity as the effect size goes to infinity, while keeping the sample size fixed. The Bayes factor is called information inconsistent if it converges to a constant in the limit. A well-known example of information inconsistency is the Bayes factor based on Zellner's g-prior (Berger & Pericchi, 2001; Zellner, 1986). Information (in)consistency when testing inequality-constrained hypotheses was first considered by Mulder (2014a) for testing means. To our knowledge, information (in)consistency has never been investigated when testing variances.

Information consistency is an important property because it ensures that the conclusion based on the Bayes factor corresponds with the conclusion by looking at the effects in the data.

5. *Large sample consistency:* A Bayes factor is called large sample consistent when the evidence for the true hypothesis against the competing hypotheses goes to infinity as the sample size goes to infinity. Large sample consistency is a crucial statistical property as it ensures that we will always select the true hypothesis as long as we collect enough data. Note that the classical p value in a null hypothesis significance test is not consistent because there is still a probability of incorrectly rejecting a true null equal to the prespecified significance level, even when we have extremely large samples.
6. *Robustness to non-normality:* The Bayes factors we present in this article are based on the assumption that the outcome of interest is normally distributed in the populations under study. Empirical data, however, may deviate in certain aspects from normality (e.g., outliers, skew). A Bayes factor should be robust to such violations of normality in the data.

5. Automatic Bayes Factors

In Sect. 3, we saw that in order to quantify the relative evidence in the data between the hypotheses of interest one needs to specify a proper prior for the free parameters under each hypothesis. However, specifying priors for the population variances under all hypotheses to be tested is a difficult and time-consuming task. First, one must elicit how plausible the values of the unique variances under each hypothesis are before observing the data. In the case of mean effects, researchers generally have an idea how plausible it is to observe a small, medium, or large effect in the case that the null is false. In the case of variance parameters, this is much more difficult because people generally have less intuition when considering variances than when considering means. Even though it is extremely difficult, suppose that it is possible to elicit prior information about the magnitude of each unique variance under the hypotheses. The next challenge is to translate this information into a prior distribution. Prior distributions can have endlessly many possible shapes (left/right skewed, little/much kurtosis, etc.). Therefore, it is practically impossible to derive a prior distribution that exactly matches one's prior beliefs about the magnitude of the variances.

Because of these difficulties, one might be tempted to use non-informative priors instead. The standard non-informative prior for a variance parameter is the Jeffreys prior, which is σ^{-2}. This prior assumes that all possible values of the variance are equally likely on a log scale. Hence, this prior does not integrate to one, and therefore, the prior is called improper. These improper priors depend on undefined constants. Although improper priors can be used in Bayesian estimation (the undefined constants cancel out in the posterior), improper priors cannot be used in Bayes factor testing because the resulting Bayes factors will depend on these undefined constants (for details see, e.g., O'Hagan, 1995). Another potential option could be to work with very vague proper priors. This would avoid the issue of undefined constants of improper priors and still allows us to compute the Bayes factor without needing subjective prior beliefs. This is also a bad idea, however, as it will result in Bartlett's paradox (Bartlett, 1957), as noted earlier. For this reason, we focus on automatic Bayes factors which can be computed in an automatic fashion without needing subjective information about the magnitude of the effects under each hypothesis. Three different types of Bayes factors will be presented below, each in a separate subsection.

5.1. Balanced Bayes Factor

The balanced Bayes factor (BBF) was introduced by Böing-Messing and Mulder (2016) as an automatic Bayes factor for testing (in)equality-constrained hypotheses on $J = 2$ variances. In

this section, we generalize this approach to the case of testing hypotheses on $J \geq 2$ variances formulated according to Eq. (1). The main idea of the BBF is to use information from the sample data to construct a proper prior in an automatic fashion such that it is balanced. We use the term "balanced" following Jeffreys (1961), who referred to an unconstrained prior for an effect as balanced if the prior probability of a positive effect is equal to the prior probability of a negative effect. The automatic prior in the BBF is based on a similar idea, namely that every possible ordering of the population variances is equally likely a priori (similar as in Mulder, Hoijtink, and Klugkist (2010) for population means and regression coefficients). This balanced prior for the population variances contains minimal information and has a scale hyperparameter that is automatically determined by the sample data to avoid the need for subjective prior information. To obtain this balanced prior, we proceed as follows. First, we fit a null model with a common variance to a small part of the sample data. The latter is obtained by taking a small fraction of the likelihood as suggested by O'Hagan (1995) in his fractional Bayes factor methodology. Next, we obtain the marginal posterior of the common variance based on this small fraction of the likelihood. We choose the fraction of the likelihood such that this marginal posterior contains minimal information. (Details will be discussed below.) Finally, this posterior is used as prior for each unique variance under the constrained hypotheses. Note that under this prior different orderings of the variances are equally likely because every unique variance has the same prior.

The technical details of our approach to constructing the automatic balanced prior in the BBF are as follows. First, we assume $H_0\colon \sigma_1^2 = \ldots = \sigma_J^2 = \sigma^2$. We then obtain a proper posterior by updating the non-informative Jeffreys prior on $\boldsymbol{\mu}$ and σ^2 with a fraction of the likelihood under H_0:

$$\pi_0^{\mathrm{B}}\left(\boldsymbol{\mu}, \sigma^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \propto \left(\prod_{j=1}^{J} f\left(\boldsymbol{x}_j | \mu_j, \sigma^2\right)^{b_j}\right) \pi_0^{N}\left(\boldsymbol{\mu}, \sigma^2\right), \tag{7}$$

where $\pi_0^N\left(\boldsymbol{\mu}, \sigma^2\right) \propto \sigma^{-2}$ is the Jeffreys prior under H_0, and we use the superscript B to refer to the BBF. The expression $f\left(\boldsymbol{x}_j | \mu_j, \sigma^2\right)^{b_j}$ denotes a fraction of the likelihood of the data from population j under H_0 (inspired by the fractional Bayes factor of O'Hagan, 1995). It is obtained by raising the likelihood of population j to the power of $b_j \in [0, 1]$. The exponent b_j is a population-specific fraction that controls how much information (in terms of the number of observations) is contained in the fraction of the likelihood of population j (Berger & Pericchi, 2001; De Santis & Spezzaferri, 2001). We use the notation $\boldsymbol{x}^{\boldsymbol{b}}$, where $\boldsymbol{b} = \begin{bmatrix} b_1 & \cdots & b_J \end{bmatrix}^T$, to indicate that the posterior in Eq. (7) contains a fraction of the information in the complete sample data. The larger the b's, the more information from the likelihood (i.e., from the sample data) is contained in the posterior.

In the next step, we integrate $\boldsymbol{\mu}$ out of the joint posterior to obtain the marginal posterior of σ^2:

$$\pi_0^{\mathrm{B}}\left(\sigma^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) = \int_{\mathbb{R}^J} \pi_0^{\mathrm{B}}\left(\boldsymbol{\mu}, \sigma^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \mathrm{d}\boldsymbol{\mu} = \text{Inv-}\chi^2\left(\sigma^2 | \nu, \tau^2\right), \tag{8}$$

where

$$\nu = \left(\sum_{j=1}^{J} b_j n_j\right) - J \quad \text{and} \quad \tau^2 = \frac{\sum_{j=1}^{J} b_j (n_j - 1) s_j^2}{\left(\sum_{j=1}^{J} b_j n_j\right) - J}. \tag{9}$$

Here Inv-$\chi^2\left(\nu, \tau^2\right)$ is the scaled inverse-χ^2 distribution with degrees of freedom parameter $\nu > 0$ and scale parameter $\tau^2 > 0$ (Gelman, Carlin, Stern, & Rubin, 2004), and $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$ is the sample variance of $\boldsymbol{x}_j$.

We then define the prior on the unique variances $\boldsymbol{\sigma}_t^2 = \left[\sigma_1^2 \cdots \sigma_{K_t}^2\right]^T$ under an (in)equality-constrained hypothesis H_t as

$$\pi_t^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) = \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \prod_{k=1}^{K_t} \pi_0^{\mathrm{B}}\left(\sigma_k^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right), \tag{10}$$

where

$$P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right) = \int_{\Omega_t} \prod_{k=1}^{K_t} \pi_0^{\mathrm{B}}\left(\sigma_k^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \tag{11}$$

is the prior probability that the inequality constraints on the unique variances hold. In Eq. (10), its inverse acts as a normalizing constant. The prior in Eq. (10) is referred to as balanced because it implies that every possible ordering of the variances is equally likely a priori. For example, under $H_1 : \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ the prior probability P^{B} in Eq. (11) equals $1/6$ because all $3! = 6$ orderings of the 3 variances are equally likely a priori.

The prior in Eq. (10) must not be too vague or else Bartlett's paradox is induced (e.g., Bartlett, 1957; Jeffreys, 1961; Liang et al., 2008; Lindley, 1957). On the other hand, the prior should not be too informative either because then it would dominate the data. A widely accepted principle that provides a solution to this problem is to let the prior contain minimal information (e.g., Berger & Pericchi, 1996; O'Hagan, 1995; Spiegelhalter & Smith, 1982). We can make the scaled inverse-χ^2 prior in Eq. (8) contain minimal information by setting the degrees of freedom to 1. This can be achieved by setting the fractions to $b_j = (1 + 1/J)/n_j$, for $j = 1, \ldots, J$. This gives us degrees of freedom of $\nu = \left(\sum_{j=1}^{J} b_j n_j\right) - J = \left(\sum_{j=1}^{J}(1 + 1/J)\right) - J = 1$ regardless of the sample sizes $n_1, \ldots, n_J$. Note that the scale parameter τ^2 in Eq. (9) can be interpreted as a weighted average of sums of squares across all populations.

The unconstrained mean vector $\boldsymbol{\mu}$ is common under all hypotheses, which is why we use the non-informative Jeffreys prior $\pi^N(\boldsymbol{\mu}) = C$ for it (Jeffreys, 1961), where C is an unspecified normalizing constant (see, e.g., O'Hagan, 1995). The joint balanced prior on the means and the variances under H_t is then given by

$$\pi_t^{\mathrm{B}}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) = \pi_t^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \pi^N(\boldsymbol{\mu}). \tag{12}$$

Eventually, we define the marginal likelihood under a constrained hypothesis H_t based on the balanced prior as

$$m_t^{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{b}) = \int_{\Omega_t} \int_{\mathbb{R}^J} f_t\left(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \pi_t^{\mathrm{B}}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2 | \boldsymbol{x}^{\boldsymbol{b}}\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2. \tag{13}$$

After some algebra (see "Appendix A"), the marginal likelihood under a constrained hypothesis as in Eq. (1) can be written in an analytic form:

$$m_t^{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{b}) = C \frac{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right)}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \left(\nu\tau^2\right)^{\frac{\nu K_t}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-K_t} \pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k\right)}{2}} \left(\prod_{k=1}^{K_t} \prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}}\right)$$
$$\prod_{k=1}^{K_t} \Gamma\left(\frac{\nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k}{2}\right) \left(\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j} - 1\right) s_{k_j}^2\right)^{-\frac{\nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k}{2}}, \tag{14}$$

where C is the unspecified normalizing constant from the Jeffreys prior on the means, $\Gamma(\cdot)$ is the gamma function, and $s_{k_j}^2$ is the sample variance of the data from the jth population sharing the unique variance σ_k^2. Furthermore,

$$P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right) = \int_{\Omega_t} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(\sigma_k^2 \,\middle|\, \nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k, \frac{\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{\nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \tag{15}$$

is the posterior probability that the inequality constraints on the unique variances hold. Note that the unspecified constant C cancels out in the computation of Bayes factors. The integral in Eq. (15) cannot be computed analytically, but it can be approximated numerically using Monte Carlo methods (see "Appendix C").

