Bayesian Analysis (2024) **19**, Number 3, pp. 699–723

Default Bayes Factors for Testing the (In)equality of Several Population Variances*

Fabian Dablander[†,**], Don van den Bergh[‡,**], Eric-Jan Wagenmakers[§], and Alexander Ly[¶,‖]

Abstract. Testing the (in)equality of variances is an important problem in many statistical applications. We develop default Bayes factor tests to assess the (in)equality of two or more population variances, as well as a test for whether a population variance equals a specific value. The resulting test can be used to check assumptions for commonly used procedures such as the t-test or ANOVA, or test substantive hypotheses concerning variances directly. We show that our Bayes factor fulfills a number of desiderata. Researchers may have directed hypotheses such as $\sigma_1^2 > \sigma_2^2$, they may want to extend $\mathcal{H}_0$ to have a null-region, or wish to combine hypotheses about equality with hypotheses about inequality, for example $\sigma_1^2 = \sigma_2^2 > (\sigma_3^2, \sigma_4^2)$. We extend our Bayes factor test to allow for these deviations from our proposed default and illustrate it on a number of practical examples. Our procedure is implemented in the R package *bfvartest*.

Keywords: Bayes factors, model selection, comparing variances.

MSC2020 subject classifications: 62F03, 62F15.

1 Introduction

Testing the (in)equality of variances is important in many sciences and applied contexts. In engineering, for example, researchers may want to assess whether a new, cheaper measurement instrument achieves the same precision as the gold standard (Sholts et al., 2011). In genetics and medicine, scientists are not only interested in studying the genetic effect on the mean of a quantitative trait, but also on its variance (Paré et al., 2010). In economics and archeology, ideas such as that increased economic production should reduce variability in products directly lead to statistical hypotheses on variances (Kvamme et al., 1996). In a court of law, one may be interested in reducing unwanted variability in civil damage awards and may want to compare how different interventions reduce this variability (Saks et al., 1997). In psychology, educational researchers may be

arXiv: 2003.06278

*FD, DvB, EJW, and AL were supported by a Vici grant no. C.2523.0278.01.

†Department of Psychological Methods, University of Amsterdam, The Netherlands, dablander.fabian@gmail.com

‡Department of Psychological Methods, University of Amsterdam, The Netherlands, donvdbergh@hotmail.com

§Department of Psychological Methods, University of Amsterdam, The Netherlands, ej.wagenmakers@gmail.com

¶Department of Psychological Methods, University of Amsterdam, The Netherlands

‖Centrum Wiskunde & Informatica, The Netherlands, alexander.ly.nl@gmail.com

**These authors share first authorship.

 https://doi.org/10.1214/23-BA1369

interested in studying how the variance in pupil's mathematical ability changes across school grades (Aunola et al., 2004).

While there exist several classical p-value tests for assessing the (in)equality of population variances (e.g., Levene, 1961; Brown and Forsythe, 1974; Gastwirth et al., 2009), testing such hypotheses has received little attention from a Bayesian perspective. Such a perspective, however, would offer practitioners the possibility to (a) quantify evidence in favor of the null hypothesis (e.g., Morey et al., 2016), (b) allow one to incorporate prior knowledge (e.g., O'Hagan et al., 2006), (c) use sequential sampling designs which in many cases is more cost-effective (e.g., than a fixed-N design, see Stefan et al., 2019), and (d) translate substantive predictions more easily into statistical hypotheses by specifying equality and inequality constraints (e.g., Böing-Messing and Mulder, 2018; Hoijtink et al., 2008).

In light of these benefits and recent recommendations to go beyond p-value testing (Wasserstein and Lazar, 2016), we develop default Bayes factor tests (e.g., Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a,b) for the (in)equality of several population variances. Our work is inspired by Jeffreys (1939, pp. 222-224), who developed a test for the "agreement of two standard errors". Equipped with our procedure, researchers are able to state graded evidence both for the case of testing assumptions of other tests (e.g., the equality of variances assumption in the Student's t-test), as well as testing order-constrained hypotheses on variances directly.

This paper is structured as follows. In Section 2, we introduce the problem setup and propose the default Bayes factor. In Section 3, we elaborate on the desiderata that the proposed Bayes factor adheres to. In Section 4, we discuss the special case with $K = 2$ groups, including directed and interval Bayes factors, compare our method to a fractional Bayes factor procedure proposed by Böing-Messing and Mulder (2018), and discuss testing all possible (in)equalities at once. We illustrate our default Bayes factor test and deviations from it on a number of practical examples in Section 5. We conclude in Section 6. All derivations and proofs can be found in the supplementary materials (Dablander et al., 2023).

2 Default Bayes Factor for K Groups

2.1 Notation and Problem Setup

The problem of testing the (in)equality of variances can be equivalently expressed in terms of variances σ_j^2 or precisions $\tau_j = \sigma_j^{-2}$. For the data we assume that $Y_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \tau_j^{-1})$, where $i \in [n_j]$ and $j \in [K]$ with the rectangular brackets embracing an integer denoting the set of positive integers up to and including that integer, e.g., $[K] := \{1, 2, \ldots, K-1, K\} \subset \mathbb{N}$.

As the K groups are assumed to be independent of each other, the data $y^{[K]}$ can be sufficiently summarized by the sample means $\boldsymbol{\bar{y}} = (\bar{y}_1, \ldots, \bar{y}_K)$, where $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}$ and the (unbiased) sample variances $\boldsymbol{s^2} = (s_1^2, \ldots, s_K^2)$, where $s_j^2 = \frac{1}{\nu_j}\sum_{i=1}^{n}(y_{ji} - \bar{y}_j)^2$ and where $\nu_j = n_j - 1$ is the degree of freedom of group j. As a convention, we denote

K-dimensional vectors in bold, whereas an arrow is used to denote a $K-1$ dimensional vector, e.g., $\boldsymbol{s^2} = (\vec{s^2}, s_K^2)$. A subscript $+$ is used to denote summation over the vector's elements, e.g., $\boldsymbol{\tau}_+ = \sum_{j=1}^{K} \tau_j$, whereas $\vec{\vartheta}_+ = \sum_{j=1}^{K-1} \vartheta_j$, since $\vec{\vartheta} \in \mathbb{R}^{K-1}$.

The null hypothesis $\mathcal{H}_0$ states that all precisions are the same, while the alternative hypothesis $\mathcal{H}_1$ includes at least one inequality. Formally, we compare

$$\mathcal{H}_0 : \tau_j = \tau_k \text{ for all } j, k \in [K], \tag{2.1}$$

$$\mathcal{H}_1 : \tau_j \neq \tau_k \text{ for some } j \neq k \in [K], \tag{2.2}$$

regardless of the nuisance parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K) \in \mathbb{R}^K$. The null hypothesis restricts the K precisions to a single but unknown precision, whereas the alternative allows all precisions to vary freely. Including the means, the null model has $K+1$ free parameters, whereas the alternative model has $2K$ free parameters.

We rephrase the model comparison by generalizing the reparametrization proposed by Jeffreys (1939, pp. 222-224); see also the supplementary materials. More specifically, in the alternative model we reparametrize the K precisions $\boldsymbol{\tau}$ in terms of an average precision $\bar{\boldsymbol{\tau}} = \frac{1}{K}\boldsymbol{\tau}_+$ and $K-1$ proportions $\vec{\vartheta}$ with $\vartheta_j = \frac{\tau_j}{\boldsymbol{\tau}_+}$. Note that this reparametrization is invertible as it should be. In this parametrization the hypotheses translate into

$$\mathcal{H}_0 : \vartheta_j = \tfrac{1}{K} \text{ for all } j \in [K-1], \tag{2.3}$$

$$\mathcal{H}_1 : \vartheta_j \neq \tfrac{1}{K} \text{ for some } j \in [K-1], \tag{2.4}$$

regardless of the values of the nuisance parameter $\boldsymbol{\mu} \in \mathbb{R}^K$ and the average precision $\bar{\boldsymbol{\tau}} > 0$, which are common to both models.