5.2. *Generalized Fractional Bayes Factor*

In the construction of the BBF, a possible objection one could have is that there is a slight issue of using the data twice: first for constructing the balanced prior and second for hypothesis testing. Note, however, that this violation is extremely small, as the balanced prior contains minimal information. Another potential issue with the BBF is that the balanced prior shrinks the posterior to the boundary of the parameter space where the variances are equal, which results in a loss of evidence in favor of a true inequality-constrained hypothesis. An alternative approach which avoids these issues is the fractional Bayes factor (FBF) of O'Hagan (1995). In this section, we apply the FBF for the first time to the testing problem formulated in Eq. (1). The FBF is constructed as the ratio of the marginal likelihoods of the complete information in the data and a fraction of the information in the data, both using improper priors:

$$m_t^{\mathrm{F}}(\boldsymbol{x}, b) = \frac{\int_{\Omega_t}\int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)\pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{\sigma}_t^2}{\int_{\Omega_t}\int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)^b\pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{\sigma}_t^2}. \tag{16}$$

Here $\pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)$ is the non-informative Jeffreys prior on the population means and variances given by

$$\pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) = C_t \prod_{k=1}^{K_t} \sigma_k^{-2}\,\mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right), \tag{17}$$

where C_t is an unspecified normalizing constant. The expression $f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)^b$ is the likelihood under H_t to the power of b, a key part of the FBF methodology. The fraction b is a proportion that determines how much of the information in the likelihood (in terms of observations) is contained in $f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)^b$. Note that b is the same under all hypotheses. This is because b effectively divides the likelihood into a training fraction and a test fraction (see, e.g., Gilks, 1995), and the size of these two fractions should be constant across hypotheses.

Choosing the fraction b is a crucial step in the application of the FBF. A popular and widely accepted approach is setting $b = m_0/n$, where m_0 is the size of a minimal training sample and n is the sample size (e.g., Berger & Mortera, 1999; O'Hagan, 1995). This way the information in the data that is used for hypothesis testing is maximal. Despite the useful properties of the FBF (O'Hagan, 1997), De Santis and Spezzaferri (2001) highlighted that the FBF may result in inconsistent behavior in the case of unbalanced data with groups of very different size. This is caused by the fact that the same fraction is used for all groups. In the case of testing mean

parameters, the authors therefore proposed a generalization of the FBF where different fractions are used for different parts of the likelihood. Here, we adopt a similar idea for testing variances using population-specific fractions $b_{k_j} = m_0/n_{k_j} = 2/n_{k_j}$, where $m_0 = 2$. Note that we need two observations from each population for the automatic prior under the unconstrained hypothesis to be proper. The fraction of the likelihood is then given by

$$f_t\left(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\sigma}_t^2\right)^{\boldsymbol{b}} = \prod_{k=1}^{K_t}\prod_{j=1}^{J_k} f\left(\boldsymbol{x}_{k_j}|\mu_{k_j},\sigma_k^2\right)^{b_{k_j}} \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right), \tag{18}$$

where we slightly abuse notation by using the vector of population-specific fractions $\boldsymbol{b}$ as a superscript. Plugging the expression above into Eq. (16), the marginal likelihood of the generalized fractional Bayes factor (GFBF) can be written as (see "Appendix B")

$$\begin{aligned} m_t^{\mathrm{GF}}(\boldsymbol{x},\boldsymbol{b}) = \frac{P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t|\boldsymbol{x}\right)}{P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t|\boldsymbol{x}^{\boldsymbol{b}}\right)} \pi^{-\frac{\sum_{k=1}^{K_t}\sum_{j=1}^{J_k}\left(1-b_{k_j}\right)n_{k_j}}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} b_{k_j}^{\frac{1}{2}}\right) \\ \prod_{k=1}^{K_t}\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}\right)\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} b_{k_j}n_{k_j}\right)-J_k}{2}\right)^{-1} \\ \left(\sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}} \left(\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{\frac{\left(\sum_{j=1}^{J_k} b_{k_j}n_{k_j}\right)-J_k}{2}}, \end{aligned} \tag{19}$$

where

$$P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t|\boldsymbol{x}^{\boldsymbol{b}}\right) = \int_{\Omega_t}\prod_{k=1}^{K_t}\text{Inv-}\chi^2\left(\sigma_k^2\left|\left(\sum_{j=1}^{J_k} b_{k_j}n_{k_j}\right)-J_k, \frac{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j}-1\right)s_{k_j}^2}{\left(\sum_{j=1}^{J_k} b_{k_j}n_{k_j}\right)-J_k}\right.\right)\mathrm{d}\boldsymbol{\sigma}_t^2 \tag{20}$$

is the prior probability that the inequality constraints on the unique variances hold. The expression for the posterior probability that the inequality constraints hold is identical to Eq. (20) with all b's equal to 1, that is,

$$P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t|\boldsymbol{x}\right) = \int_{\Omega_t}\prod_{k=1}^{K_t}\text{Inv-}\chi^2\left(\sigma_k^2\left|\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k, \frac{\sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}\right.\right)\mathrm{d}\boldsymbol{\sigma}_t^2. \tag{21}$$

As for the BBF, the integrals in Eqs. (20) and (21) can be approximated using Monte Carlo methods (see "Appendix C").

5.3. Adjusted Fractional Bayes Factor

A property of the (generalized) FBF that has been criticized is that it may not function as an Occam's razor when testing hypotheses with inequality constraints on the parameters (e.g., Mulder, 2014b). This can be explained by the fact that the implicit automatic prior is concentrated around the likelihood. If an inequality-constrained hypothesis is strongly supported by the

data, which implies that the likelihood and fraction of the likelihood are completely concentrated in the inequality-constrained subspace, the posterior probability and the automatic prior probability in Eq. (19) are approximately equal to 1. Consequently, the marginal likelihood of an inequality-constrained hypothesis that is supported by the data is approximately equal to the marginal likelihood of an unconstrained hypothesis. Thus, the Bayes factor is indecisive even though the data strongly support the more parsimonious inequality-constrained hypothesis. This will be shown in a numerical simulation in the next section.

To circumvent this issue, Böing-Messing et al. (2017) proposed an adjustment of the GFBF. In this adjusted fractional Bayes factor (AFBF) approach the marginal likelihood under hypothesis H_t is defined as

$$m_t^{\mathrm{AF}}(\boldsymbol{x}, \boldsymbol{b}) = \frac{\int_{\Omega_t} \int_{\mathbb{R}^J} f_u\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \pi_u^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2}{\int_{\Omega_t^a} \int_{\mathbb{R}^J} f_u\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)^{\boldsymbol{b}} \pi_u^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2}, \tag{22}$$

where we use the same population-specific fractions $\boldsymbol{b}$ as in the GFBF, that is, we set $b_{k_j} = 2/n_{k_j}$. Apart from the fractions $\boldsymbol{b}$, the formulation above features two differences from the marginal likelihood in the FBF approach in Eq. (16). First, in the denominator the integration region is an adjusted parameter space Ω_t^a given by

$$\Omega_t^a = \left\{\boldsymbol{\sigma}_t^2 : \boldsymbol{R}_t^I \left[a_1\sigma_1^2 \cdots a_{K_t}\sigma_{K_t}^2\right]^T > \boldsymbol{0}\right\}, \tag{23}$$

where $a_1, \ldots, a_{K_t}$ are tuning parameters given by

$$a_k = \frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}, \tag{24}$$

for $k = 1, \ldots, K_t$.

The second difference in Eq. (22) is that the unconstrained likelihood and Jeffreys prior are used instead of the inequality-constrained likelihood and Jeffreys prior under H_t. The unconstrained Jeffreys prior is given by $\pi_u^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) = C_{t,u} \prod_{k=1}^{K_t} \sigma_k^{-2}$. Using the unconstrained likelihood and Jeffreys prior is necessary to ensure that we integrate over the complete adjusted parameter space Ω_t^a in the denominator in Eq. (22). Because the unconstrained Jeffreys prior is used in the denominator, it should also be used in the numerator to ensure that the unspecified normalizing constant cancels out. Despite this adjustment, it is important to note that the numerator in the AFBF is still equal to the marginal likelihood under H_t based a non-informative improper prior, similar as in the original FBF.

The final expression for the marginal likelihood in the AFBF approach is identical to that of the GFBF given in Eq. (19), except that the prior probability that the inequality constraints hold is given by

$$\begin{aligned} P^{\mathrm{AF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t^a | \boldsymbol{x}^{\boldsymbol{b}}\right) &= \int_{\Omega_t^a} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(\sigma_k^2 \,\middle|\, \left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k, \frac{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\ &= \int_{\Omega_t} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(a_k\sigma_k^2 \,\middle|\, \left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k, a_k \frac{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}\right) \mathrm{d}\boldsymbol{\sigma}_t^2, \end{aligned} \tag{25}$$

where the two integrals are equal due to the mathematical result that if $\sigma^2 \sim \text{Inv-}\chi^2\left(\nu, \tau^2\right)$, then $a\sigma^2 \sim \text{Inv-}\chi^2\left(\nu, a\tau^2\right)$. As with the BBF and GFBF, the integrals above can be approximated using Monte Carlo methods (see "Appendix C").

6. Evaluation of the Bayes Factors

In this section, we evaluate the three automatic Bayes factors based on the six desirable properties discussed in Sect. 4. We provide theoretical arguments showing that the priors contain minimal information and that the Bayes factors are scale invariant and large sample consistent. In addition to the theoretical evaluation of large sample consistency, we present a simulation study investigating how fast the evidence in favor of the true hypothesis increases as the sample size increases. Information consistency and robustness to non-normality will be examined by means of simulation studies as well. We use a combination of simulations and theoretical arguments to check the Occam's razor property of the Bayes factors.

6.1. Minimal Information Prior

In the construction of the BBF, it was explicitly taken into account that the balanced prior contains minimal information. Although there is no explicit prior in the GFBF and the AFBF, Gilks (1995) showed that the information in the fraction of the likelihood can be viewed as the information in an implicit underlying prior. Because minimal fractions in the GFBF and the AFBF were considered, we can therefore argue that this property is satisfied by these automatic Bayes factors.

6.2. Scale Invariance

To show that the BBF is scale invariant, we can proceed as follows. Let $w\boldsymbol{x}_{k_j}$ be the rescaled data of the jth group sharing the unique variance σ_k^2, where w is a constant. Then, the sample variance of $w\boldsymbol{x}_{k_j}$ is given by $w^2 s_{k_j}^2$. If we substitute $s_{k_j}^2$ in Eq. (14) with $w^2 s_{k_j}^2$, it can be shown that the marginal likelihood based on the rescaled data is equal to a hypothesis-independent constant times the marginal likelihood based on the original data. The hypothesis-independent constant cancels out in the computation of Bayes factors and posterior probabilities of the hypotheses. Thus, the Bayes factors and posterior probabilities based on the rescaled data are equal to those based on the original data, which shows that the BBF is scale invariant. In a similar manner, it can be shown that the GFBF and the AFBF are scale invariant.