From a Bayesian perspective, we assess the relative merits of $\mathcal{H}_0$ and $\mathcal{H}_1$ by virtue of how well they predict the data, that is, by their respective marginal likelihoods. The ratio of marginal likelihoods is known as the Bayes factor (Kass and Raftery, 1995), and its specification requires assigning priors to both the free parameters of the null and the alternative model. For the models being compared this implies one prior on the $2K$ free parameters of the alternative model, and another prior on the $K+1$ free parameters of the null model. To simplify matters, we mimic the nesting of the null model into the alternative model and choose $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$. The Bayes factor we propose is constructed from a right Haar prior $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ on the common parameters and from a (proper) Dirichlet prior $\pi_1(\vec{\vartheta})$ on the test-relevant parameters $\vec{\vartheta}$ with hyperparameters $\boldsymbol{u}$, where $u_j > 0$ for all $j \in [K]$.

In the remainder of this section we show that this choice of priors results in a Bayes factor that is analytic. In Section 3 we show that the proposed Bayes factor fulfills certain Bayesian model comparison desiderata.

2.2 The Proposed Bayes Factor

The choice for $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ is based on the observation that the hypotheses to be tested are invariant under (1) scalar multiplications of all the data points, and (2) location shifts of the data points of each sample/group. The nesting $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$ makes the use of the improper priors $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ permissible as a limit of proper priors with normalization constants canceling due to their appearances in both the numerator and denominator of the Bayes factor (see also Hendriksen et al., 2021; Ly et al., 2016b; Robert, 2016). The derivations in the supplementary materials show that with $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ on the nuisance parameters, the Bayes factor simplifies to

$$\mathrm{BF}_{10}(y^{[K]}) = \frac{\int\limits_{\Theta} \left(\int\limits_{\mathbb{R}_{>0}} \int\limits_{\mathbb{R}^K} f(y^{[K]} \,|\, \boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\bar{\boldsymbol{\tau}} \right) \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}}{\int\limits_{\mathbb{R}_{>0}} \int\limits_{\mathbb{R}^K} f(y^{[K]} \,|\, \boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta} = \frac{1}{K}) \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\bar{\boldsymbol{\tau}}} = \int_{\Theta} h(\boldsymbol{s^2} \,|\, \vec{\vartheta}) \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}, \tag{2.5}$$

where $\mathbb{R}_{>0}$ denotes the positive reals, $\Theta := \{\vec{\theta} \in \mathbb{R}^{K-1} \,|\, \vec{\theta}_+ < 1\} \subset \mathbb{R}_{>0}^{K-1}$, and where we refer to $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ as the reduced likelihood, which is given by

$$h(\boldsymbol{s^2} \,|\, \vec{\vartheta}) := \Big(1 + \sum_{j=1}^{K-1} \tfrac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} \Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1 - \vec{\vartheta}_+)^{\frac{\nu_K}{2}} \Big(1 - \sum_{j=1}^{K-1} [1 - \tfrac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}, \tag{2.6}$$

where $\boldsymbol{\nu}_+ = \sum_{j=1}^{K} \nu_j$, and $\vec{\vartheta}_+ := \sum_{j=1}^{K-1} \vartheta_j$. Note that, for any proper prior $\pi_1(\vec{\vartheta})$, the nesting and the choice $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ leads to a measurement invariant Bayes factor, as desired. This is because $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ and therefore $\mathrm{BF}_{10}(y^{[K]}) = \mathrm{BF}_{10}(\boldsymbol{s^2})$ only depend on the data via the ratios of sums of squares $\frac{\nu_j s_j^2}{\nu_K s_K^2}$, and because each s_k^2 is invariant under location shifts within sample/group k.