6.3. Occam's Razor When Testing Inequality-Constrained Hypotheses

To illustrate the testing behavior of the three automatic Bayes factors for inequality-constrained hypotheses in particular, we test the order-constrained hypothesis $H_1 : \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ against the unconstrained hypothesis $H_u : \sigma_1^2, \sigma_2^2, \sigma_3^2$. The top row of Fig. 1 shows the BBF (red line), the GFBF (green line), and the AFBF (blue line) of H_1 against H_u for common sample sizes of $n_1 = n_2 = n_3 = n = 5$ (left plot) and $n = 20$ (right plot) and sample variances of $\left[s_1^2\ s_2^2\ s_3^2\right]^T = \left[1\ s\ s^2\right]^T$. We let s^2 go from $\exp(0) = 1$ to $\exp(10) = 22{,}026.47$. Thus, the larger s^2, the larger the size of the order effect. Note that setting $s_2^2 = s$ results in equal sample variance ratios of $s_2^2/s_1^2 = s_3^2/s_2^2 = s$. For the BBF, we set $b_k = (1 + 1/3)/n$, whereas for the GFBF and the AFBF we set $b_k = 2/n$.

Now, according to the Occam's razor principle H_1 should be favored over H_u if the constraints under H_1 are supported by the data since H_1 is more parsimonious than H_u (in the sense that the admissible parameter space under H_1 is a subset of the unconstrained space under H_u). It

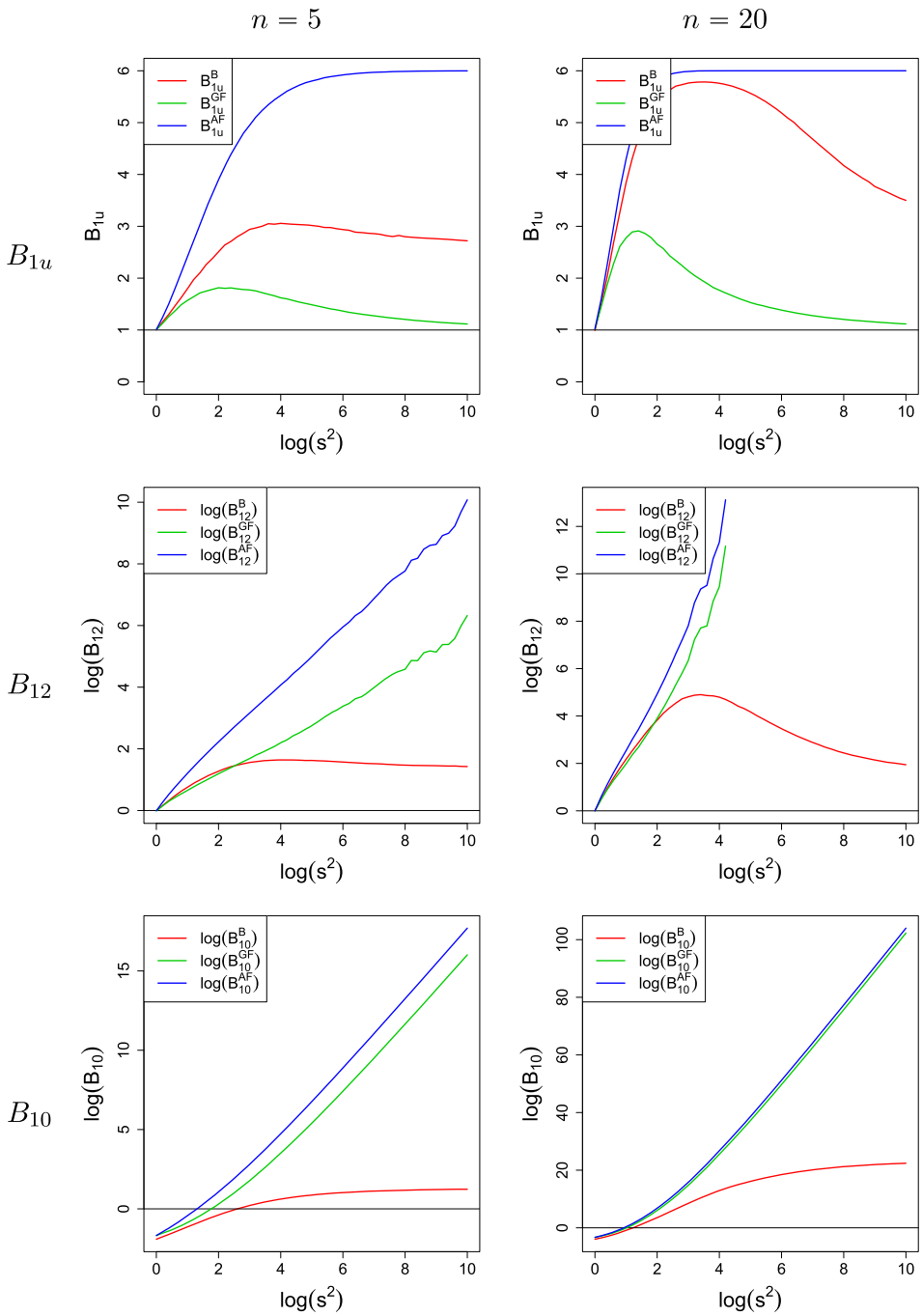


FIGURE 1.
The BBF (red line), GFBF (green line), and AFBF (blue line) testing $H_1 : \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ against $H_u : \sigma_1^2, \sigma_2^2, \sigma_3^2$ (top row), $H_2 : \neg H_1$ (middle row), and $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ (bottom row). The Bayes factors are plotted for common sample sizes of $n_1 = n_2 = n_3 = n = 5$ (left column) and $n = 20$ (right column) and sample variances of $\left[s_1^2\ s_2^2\ s_3^2\right]^T = \left[1\ s\ s^2\right]^T$, where $s^2 \in \left[\exp(0), \exp(10)\right]$. For the BBF, we set $b_k = (1 + 1/3)/n$, whereas for the GFBF and the AFBF we set $b_k = 2/n$ (Color figure online).

can be seen, however, that the GFBF approaches 1 as s^2 grows very large. This means that the GFBF is undecided about H_1 and H_u despite the fact that the data strongly support H_1, which suggests that the GFBF does not function as an Occam's razor in this case. This can be explained as follows. From the expression for the marginal likelihood in Eq. (19), it follows that the GFBF of H_1 against H_u can be written as

$$B_{1u}^{\mathrm{GF}} = \frac{P^{\mathrm{GF}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x}\right)}{P^{\mathrm{GF}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x^b}\right)} \rightarrow \frac{1}{1} = 1, \tag{26}$$

where $P^{\mathrm{GF}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x}\right)$ and $P^{\mathrm{GF}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x^b}\right)$ are the posterior and the prior probability that the inequality constraints under H_1 hold, respectively. Now, for very large effects both probabilities converge to 1, which results in a Bayes factor that converges to 1.

The BBF and the AFBF do not converge to 1 but to a value strictly larger than 1 as s^2 goes to infinity. The explanation for the BBF and the AFBF converging to constants greater than 1 is similar. First, similar to the GFBF, it holds that

$$B_{1u}^{\mathrm{B}} = \frac{P^{\mathrm{B}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x}\right)}{P^{\mathrm{B}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x^b}\right)} \rightarrow \frac{P^{\mathrm{B}^*}}{1/6} = 6 \times P^{\mathrm{B}^*} \tag{27}$$

and

$$B_{1u}^{\mathrm{AF}} = \frac{P^{\mathrm{AF}}\left(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \boldsymbol{x}\right)}{P^{\mathrm{AF}}\left(a_1\sigma_1^2 < a_2\sigma_2^2 < a_3\sigma_3^2 | \boldsymbol{x^b}\right)} \rightarrow \frac{1}{1/6} = 6. \tag{28}$$

For the BBF, the posterior probability converges to $P^{\mathrm{B}^*} \approx 0.45$ for $n = 5$ and $P^{\mathrm{B}^*} \approx 0.50$ for $n = 20$ as the effect size increases. (It does not converge to 1 due to prior shrinkage.) The prior probability always equals $1/6$ since in the BBF approach the prior is the product of three identical distributions, such that each of the 6 possible orderings of the 3 variances is equally likely a priori. Consequently, the BBF converges to $6 \times P^{\mathrm{B}^*} \approx 2.69$ for $n = 5$ and $6 \times P^{\mathrm{B}^*} \approx 3.00$ for $n = 20$ as the effect size increases. In the AFBF approach, the posterior probability goes to 1 as the effect size increases, and the tuning parameters a_1, a_2, a_3 adapt to the sample sizes and sample variances such that the prior probability always equals $1/6$. As a result, the AFBF converges to 6. Thus, contrary to the GFBF, the BBF and the AFBF function as an Occam's razor by favoring the more parsimonious inequality-constrained hypothesis H_1 over the unconstrained hypothesis H_u if the former is strongly supported by the data.

It is important to note that the automatic prior probability that the inequality constraints hold under the AFBF is not always equal to the reciprocal of the number of possible orderings as above. For example, under the inequality-constrained hypothesis $H\colon \sigma_1^2 = \sigma_2^2 < \sigma_3^2$ the prior probability that the inequality constraint holds is computed using scaled inverse-χ^2 distributions with identical scale parameters but different degrees of freedom (due to the equality constraint; cf. Eq. (25)). Thus, since the distributions are not identical, the automatic prior probability that the inequality constraint holds is not equal to the reciprocal of the number of possible orderings.

6.4. Information Consistency

We evaluate information consistency for two different tests on variances: (i) testing an inequality-constrained hypothesis against its complement and (ii) testing an inequality-constrained hypothesis against the null hypothesis. We will call the Bayes factor B_{12} (B_{10}) of an inequality-constrained hypothesis $H_1\colon \boldsymbol{R}_1^I\boldsymbol{\sigma}^2 > \boldsymbol{0}$ against $H_2\colon \neg H_1$ ($H_0\colon \sigma_1^2 = \cdots = \sigma_J^2$) information consistent if $B_{12} \rightarrow \infty$ ($B_{10} \rightarrow \infty$) as each element in $\hat{\boldsymbol{\xi}} = \boldsymbol{R}_1^I\left[\log\left(\hat{\sigma}_1^2\right) \cdots \log\left(\hat{\sigma}_J^2\right)\right]^T$

goes to infinity, while keeping the sample size fixed. If the Bayes factor converges to a constant $B_{12}^* < \infty$ ($B_{10}^* < \infty$) instead, then it is referred to as information inconsistent.

First, we investigate information consistency when testing an inequality-constrained hypothesis against its complement. The middle row of Fig. 1 shows the logarithm of the BBF, the GFBF, and the AFBF of $H_1 \colon \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ against $H_2 \colon \neg H_1$ as $\hat{\boldsymbol{\xi}} = \left[\log\left(s_2^2/s_1^2\right) \log\left(s_3^2/s_2^2\right)\right]^T$ increases from $\left[0\ 0\right]^T$ to $\left[5\ 5\right]^T$. The results indicate that the GFBF and the AFBF are information consistent, whereas the BBF is information inconsistent. As the effect size increases, the BBF converges to a constant $B_{12}^{\mathrm{B}^*} \approx \exp(1.399) = 4.05$ for $n = 5$ and $B_{12}^{\mathrm{B}^*} \approx \exp(1.607) = 4.99$ for $n = 20$. This behavior of the BBF can be explained by the fact that the posterior probability that the inequality constraints under H_1 hold converges to $P^{\mathrm{B}^*} \approx 0.45$ for $n = 5$ and $P^{\mathrm{B}^*} \approx 0.50$ for $n = 20$ (as was found in Sect. 6.3), which implies that the probability that the inequality constraints under H_1 do not hold (as is stated in H_2) converges to $1 - P^{\mathrm{B}^*} \approx 0.55$ for $n = 5$ and $1 - P^{\mathrm{B}^*} \approx 0.50$ for $n = 20$. The BBF of H_1 against H_2 thus converges to $B_{12}^{\mathrm{B}^*} = B_{1u}^{\mathrm{B}^*} / B_{2u}^{\mathrm{B}^*} \approx \frac{P^{\mathrm{B}^*}}{1/6} \Big/ \frac{1-P^{\mathrm{B}^*}}{5/6}$, which equals 4.05 for $n = 5$ and 4.99 for $n = 20$. Note that H_1 is more parsimonious than H_2 because H_1 covers $1/6$ of the unconstrained parameter space while H_2 covers $5/6$ of the unconstrained space. The results show that the AFBF indicates stronger evidence in favor of the more parsimonious hypothesis H_1 than the GFBF. This again illustrates that the AFBF functions as an Occam's razor, whereas the GFBF does not.

Next, we investigate information consistency when testing the order-constrained hypothesis $H_1 \colon \sigma_1^2 < \sigma_2^2 < \sigma_3^2$ against the null hypothesis $H_0 \colon \sigma_1^2 = \sigma_2^2 = \sigma_3^2$. The bottom row of Fig. 1 shows the logarithm of the Bayes factor B_{10} as a function of the effect size s^2. The results indicate that the GFBF and the AFBF are information consistent since for these two Bayes factors the evidence in favor of H_1 goes to infinity as the size of the order effect increases. Again, the AFBF indicates stronger evidence in favor of H_1 than the GFBF. The BBF, on the other hand, does not show information consistent behavior in this case either. The inconsistent behavior of the BBF illustrates that it may not provide a good quantification of the relative evidence in the data between (in)equality-constrained hypotheses in the case of small samples and large effects.

6.5. Large Sample Consistency

Using the same argument as O'Hagan (1995), it can be shown that the BBF, GFBF, and AFBF are large sample consistent when testing equality-constrained hypotheses. When testing inequality-constrained hypotheses, the automatic Bayes factors are also consistent. This can be seen from the expressions of the marginal likelihoods, which, in the case of inequality constraints, depend on the ratio of the posterior probability that the inequality constraints hold and the automatic prior probability that the inequality constraints hold (e.g., $P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right) / P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)$ in the GFBF). Under the true hypothesis, the posterior probability that the inequality constraints hold goes to 1 as the sample size goes to infinity, whereas under the incorrect hypotheses it goes to 0. Similarly, the prior probability that the inequality constraints hold converges to a constant unequal to 0. As a result, the Bayes factor, that is, the ratio of the marginal likelihoods, goes to infinity toward the true hypothesis. Consequently, the posterior probability of the true hypothesis goes to 1 while the posterior probabilities of the incorrect hypotheses go to 0. This implies consistency.

In addition to this theoretical result, we also performed a simulation study to investigate how fast the evidence for the true hypothesis accumulates as the sample size grows. This will provide some insight into how much data are needed to obtain strong evidence for a true constrained hypothesis on variances in different scenarios.

6.5.1. Simulation Setup In this simulation, we tested hypotheses on the variances of $J \in \{4, 6\}$ populations. The results for 6 populations are similar to those for 4 populations, which is why we

TABLE 1.
Overview of the population variances used in the simulation study with 4 populations.

Population	Effect	σ_1^2	σ_2^2	σ_3^2	σ_4^2
Null	No	1.00	1.00	1.00	1.00
Order	Small	1.00	1.14	1.30	1.49
	Medium	1.00	1.36	1.84	2.50
	Large	1.00	1.58	2.48	3.91
Mixed	Small	1.00	1.00	1.33	1.33
	Medium	1.00	1.00	2.00	2.00
	Large	1.00	1.00	2.94	2.94
Near order	Small	1.14	1.00	1.30	1.49
	Medium	1.36	1.00	1.84	2.50
	Large	1.58	1.00	2.48	3.91
Reverse order	Small	1.49	1.30	1.14	1.00
	Medium	2.50	1.84	1.36	1.00
	Large	3.91	2.48	1.58	1.00

present the former in "Appendix D." For the simulation with 4 populations, we used a simulation design with three factors:

1. *Pattern of the population variances:* We considered five different variance patterns, referred to as null pattern, order pattern, mixed pattern, near-order pattern, and reverse-order pattern. In the null pattern, all population variances were equal, $\sigma_1^2 = \cdots = \sigma_4^2$. In the order pattern, the variances followed an increasing order, $\sigma_1^2 < \cdots < \sigma_4^2$. In the mixed pattern, the structure of the variances was $\sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2$. Note that such a pattern could emerge in a 2×2 factorial study where there is an effect of only one of the two factors (similar to our third motivating example in Sect. 2). The near-order pattern was similar to the order pattern with the difference that the variances of populations 1 and 2 were interchanged: $\sigma_2^2 < \sigma_1^2 < \sigma_3^2 < \sigma_4^2$. Finally, in the reverse-order pattern the variances were ordered as $\sigma_4^2 < \cdots < \sigma_1^2$.
2. *Effect size:* In all patterns but the null pattern we considered three effect sizes for the population variances: small, medium, and large. The effect size is given by the ratio of the largest to the smallest population variance (e.g., Ruscio & Roche, 2012). To determine the actual values of the population variances, we followed the approach of Böing-Messing et al. (2017). The authors set the smallest variance equal to 1 and determine the largest variance under different effect sizes based on established guidelines for testing equality of means. The intermediate variances are then determined such that the ratio of adjacent variances is constant. The resulting values of the population variances are given in Table 1. Note that in the null pattern we set all variances equal to 1.
3. *Sample size:* We drew samples of common size $n_1 = \cdots = n_4 = n \in \{5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10{,}000\}$ from the four populations.

Hence, there were 11 conditions for the null pattern and 4 (patterns) $\times$ 3 (effect sizes) $\times$ 11 (sample sizes) $=$ 132 conditions for the remaining four patterns, resulting in a total of 143 conditions. In each condition, we drew 1000 samples $\boldsymbol{x}^{(m)} = \left[\boldsymbol{x}_1^{(m)} \cdots \boldsymbol{x}_J^{(m)}\right]$, $m = 1, \ldots, 1000$. Here $\boldsymbol{x}_j^{(m)} = \left[x_{1j}^{(m)} \cdots x_{nj}^{(m)}\right]^T$, where $x_{ij}^{(m)}$ is distributed as in Eq. (2). We specified the population variances according to Table 1 and set all population means equal to 0 (note that the three Bayes

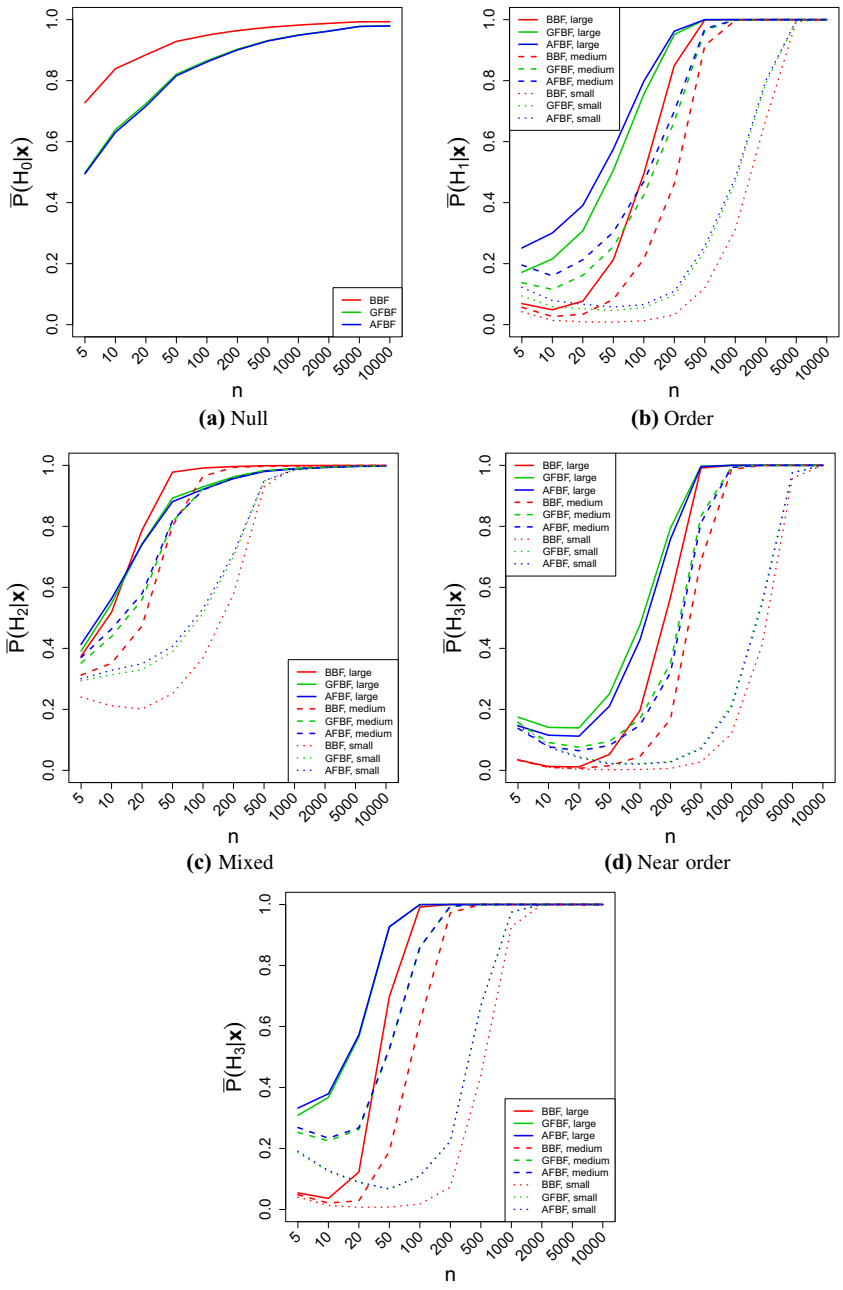
factors do not depend on the sample means, cf. Eqs. (14), (19), and (25)). In each of the 1000 samples per condition, we tested four hypotheses using the three different Bayes factors:

$$\begin{aligned} H_0 &: \sigma_1^2 = \cdots = \sigma_4^2, \\ H_1 &: \sigma_1^2 < \cdots < \sigma_4^2, \\ H_2 &: \sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2, \\ H_3 &: \neg\,(H_0 \vee H_1 \vee H_2). \end{aligned} \tag{29}$$

Here, H_3 is the complement which comprises all possible hypotheses except H_0, H_1, and H_2. Note that the marginal likelihood under H_3 is equal to the marginal likelihood under the hypothesis $H_4 : \neg H_1$ because the probability of the event that two or more variances are exactly equal is 0. Furthermore, note that for the near-order and reverse-order patterns the true hypothesis is contained in the complement H_3 (cf. Table 1). In each sample, we then used the marginal likelihoods under all four hypotheses to compute the posterior probability of the true hypothesis H_t as $P\left(H_t|\boldsymbol{x}^{(m)}\right) = m_t\left(\boldsymbol{x}^{(m)}\right) \Big/ \sum_{t'=0}^{3} m_{t'}\left(\boldsymbol{x}^{(m)}\right)$, where we assumed equal prior probabilities of the hypotheses. Eventually, we computed the expected posterior probability of H_t as $\bar{P}(H_t|\boldsymbol{x}) = \frac{1}{1000}\sum_{m=1}^{1000} P\left(H_t|\boldsymbol{x}^{(m)}\right)$.