The Dirichlet prior $\pi_1(\vec{\vartheta})$ on the test-relevant parameters is inspired by the form of $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ and makes the proposed Bayes factor analytic. By definition of the integral form of the type D Lauricella function, the proposed Bayes factor is

$$\mathrm{BF}_{10}(\boldsymbol{s^2}) = \frac{\mathcal{B}(\frac{\boldsymbol{\nu}}{2} + \boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} \Big(1 + \sum_{j=1}^{K-1} \tfrac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} F_D\Big(\tfrac{\boldsymbol{\nu}_+}{2} \,;\, \tfrac{\vec{\nu}}{2} + \vec{u} \,;\, \tfrac{\boldsymbol{\nu}_+}{2} + \boldsymbol{u}_+ \,;\, \vec{1} - \tfrac{\overrightarrow{\nu s^2}}{\nu_K s_K^2}\Big), \tag{2.7}$$

where $\mathcal{B}(\boldsymbol{u}) = \frac{\Gamma(u_1) \cdots \Gamma(u_K)}{\Gamma(u_+)}$ is the multivariate beta function, $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^{K-1}$, $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1} s_{K-1}^2)$ is the $K-1$ vector of sums of squares, and where F_D is a type D Lauricella function which has the integral representation $F_D(a \,;\, \vec{b} \,;\, d \,;\, \vec{x}) = \frac{\Gamma(d)}{\Gamma(a)\Gamma(d-a)} \int_0^1 t^{a-1}(1-t)^{d-a-1}(1-x_1 t)^{-b_1} \cdots (1-x_{K-1}t)^{-b_{K-1}} \mathrm{d}t$ whenever $d > a$, which holds trivially since $u > 0$ always. Observe that, with Equation (2.7) at hand, we also have an analytic marginal posterior for $\vec{\vartheta}$, namely,

$$\pi_1(\vec{\vartheta} \,|\, y^{[K]}) = \frac{\Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1 - \vec{\vartheta}_+)^{\frac{\nu_K}{2}} \Big(1 - \sum_{j=1}^{K-1} [1 - \frac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}}{\mathcal{B}(\frac{\boldsymbol{\nu}}{2} + \boldsymbol{u}) F_D\Big(\frac{\boldsymbol{\nu}_+}{2} \,;\, \frac{\vec{\nu}}{2} + \vec{u} \,;\, \frac{\boldsymbol{\nu}_+}{2} + \boldsymbol{u}_+ \,;\, \vec{1} - \frac{\overrightarrow{\nu s^2}}{\nu_K s_K^2}\Big)}. \tag{2.8}$$

The proposed Bayes factor can be computed from the sample variances and sample sizes directly. This makes it possible to re-evaluate the published literature without the need to have access to the raw data, as shown in Section 5. In the next section, we show that the proposed Bayes factor fulfills a number of desiderata; all proofs can be found in the supplementary materials.

3 Properties of the Proposed Bayes Factor

An important result of this paper is that our proposed Bayes factor fulfills a number of desiderata (Bayarri et al., 2012; Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a,b). More specifically, we show that the proposed Bayes factor has the finite-sample properties of being (i) labelling invariant, (ii) (exactly) predictively matched, and (iii) information consistent. It also has the asymptotic properties of being (iv) model selection consistent and (v) limit and across-sample consistent. Information consistency requires $u_j \leq 1/2$ for $j \in [K]$ while labelling invariance requires $u_i = u_j$ for all $i, j \in [K]$, suggesting the default choice of $u_j = 1/2$ for all $j \in [K]$.[1]

3.1 Labelling Invariance

A Bayes factor is labelling invariant if it is independent of the arbitrary choice of which group is labelled K.

Theorem 3.1 (Labelling invariance)**.** *The proposed Bayes factor with $u_i = u_j$ for all $i, j \in [K]$ is labelling invariant.* $\diamond$

3.2 Predictive Matching

A Bayes factor is (exactly) predictively matched if it equals 1 for all data sets of insufficient size, that is, $\text{BF}_{10}(y^{[K]}) = 1$ for all $y^{[K]}$ with $\boldsymbol{n} = (n_1, \ldots, n_K)$ smaller than the minimal sample sizes (Bayarri et al., 2012). The insufficient sizes are: (a) $n_1 = \ldots = n_K = 1$ as then $\nu_j s_j^2 = 0$ for all $j \in [K]$ regardless of the observations, and (b) $n_k = 2$ for some $k \in [K]$ and $n_j = 1$ for all $j \in [K] \setminus \{k\}$, in which case there is no other sample variance to compare s_k^2 to.