6.5.2. Results Figure 2 shows the simulation results for the five variance patterns. The plots show the expected posterior probability of the true hypothesis H_t, $t = 0, 1, 2, 3$, as a function of the common sample size n for the BBF (red lines), the GFBF (green lines), and the AFBF (blue lines). Figure 2b–e shows the results for a small effect (dotted lines), medium effect (dashed lines), and large effect (solid lines). Note that in Fig. 2a and e the lines for the GFBF and AFBF largely overlap. It can be seen that under all variance patterns and effect sizes the lines approach 1 as the sample size increases, which is a result of the large sample consistency of the three automatic Bayes factors. Naturally, the expected posterior probability of the true hypothesis goes to 1 fastest under a large effect because small effects are more difficult to detect for a given sample size. Moreover, the plots show that the BBF converges fastest to a true hypothesis if two or more population variances are equal (see the null and the mixed pattern in Fig. 2a and c, respectively), whereas the GFBF and the AFBF converge fastest to the true hypotheses if none of the population variances are equal (see the order patterns in Fig. 2b, d, and e). Furthermore, it can be seen that the GFBF and the AFBF behave similarly. The GFBF converges slightly faster to a true null hypothesis (see Fig. 2a), whereas the AFBF converges somewhat faster to a true inequality-constrained hypothesis (see Fig. 2b and c).

Under the null pattern (Fig. 2a), sample sizes of 10 (BBF) and 50 (GFBF, AFBF) result in posterior probabilities of the true null hypothesis H_0 of at least 0.8. Under the order pattern (Fig. 2b), we need considerably larger samples to obtain posterior probabilities of the true order-constrained hypothesis H_1 of at least 0.8. While under a large effect sample sizes of 200 are sufficient for reaching a value of at least 0.8, under a small effect we need sample sizes of 5000 (BBF) and 2000 (GFBF, AFBF), respectively. Under the mixed pattern (Fig. 2c), sample sizes of 50 result in a posterior probability of the true mixed hypothesis H_2 of at least 0.8 if the effect is large, whereas under a small effect sample sizes of 500 are required. In Fig. 2d, it can be seen that rather large samples are necessary to detect that the order of the first two population variances is reversed. While under a large effect sample sizes of 200 (GFBF) and 500 (BBF, AFBF), respectively, result in a posterior probability of the true complement H_3 of at least 0.8, under a small effect this value is only reached for sample sizes of 5000. Figure 2e shows that it is easier for the three Bayes factors to detect that the order of the four population variances is reversed. If the effect is large, posterior probabilities of the true complement H_3 of at least 0.8



(a) Null **(b)** Order

(c) Mixed **(d)** Near order

(e) Reverse order

FIGURE 2.

Results of a simulation study comparing the performance of the three automatic Bayes factors in testing variances of 4 populations. We examined five different patterns of the population variances: **a** $\sigma_1^2 = \cdots = \sigma_4^2$, **b** $\sigma_1^2 < \cdots < \sigma_4^2$, **c** $\sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2$, **d** $\sigma_2^2 < \sigma_1^2 < \sigma_3^2 < \sigma_4^2$, and **e** $\sigma_4^2 < \cdots < \sigma_1^2$. In patterns **b** to **e** we considered three different sizes of the order effect: small, medium, and large. For each combination of pattern and effect size, we drew 1000 samples of size $n_1 = \cdots = n_4 = n$. In each sample we then tested four hypotheses: $H_0 : \sigma_1^2 = \cdots = \sigma_4^2$, $H_1 : \sigma_1^2 < \cdots < \sigma_4^2$, $H_2 : \sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2$, and $H_3 : \neg(H_0 \vee H_1 \vee H_2)$. Eventually, we computed the expected posterior probability of the true hypothesis $\bar{P}(H_t|\boldsymbol{x})$ across the 1000 samples. The plots show $\bar{P}(H_t|\boldsymbol{x})$ as a function of the common sample size n for the BBF (red lines), GFBF (green lines), and AFBF (blue lines) under a small effect (dotted lines), medium effect (dashed lines), and large effect (solid lines) (Color figure online).

TABLE 2.
Sample sizes and sample variances for three examples.

Example	Group	n	s^2
Example 1	1: Treatment 1	7	0.30
	2: Treatment 2	5	0.79
	3: Treatment 3	8	2.89
	4: Treatment 4	6	3.61
Example 2	1: Controls	17	15.52
	2: Tourette's patients	17	20.07
	3: ADHD patients	17	38.81
Example 3	1: Male leader, appointed at random	30	3.46
	2: Female leader, appointed at random	30	1.32
	3: Male leader, appointed on ability	30	3.20
	4: Female leader, appointed on ability	30	2.10

TABLE 3.
Results for three examples. The posterior probabilities of the hypotheses were computed assuming equal prior probabilities. In some cases the posterior probabilities do not sum to 1 due to rounding.

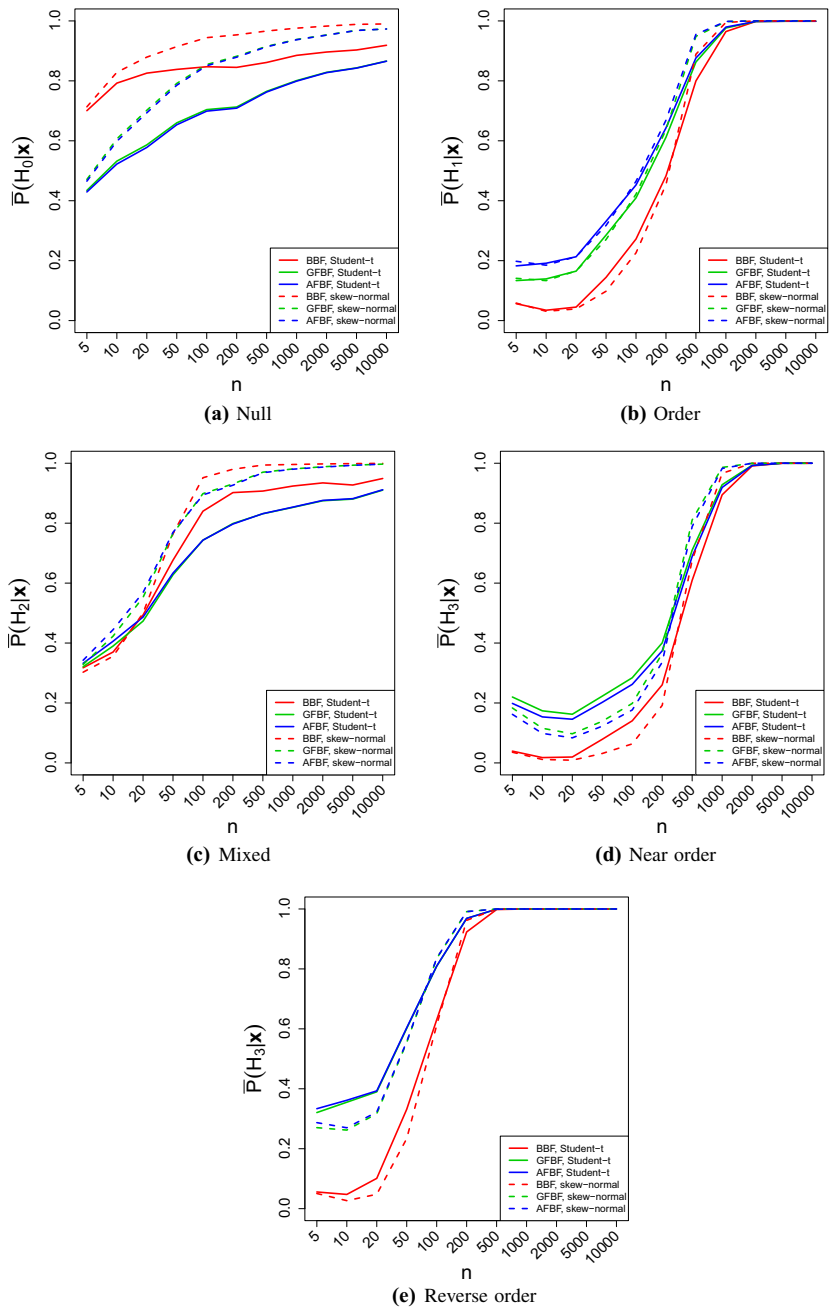
Example	Bayes factor	$P(H_0\|\boldsymbol{x})$	$P(H_1\|\boldsymbol{x})$	$P(H_2\|\boldsymbol{x})$	$P(H_3\|\boldsymbol{x})$
Example 1	BBF	0.74	0.23	0.04	—
	GFBF	0.12	0.72	0.17	—
	AFBF	0.04	0.91	0.05	—
Example 2	BBF	0.35	0.48	0.14	0.03
	GFBF	0.28	0.40	0.25	0.07
	AFBF	0.24	0.43	0.28	0.06
Example 3	BBF	0.37	0.62	0.00	—
	GFBF	0.16	0.82	0.03	—
	AFBF	0.12	0.86	0.02	—

are reached for sample sizes of 100 (BBF) and 50 (GFBF, AFBF), while under a small effect we need sample sizes of 1000 to surpass this mark.

6.6. Robustness to Non-normality

To check for robustness of the proposed Bayes factors, we repeated the simulation study from the previous section with non-normal data. We considered t-distributed as well as skew-normally distributed data. The simulation setup was the same as in Sect. 6.5.1, except that the data were sampled from a $t(\mu_j = 0, \sigma_j, \nu = 5)$ distribution and a $SN(\mu_j = 0, \sigma_j, \alpha = 4)$ distribution, respectively, where α is the shape parameter of the skew-normal distribution and the scale parameters $\sigma_1, \ldots, \sigma_4$ were specified according to Table 1. A Bayes factor is robust in this setting if it shows large sample consistent behavior in the sense that the expected posterior probability of the true hypothesis goes to 1 as the sample size increases despite the data coming from a non-normal distribution.

For the sake of brevity, we only present the simulation results for 4 populations and a medium effect in this section. The results are shown in Fig. 3. The plots show the expected posterior probability of the true hypothesis for data coming from a t-distribution (solid lines) and a skew-



(a) Null **(b)** Order

(c) Mixed **(d)** Near order

(e) Reverse order

FIGURE 3.

Results of a simulation study investigating the robustness of the three automatic Bayes factors to non-normality when testing variances of 4 populations. We considered data coming from a t-distribution (dashed lines) and a skew-normal distribution (solid lines). The scale parameters of the distributions were specified according to five patterns: **a** $\sigma_1^2 = \cdots = \sigma_4^2$, **b** $\sigma_1^2 < \cdots < \sigma_4^2$, **c** $\sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2$, **d** $\sigma_2^2 < \sigma_1^2 < \sigma_3^2 < \sigma_4^2$, and **e** $\sigma_4^2 < \cdots < \sigma_1^2$. In patterns **b** to **e**, we used a medium size of the order effect. For each combination of pattern and distribution we drew 1000 samples of size $n_1 = \cdots = n_4 = n$. In each sample, we then tested four hypotheses: $H_0 : \sigma_1^2 = \cdots = \sigma_4^2$, $H_1 : \sigma_1^2 < \cdots < \sigma_4^2$, $H_2 : \sigma_1^2 = \sigma_2^2 < \sigma_3^2 = \sigma_4^2$, and $H_3 : \neg(H_0 \vee H_1 \vee H_2)$. Eventually, we computed the expected posterior probability of the true hypothesis $\bar{P}(H_t|\boldsymbol{x})$ across the 1000 samples. The plots show $\bar{P}(H_t|\boldsymbol{x})$ as a function of the common sample size n for the BBF (red lines), GFBF (green lines), and AFBF (blue lines) (Color figure online).

normal distribution (dashed lines), respectively. In general, the Bayes factors appeared to be robust to non-normality, as can be seen from the posterior probabilities approaching 1 as the sample size increases. Furthermore, the differences between the three Bayes factors were the same as in the simulation study with normally distributed data (cf. Fig. 2): On the one hand, the BBF provided stronger evidence in favor of true hypotheses containing equality constraints (Fig. 3a and c), except under the mixed pattern and small samples (see Fig. 3c). The GFBF and AFBF, on the other hand, yielded stronger evidence in favor of a true order-constrained hypothesis (Fig. 3b) and complement (Fig. 3d and e). The results for the remaining conditions in the simulation study were similar. Most importantly, the three automatic Bayes factors showed robust behavior in these conditions as well.

7. Motivating Examples (Continued)

We next apply the three automatic Bayes factors to actual data from the three motivating examples introduced in Sect. 2. We begin with the hypothetical study with four treatment groups from Weerahandi (1995). Here, we formulated the following three hypotheses on the group variances: $H_0\colon \sigma_1^2 = \cdots = \sigma_4^2$, $H_1\colon \sigma_1^2 < \cdots < \sigma_4^2$, and $H_2\colon \neg(H_0 \vee H_1)$. Table 2 (Example 1) shows the sample sizes and sample variances of the four treatment groups as reported by Weerahandi. It appears that the data support H_1 since the sample variances follow an increasing pattern. We applied the Bayes factors to the data to determine the evidence in favor of the three competing hypotheses. The results are shown in Table 3 (Example 1). The posterior probabilities of the hypotheses were computed assuming equal prior probabilities. It can be seen that the BBF favors H_0. The GFBF and the AFBF, on the other hand, favor H_1, with the AFBF indicating considerably weaker evidence in favor of H_0 and H_2. Overall, the results are in line with the findings of the simulation study, where the BBF provides stronger evidence in favor of the null hypothesis, whereas the GFBF and the AFBF yield stronger evidence in favor of inequality-constrained hypotheses. The fact that the GFBF and the AFBF support the order-constrained hypothesis H_1 despite the small sample sizes is due to the large effect size of $s_4^2/s_1^2 = 3.61/0.30 = 11.93$. Comparing the logarithm of this effect size with the results in the bottom row of Fig. 1 indicates that the preference of the BBF for the null hypothesis may be a result of information inconsistency: From the plots, it can be seen that for an effect size of $\log(11.93) = 2.48$ the BBF already shows information inconsistent behavior. This suggests relying on the results of the GFBF or the AFBF, which indicate evidence in favor of the order-constrained hypothesis H_1 stating that the variance increases across the treatment groups.

In our second motivating example (taken from Silverstein et al., 1995), we formulated the following hypotheses on the variances of the attentional performances of unaffected controls (group 1), Tourette's patients (group 2), and ADHD patients (group 3): $H_0\colon \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, $H_1\colon \sigma_1^2 = \sigma_2^2 < \sigma_3^2$, $H_2\colon \sigma_1^2 < \sigma_2^2 = \sigma_3^2$, and $H_3\colon \neg(H_0 \vee H_1 \vee H_2)$. Table 2 (Example 2) shows the sample variances of the attentional performances in the three groups. The results are shown in Table 3 (Example 2). It can be seen that the three automatic Bayes factors produce similar results. In particular, the three Bayes factors favor H_1, which states that Tourette's patients are as heterogeneous as unaffected controls, and both groups are less heterogeneous than ADHD patients. However, while we can rule out the complement H_3, the posterior probabilities indicate some evidence in favor of H_0 and H_2. It can be seen that the AFBF provides somewhat stronger evidence in favor of the inequality-constrained hypotheses than the GFBF. This behavior was also observed in the numerical studies in Sect. 6.

In our final motivating example (taken from Lucas, 2003), the following hypotheses were formulated on the variances of the group leaders' influence (as measured by the number of times that a participant changed his/her opinion to match the group leader's opinion): $H_0\colon \sigma_1^2 = \cdots = \sigma_4^2$,

$H_1\colon \sigma_2^2 = \sigma_4^2 < \sigma_1^2 = \sigma_3^2$, and $H_2\colon \neg(H_0 \vee H_1)$, where $\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\}$ are the variances of the groups whose leader is {(male, random), (female, random), (male, based on ability), (female, based on ability)}. Table 2 (Example 3) shows the sample variances of the counts in the four experimental groups. The results of the multiple hypothesis test are shown in Table 3 (Example 3). It can be seen that H_1 receives strongest support from all three automatic Bayes factors. While there is some evidence in favor of H_0 (especially for the BBF), the complement H_2 can be ruled out given posterior probabilities close to 0. In conclusion, the Bayes factors indicate that the variance is greater when the leader is male rather than female and that there is no effect of the way the leader was appointed.

8. Conclusion

In this article, we presented three automatic Bayes factors for testing (in)equality-constrained hypotheses on variances. We first introduced the balanced Bayes factor, which is based on identical automatic priors for the unique variances under each hypothesis. The hyperparameters of this prior are determined automatically using information from the sample data. The second Bayes factor is the fractional Bayes factor of O'Hagan (1995), which we derived for testing (in)equality-constrained hypotheses on variances. We proposed a generalization of the fractional approach using population-specific fractions instead of a common fraction. The third Bayes factor we presented is an adjustment of the fractional Bayes factor such that the parsimony of inequality-constrained hypotheses is incorporated. The three Bayes factors are fully automatic for testing multiple hypotheses with equality and inequality constraints on the population variances. There is no need for the user to specify priors under all hypotheses to be tested. Instead, the user only needs to provide the sample sizes and sample variances.

The Bayes factors were evaluated based on six criteria. First, the (implicit) prior in each Bayes factor contains minimal information. Second, all three Bayes factors are scale invariant. Third, results of numerical studies indicated that the GFBF does not properly function as an Occam's razor when testing inequality-constrained hypotheses on variances. The BBF and the AFBF, on the other hand, always behaved as an Occam's razor in this situation. Fourth, numerical results indicated that the BBF is information inconsistent when testing (in)equality-constrained hypotheses. The GFBF and AFBF, on the other hand, showed information consistent behavior. Fifth, all three Bayes factors are large sample consistent. Sixth, results of a simulation study indicated that the three Bayes factors are robust to violations of normality in the data. Based on our findings we recommend the AFBF for quantifying the relative evidence in the data between multiple constrained hypotheses on variances when prior information about the magnitude of the effects is unavailable.

In this article, we tested hypotheses involving equality and inequality constraints with equal coefficients for the variances. Thus, each hypothesis states whether a certain variance is larger than, equal to, or smaller than another variance. Experience has shown that relationships between variances can often be appropriately described using hypotheses of this type, as we illustrated with our motivating examples. In fact, other popular models in the psychological sciences imply similar relationships between the variances. For example, in the random slope model the variance may either decrease over time, increase over time, or first decrease and then increase over time (e.g., Snijders & Bosker, 2012). In practice, however, the true hypothesis might be a more complicated function of the variances such as $\sigma_1^2/\sigma_2^2 < \sigma_3^2/\sigma_4^2$. While such hypotheses cannot be tested directly with our Bayes factor approaches, there is a way to safeguard against making erroneous conclusions in case the true hypothesis is a complicated function of the variances. When testing a set of (in)equality-constrained hypotheses on variances, it is advisable to include the complement of the hypotheses under consideration. This way, the complete parameter space is covered by the

hypotheses. Allowing the user to directly specify more complicated hypotheses is an interesting topic for future research. Here, the challenge is to incorporate the complicated constraints when computing the marginal likelihoods.

The Bayes factors we presented in this article can also be used to test the assumption of homogeneity of variances before conducting a classical F-test in an ANOVA setting. Here the null hypothesis stating homogeneity of variances can be tested against the unconstrained alternative hypothesis. This approach has two advantages: First, there is no need to adjust the significance level in the F-test for multiple testing. Second, the Bayes factors are able to quantify the evidence in favor of homogeneity of variances, which is a useful property for determining whether this assumptions holds. A natural extension of our testing approach would be to consider hypotheses with constraints on the population variances as well as the population means in the ANOVA setting. Such a method would be useful when a researcher would like to simultaneously test (in)equality constraints on the variances and the means. An example of a multiple hypothesis test with (in)equality constraints on the means in an ANOVA setting can be found in Mulder (2014b). The author computed the Bayes factors under the assumption of homogeneity of variances. This assumption could be relaxed, allowing for the specification of constrained hypotheses on both the means and the variances. A joint prior distribution on the means and the variances could then be specified using a combination of the methods discussed in this article and in Mulder (2014b).

Further extensions of our approach to testing (in)equality-constrained hypotheses on variances are conceivable. The problem of testing constraints on variances also naturally arises for repeated measurements and other types of data where there is a dependency between the observations. Such data can be analyzed with different kinds of models. First, one might consider using a multivariate regression model where the errors are assumed to follow a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$. Equality- and inequality-constrained hypotheses could then be formulated on the variances on the main diagonal of $\boldsymbol{\Sigma}$. A second option would be using a random effects model to take into account that observations are dependent. In such a model, (in)equality-constrained hypotheses could be formulated on the variances of the random effects (in the spirit of Mulder & Fox, 2013) or the errors. Similarly, (in)equality-constrained hypotheses could be formulated on the random effects variances in item response models (building on the work of, e.g., Fox, Mulder, & Sinharay, 2017, and Verhagen & Fox, 2013). Another area where dependencies between variables play an important role is in structural equation modeling. Here, (in)equality-constrained hypotheses on variances are conceivable as well. For example, in a factor analytic model one might be interested in testing constraints on the variances of the indicators' errors. The current article will be a good starting point for testing variance components in these more complex models.