Theorem 3.2 (Predictive matching)**.** *A Bayes factor constructed from the pair of priors $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$ and $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ with $\pi_1(\vec{\vartheta})$ proper is predictively matched. This holds for our proposed Bayes factor.* $\diamond$

3.3 Information Consistency

Information consistency implies that for all data sets of sufficient size, that is, fixed $\boldsymbol{n} = (n_1, \ldots, n_K)$ with at least two indexes $j \neq k \in [K]$ such that $n_j, n_k \geq 2$, the

[1]Values $0 < u < 1/2$ would also fulfill all desiderata, but would put even more mass on large differences between the variances; we therefore use $u = 1/2$ as our default choice.

Bayes factor in favor of the alternative over the null should tend to infinity whenever it becomes abundantly clear that the null cannot hold true. This occurs in the limit $s_j^2/s_K^2 \to 0$, that is, when the observed variance s_K^2 is of a much higher order than another sample variance s_j^2.

Theorem 3.3 (Information consistency)**.** *The proposed Bayes factor is information consistent if $u_j \leq 1/2$ for $j \in [K]$.* $\diamond$

3.4 Model Selection Consistency

A Bayes factor is model selection consistent if it selects the correct model as $\boldsymbol{n} \to \infty$, that is, if

$$\mathrm{BF}_{10}(Y^{[K]}, \boldsymbol{n}) \xrightarrow{\mathbb{P}} 0 \text{ if } \mathbb{P} \in \mathcal{M}_0, \text{ and } \mathrm{BF}_{01}(Y^{[K]}, \boldsymbol{n}) \xrightarrow{\mathbb{P}} 0 \text{ if } \mathbb{P} \in \mathcal{M}_1, \tag{3.1}$$

where $\mathbb{P}$ refers to the data generating distribution, and where $X_n \xrightarrow{\mathbb{P}} X$ denotes convergence in probability, that is, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$ for all $\epsilon > 0$.

To state the theorem and to allow the K sample sizes go to infinity independently of each other, we let $n_K := n$ and $n_j := c_j n$ for $c_j > 0$, $j \in [K]$, thus, $c_K = 1$ by definition. To also allow the (data-governing) variances to differ arbitrarily as well, we let γ_j be the relative size of the variance σ_j^2 with respect to σ_K^2, that is, $\sigma_j^2 := \gamma_j \sigma_K^2$ where $\gamma_j > 0$ for $j \in [K]$, thus, $\gamma_K = 1$ by definition. Note that the null hypothesis is equivalent to $\boldsymbol{\gamma} = \mathbf{1} \in \mathbb{R}^K$, whereas under the alternative there exists at least one $j \in [K]$ such that $\gamma_j \neq 1$.

Theorem 3.4 (Model selection consistency)**.** *The proposed Bayes factor is model selection consistent. Furthermore, let $Y_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \sigma_j^2)$ where $\sigma_j^2 = \gamma_j \sigma_K^2$ for $i \in [n_j]$, $n_j = c_j n$, and $n_K = n$ for $j \in [K]$, then as all the sample sizes tend to infinity, the Bayes factor behaves as*

$$\mathrm{BF}_{10}(\boldsymbol{s}^2, n) = C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma}) n^{\frac{1-K}{2}} \left(\tfrac{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle}{\boldsymbol{c}_+}\right)^{\frac{\boldsymbol{c}_+}{2} n} \Big(\prod_{j=1}^{K-1} \gamma_j^{-\frac{c_j}{2} n} \Big) \exp(V(n)), \tag{3.2}$$