Appendix A: Computation of $m_t^{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{b})$

The final expression for the marginal likelihood under an (in)equality-constrained hypothesis H_t in the balanced Bayes factor can be derived as follows:

$$
\begin{aligned}
m_t^{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{b}) &= \int_{\Omega_t} \int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \pi_t^{\mathrm{B}}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2|\boldsymbol{x}^{\boldsymbol{b}}\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= \int_{\Omega_t} \int_{\mathbb{R}^J} \left(\prod_{k=1}^{K_t} \prod_{j=1}^{J_k} f\left(\boldsymbol{x}_{k_j}|\mu_{k_j}, \sigma_k^2\right) \right) \\
&\qquad C \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t|\boldsymbol{x}^{\boldsymbol{b}}\right)} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(\sigma_k^2|\nu, \tau^2\right) \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2
\end{aligned}
$$

$$
\begin{aligned}
&= C \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \int_{\Omega_t} \prod_{k=1}^{K_t} \left(\frac{\nu\tau^2}{2}\right)^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-1} \left(\sigma_k^2\right)^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{\nu\tau^2}{2\sigma_k^2}\right) \\
&\quad \prod_{j=1}^{J_k} \int_{\mathbb{R}} \left(\sigma_k^2 2\pi\right)^{-\frac{n_{k_j}}{2}} \exp\left(-\frac{1}{2\sigma_k^2}\left(\left(n_{k_j}-1\right)s_{k_j}^2 + n_{k_j}\left(\bar{x}_{k_j}-\mu_{k_j}\right)^2\right)\right) \mathrm{d}\mu_{k_j}\, \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \left(\frac{\nu\tau^2}{2}\right)^{\frac{\nu K_t}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-K_t} \int_{\Omega_t} \prod_{k=1}^{K_t} \left(\sigma_k^2\right)^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{\nu\tau^2}{2\sigma_k^2}\right) \\
&\quad \prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}} \left(\sigma_k^2 2\pi\right)^{-\frac{n_{k_j}-1}{2}} \exp\left(-\frac{\left(n_{k_j}-1\right)s_{k_j}^2}{2\sigma_k^2}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \left(\frac{\nu\tau^2}{2}\right)^{\frac{\nu K_t}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-K_t} (2\pi)^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k\right)}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}}\right) \\
&\quad \int_{\Omega_t} \prod_{k=1}^{K_t} \left(\sigma_k^2\right)^{-\left(\frac{\nu+\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}+1\right)} \exp\left(-\frac{\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{2\sigma_k^2}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C \frac{1}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \left(\nu\tau^2\right)^{\frac{\nu K_t}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-K_t} \pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k\right)}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}}\right) \\
&\quad \left(\prod_{k=1}^{K_t} \Gamma\left(\frac{\nu+\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}\right)\left(\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\nu+\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}}\right) \\
&\quad \int_{\Omega_t} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(\sigma_k^2 \,\middle|\, \nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k, \frac{\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{\nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k}\right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C \frac{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right)}{P^{\mathrm{B}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)} \left(\nu\tau^2\right)^{\frac{\nu K_t}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-K_t} \pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k\right)}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}}\right) \\
&\quad \prod_{k=1}^{K_t} \Gamma\left(\frac{\nu+\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}\right)\left(\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\nu+\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}}, \qquad (30)
\end{aligned}
$$

where in the third line we may drop the indicator function because the integration region for the variances is already restricted to Ω_t, and the integrand in the fifth line is a product of kernels of scaled inverse-χ^2 distributions with degrees of freedom parameters $\nu_k = \nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k$ and scale parameters $\tau_k^2 = \frac{\nu\tau^2 + \sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{\nu + \left(\sum_{j=1}^{J_k} n_{k_j}\right) - J_k}$.

Appendix B: Computation of $m_t^{\mathrm{GF}}(\boldsymbol{x}, \boldsymbol{b})$

In the generalized fractional Bayes factor, the marginal likelihood under an (in)equality-constrained hypothesis H_t is defined as

$$m_t^{\mathrm{GF}}(\boldsymbol{x}, \boldsymbol{b}) = \frac{\int_{\Omega_t} \int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2}{\int_{\Omega_t} \int_{\mathbb{R}^J} f_t\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right)^{\boldsymbol{b}} \pi_t^N\left(\boldsymbol{\mu}, \boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2} = \frac{m_t^N(\boldsymbol{x})}{m_t^N(\boldsymbol{x}^{\boldsymbol{b}})}. \tag{31}$$

We first derive the denominator:

$$\begin{aligned}
m_t^N(\boldsymbol{x}^{\boldsymbol{b}}) &= \int_{\Omega_t} \int_{\mathbb{R}^J} \left(\prod_{k=1}^{K_t} \prod_{j=1}^{J_k} f\left(\boldsymbol{x}_{k_j}|\mu_{k_j}, \sigma_k^2\right)^{b_{k_j}} \right) C_t \prod_{k=1}^{K_t} \sigma_k^{-2}\, \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^2\right) \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C_t \int_{\Omega_t} \prod_{k=1}^{K_t} \sigma_k^{-2} \prod_{j=1}^{J_k} \int_{\mathbb{R}} \left(\sigma_k^2 2\pi\right)^{-\frac{b_{k_j} n_{k_j}}{2}} \\
&\quad \exp\left(-\frac{b_{k_j}}{2\sigma_k^2} \left(\left(n_{k_j} - 1\right) s_{k_j}^2 + n_{k_j} \left(\bar{x}_{k_j} - \mu_{k_j}\right)^2 \right) \right) \mathrm{d}\mu_{k_j}\, \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C_t \int_{\Omega_t} \prod_{k=1}^{K_t} \sigma_k^{-2} \prod_{j=1}^{J_k} \left(b_{k_j} n_{k_j}\right)^{-\frac{1}{2}} \left(\sigma_k^2 2\pi\right)^{-\frac{b_{k_j} n_{k_j} - 1}{2}} \exp\left(-\frac{b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}{2\sigma_k^2} \right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C_t\, (2\pi)^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k\right)}{2}} \left(\prod_{k=1}^{K_t} \prod_{j=1}^{J_k} \left(b_{k_j} n_{k_j}\right)^{-\frac{1}{2}} \right) \\
&\quad \int_{\Omega_t} \prod_{k=1}^{K_t} \left(\sigma_k^2\right)^{-\left(\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}{2} + 1\right)} \exp\left(-\frac{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}{2\sigma_k^2} \right) \mathrm{d}\boldsymbol{\sigma}_t^2 \\
&= C_t\, \pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k\right)}{2}} \left(\prod_{k=1}^{K_t} \prod_{j=1}^{J_k} \left(b_{k_j} n_{k_j}\right)^{-\frac{1}{2}} \right) \\
&\quad \left(\prod_{k=1}^{K_t} \Gamma\left(\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}{2} \right) \left(\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2 \right)^{-\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k}{2}} \right) \\
&\quad \int_{\Omega_t} \prod_{k=1}^{K_t} \text{Inv-}\chi^2\left(\sigma_k^2 \,\middle|\, \left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j} \right) - J_k,\, \frac{\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j} - 1\right) s_{k_j}^2}{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right) - J_k} \right) \mathrm{d}\boldsymbol{\sigma}_t^2
\end{aligned}$$

$$= C_t\,\pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right)-J_k\right)}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k}\left(b_{k_j} n_{k_j}\right)^{-\frac{1}{2}}\right)$$
$$\left(\prod_{k=1}^{K_t}\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right)-J_k}{2}\right)\left(\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right)-J_k}{2}}\right)$$
$$P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right). \tag{32}$$

The expression for the numerator in Eq. (31) is identical to the final expression in Eq. (32) with all b's equal to 1, that is,

$$m_t^N(\boldsymbol{x}) = C_t\,\pi^{-\frac{\sum_{k=1}^{K_t}\left(\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k\right)}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} n_{k_j}^{-\frac{1}{2}}\right)$$
$$\left(\prod_{k=1}^{K_t}\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}\right)\left(\sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}}\right) \tag{33}$$
$$P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right),$$

where

$$P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right) = \int_{\Omega_t}\prod_{k=1}^{K_t}\text{Inv-}\chi^2\left(\sigma_k^2 \,\middle|\, \left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k, \frac{\sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2}{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}\right)\mathrm{d}\boldsymbol{\sigma}_t^2. \tag{34}$$

The final expression for the marginal likelihood in Eq. (31) is then given by

$$m_t^{\mathrm{GF}}(\boldsymbol{x},\boldsymbol{b}) = \frac{m_t^N(\boldsymbol{x})}{m_t^N(\boldsymbol{x}^{\boldsymbol{b}})}$$
$$= \frac{P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}\right)}{P^{\mathrm{GF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t | \boldsymbol{x}^{\boldsymbol{b}}\right)}\,\pi^{-\frac{\sum_{k=1}^{K_t}\sum_{j=1}^{J_k}\left(1-b_{k_j}\right)n_{k_j}}{2}} \left(\prod_{k=1}^{K_t}\prod_{j=1}^{J_k} b_{k_j}^{\frac{1}{2}}\right)$$
$$\prod_{k=1}^{K_t}\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}\right)\Gamma\left(\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right)-J_k}{2}\right)^{-1}$$
$$\left(\sum_{j=1}^{J_k}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{-\frac{\left(\sum_{j=1}^{J_k} n_{k_j}\right)-J_k}{2}} \left(\sum_{j=1}^{J_k} b_{k_j}\left(n_{k_j}-1\right)s_{k_j}^2\right)^{\frac{\left(\sum_{j=1}^{J_k} b_{k_j} n_{k_j}\right)-J_k}{2}}. \tag{35}$$

TABLE 4.
Overview of the population variances used in the simulation study with 6 populations.

Population	Effect	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2
Null	No	1.00	1.00	1.00	1.00	1.00	1.00
Order	Small	1.00	1.09	1.19	1.30	1.42	1.54
	Medium	1.00	1.22	1.48	1.80	2.19	2.66
	Large	1.00	1.33	1.78	2.38	3.17	4.23
Mixed	Small	1.00	1.00	1.00	1.33	1.33	1.33
	Medium	1.00	1.00	1.00	2.00	2.00	2.00
	Large	1.00	1.00	1.00	2.94	2.94	2.94
Near order	Small	1.09	1.00	1.19	1.30	1.42	1.54
	Medium	1.22	1.00	1.48	1.80	2.19	2.66
	Large	1.33	1.00	1.78	2.38	3.17	4.23
Reverse order	Small	1.54	1.42	1.30	1.19	1.09	1.00
	Medium	2.66	2.19	1.80	1.48	1.22	1.00
	Large	4.23	3.17	2.38	1.78	1.33	1.00

Appendix C: Computing the Probability That $\boldsymbol{\sigma}_t^2 \in \Omega_t$

The integrals in Eqs. (15), (20), (21), and (25) can be approximated numerically using the following Monte Carlo approach. For the BBF and the GFBF, we first sample $\sigma_k^{2(s)} \sim \text{Inv-}\chi^2\left(\nu_k, \tau_k^2\right)$, for $k = 1, \ldots, K_t$, where $\sigma_k^{2(s)}$ is the sth draw from Inv-$\chi^2\left(\nu_k, \tau_k^2\right)$, for $s = 1, \ldots, S$, and ν_k and τ_k^2 are as in Eqs. (15), (20), and (21), respectively. An approximation of the probability that the inequality constraints on the unique variances hold is then given by the proportion of draws that fall in Ω_t, that is,

$$P\left(\boldsymbol{\sigma}_t^2 \in \Omega_t\right) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^{2(s)}\right), \tag{36}$$

where $\boldsymbol{\sigma}_t^{2(s)} = \left[\sigma_1^{2(s)} \cdots \sigma_{K_t}^{2(s)}\right]^T$, and $\mathbf{1}_{\Omega_t}\left(\boldsymbol{\sigma}_t^{2(s)}\right)$ is the indicator function which is 1 if $\boldsymbol{\sigma}_t^{2(s)} \in \Omega_t$ and 0 otherwise.

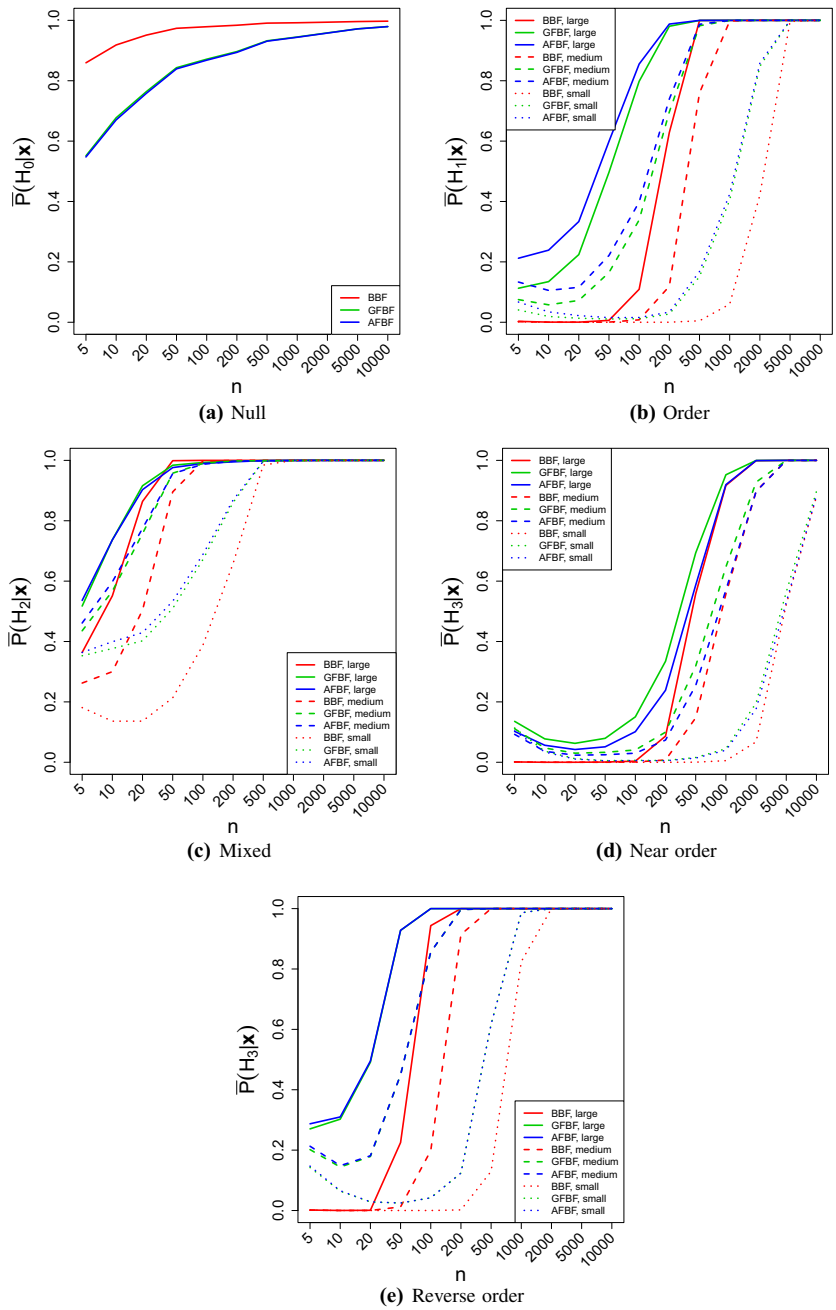
For the AFBF, let $\phi_k = a_k\sigma_k^2$. We then proceed analogously to the BBF and the GFBF: First, we sample $\phi_k^{(s)} \sim \text{Inv-}\chi^2\left(\nu_k, \tau_k^2\right)$, for $k = 1, \ldots, K_t$ and $s = 1, \ldots, S$, where ν_k and τ_k^2 are as in the second row of Eq. (25). Then

$$P^{\text{AF}}\left(\boldsymbol{\sigma}_t^2 \in \Omega_t^a \middle| \boldsymbol{x}^{\boldsymbol{b}}\right) = P^{\text{AF}}\left(\boldsymbol{\phi}_t \in \Omega_t \middle| \boldsymbol{x}^{\boldsymbol{b}}\right) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}_{\Omega_t}\left(\boldsymbol{\phi}_t^{(s)}\right), \tag{37}$$

where $\boldsymbol{\phi}_t = \left[\phi_1 \cdots \phi_{K_t}\right]^T$ and $\boldsymbol{\phi}_t^{(s)} = \left[\phi_1^{(s)} \cdots \phi_{K_t}^{(s)}\right]^T$.

Appendix D: Simulation Results for $J = 6$ Populations

In the simulation with 6 populations, we considered the same factors as in the simulation with 4 populations (cf. Sect. 6.5.1). First, we used the same patterns of the population variances: null



(a) Null **(b)** Order **(c)** Mixed **(d)** Near order **(e)** Reverse order

FIGURE 4.

Results of a simulation study comparing the performance of the three automatic Bayes factors in testing variances of 6 populations. We examined five different patterns of the population variances: **a** $\sigma_1^2 = \cdots = \sigma_6^2$, **b** $\sigma_1^2 < \cdots < \sigma_6^2$, **c** $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 < \sigma_4^2 = \sigma_5^2 = \sigma_6^2$, **d** $\sigma_2^2 < \sigma_1^2 < \sigma_3^2 < \cdots < \sigma_6^2$, and **e** $\sigma_6^2 < \cdots < \sigma_1^2$. In patterns **b** to **e** we considered three different sizes of the order effect: small, medium, and large. For each combination of pattern and effect size, we drew 1000 samples of size $n_1 = \cdots = n_6 = n$. In each sample we then tested four hypotheses: $H_0 : \sigma_1^2 = \cdots = \sigma_6^2$, $H_1 : \sigma_1^2 < \cdots < \sigma_6^2$, $H_2 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 < \sigma_4^2 = \sigma_5^2 = \sigma_6^2$, and $H_3 : \neg(H_0 \vee H_1 \vee H_2)$. Eventually, we computed the expected posterior probability of the true hypothesis $\bar{P}(H_t|\boldsymbol{x})$ across the 1000 samples. The plots show $\bar{P}(H_t|\boldsymbol{x})$ as a function of the common sample size n for the BBF (red lines), GFBF (green lines), and AFBF (blue lines) under a small effect (dotted lines), medium effect (dashed lines), and large effect (solid lines) (Color figure online).

($\sigma_1^2 = \cdots = \sigma_6^2$), order ($\sigma_1^2 < \cdots < \sigma_6^2$), mixed ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 < \sigma_4^2 = \sigma_5^2 = \sigma_6^2$), near order ($\sigma_2^2 < \sigma_1^2 < \sigma_3^2 < \cdots < \sigma_6^2$), and reverse order ($\sigma_6^2 < \cdots < \sigma_1^2$). Second, we again used the approach of Böing-Messing et al. (2017) to determine the population variances for a small, medium, and large effect. The resulting values of the variances are shown in Table 4. Note that the values for the variances in the mixed pattern are the same as in the simulation with 4 populations (cf. Table 1) because in both cases there are only two unique variances. Third, we used common sample sizes $n \in \{5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10{,}000\}$. The hypotheses we tested in each condition were analogous to those in the simulation with 4 populations (cf. Eq. (29)): $H_0\colon \sigma_1^2 = \cdots = \sigma_6^2$, $H_1\colon \sigma_1^2 < \cdots < \sigma_6^2$, $H_2\colon \sigma_1^2 = \sigma_2^2 = \sigma_3^2 < \sigma_4^2 = \sigma_5^2 = \sigma_6^2$, and $H_3\colon \neg(H_0 \vee H_1 \vee H_2)$. The results of the simulation with 6 populations are shown in Fig. 4. A notable difference between the results of the simulations with 4 and 6 populations is that under the near-order pattern with 6 populations even larger samples are needed to detect that the complement H_3 is true (cf. Figs. 2d and 4d). This is because the ratio of adjacent variances is smaller in the case of 6 populations (cf. Tables 1 and 4), which makes it more difficult for the Bayes factors to detect that the order of the first two population variances is reversed. Note that in Figs. 4a and e the lines for the GFBF and AFBF overlap to a large extent.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd international symposium on information theory* (pp. 267–281). Budapest: Akadémiai Kiadó.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*(1), 39–48.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*(4), 699–713.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*(3–4), 533–534.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402.

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*(446), 542–554.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In P. Lahiri (Ed.), *Model selection* (pp. 135–207). Beachwood, OH: Institute of Mathematical Statistics.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *P* values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122.

Böing-Messing, F., & Mulder, J. (2016). Automatic Bayes factors for testing variances of two independent normal distributions. *Journal of Mathematical Psychology*, *72*, 158–170.

Böing-Messing, F., van Assen, M. A. L. M., Hofman, A. D., Hoijtink, H., & Mulder, J. (2017). Bayesian evaluation of constrained hypotheses on variances of multiple independent groups. *Psychological Methods*, *22*(2), 262–287.

Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics*, *59*(2), 211–220.

De Santis, F., & Spezzaferri, F. (2001). Consistent fractional Bayes factor for nested normal linear models. *Journal of Statistical Planning and Inference*, *97*(2), 305–321.

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*(4), 979–1006.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gilks, W. R. (1995). Discussion of O'Hagan. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 118–120.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165.

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

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*(1), 64–72.

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

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., et al. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review*, *33*(6), 795–811.

Lehre, A.-C., Lehre, K. P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology*, *51*(2), 198–206.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.
Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1–2), 187–192.
Lucas, J. W. (2003). Status processes and the institutionalization of women as leaders. *American Sociological Review*, *68*(3), 464–480.
Mulder, J. (2014a). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171.
Mulder, J. (2014b). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463.
Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.
Mulder, J., & Fox, J.-P. (2013). Bayesian tests on components of the compound symmetry covariance matrix. *Statistics and Computing*, *23*(1), 109–122.
Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(2), 1–39.
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.
Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*(6), 530–546.
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.
O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 99–138.
O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, *6*(1), 101–118.
Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology*, *8*(1), 1–11.
Russell, V. A., Oades, R. D., Tannock, R., Killeen, P. R., Auerbach, J. G., Johansen, E. B., et al. (2006). Response variability in attention-deficit/hyperactivity disorder: A neuronal and glial energetics hypothesis. *Behavioral and Brain Functions*, *2*(1), 1–25.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
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.
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.
Spiegelhalter, D. J., & Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(3), 377–387.
Verhagen, A. J., & Fox, J.-P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.
Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics*, *51*(2), 589–599.
Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). Amsterdam, The Netherlands: Elsevier.

Manuscript Received: 10 AUG 2016
Final Version Received: 28 MAR 2018