where $\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle := \sum_{j=1}^K c_j \gamma_j$, $V(n) = \mathcal{O}_P(n^{-1/2})$ under the null and $V(n) = \mathcal{O}_P(n^{1/2})$ under the alternative, and where

$$C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma}) = \frac{(4\pi)^{\frac{K-1}{2}} \boldsymbol{c}_+^{\frac{1}{2}} \Big(\prod_{j=1}^{K-1} \gamma_j^{-u_j} \Big)}{\mathcal{B}(\boldsymbol{u}) \Big(\prod_{j=1}^{K-1} c_j^{\frac{1}{2}} \Big) (\boldsymbol{c}_+ - \sum_{j=1}^{K-1} \frac{c_j \gamma_j - 1}{\gamma_j})^{\boldsymbol{u}_+}}. \tag{3.3}$$

This means that under the alternative, $\mathcal{H}_1 : \gamma_j \neq 1$ for some $j \in [K-1]$, we have that

$$\begin{aligned} \log(\mathrm{BF}_{10}(\boldsymbol{s}^2, n)) &= \log\big(C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma})\big) + \tfrac{1-K}{2} \log(n) \\ &\quad + \Big(\boldsymbol{c}_+ \log\big(\tfrac{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle}{\boldsymbol{c}_+}\big) - \sum_{j=1}^{K-1} c_j \log(\gamma_j) \Big) \frac{n}{2} + \mathcal{O}_P(n^{1/2}). \end{aligned} \tag{3.4}$$

Under the null, $\mathcal{H}_0 : \vec{\gamma} = \vec{1}$, this simplifies drastically, and the logarithm of the Bayes factor then behaves as

$$\log(\mathrm{BF}_{10}(\boldsymbol{s}^2, n)) = \tfrac{1-K}{2}\Big(\log(n) - \log(4\pi)\Big) + \tfrac{1}{2}\Big(\log(\boldsymbol{c}_+) - \sum_{j=1}^{K-1}\log(c_j)\Big)$$
$$- \boldsymbol{u}_+ \log(K) - \log \mathcal{B}(\boldsymbol{u}) + \mathcal{O}_P(n^{-1/2}). \tag{3.5}$$

Hence, $\mathrm{BF}_{10}(\boldsymbol{s}^2, n)$ *converges relatively slowly to zero under the null compared to the exponential decay of* $\mathrm{BF}_{01}(\boldsymbol{s}^2, n)$ *under the alternative.* ◇

Illustrating the Rate of Convergence

We illustrate the rate of convergence of our default Bayes factor by visualizing Equations (3.4) and (3.5) as a function of $K \in [2, 12]$ and $\gamma_1 \in [2, \ldots, 11]$ with $\gamma_2 = \ldots = \gamma_K = 1$ and $\sigma_K^2 = 1$. Equation (3.4) shows that under the alternative the asymptotic behavior of $\log(\mathrm{BF}_{10})$ is mostly linear in n. The left panel in Figure 1 shows the slope of this linear increase — termed the log Bayes factor growth — as a function of K and γ_1. We arrive at this slope by computing Equation (3.4) for a large number of n and regressing the result on n. When $\mathcal{H}_1$ is true, the rate of convergence of the Bayes factor is exponential, and so the log Bayes factor grows linearly. We visualize the slope of how the log Bayes factor grows across the number of groups, with larger values indicating more rapid exponential growth. We find that, as the number of groups increases, the log Bayes factor grows more quickly. This increase is also dependent on γ_1; for larger values, the Bayes factor grows more quickly with increasing number of groups.

The right panel in Figure 1 illustrates $\log(\mathrm{BF}_{01})$ as a function of the sample size per group for different number of groups K under the null hypothesis, using Equation (3.5). In contrast to the scenario when $\mathcal{H}_1$ is true, the rate of convergence when $\mathcal{H}_0$ is true is no longer exponential (see also Johnson and Rossell, 2010; Jeffreys, 1961; Bahadur and Bickel, 2009).

3.5 Limit and Across-Sample Consistency

A Bayes factor is limit consistent if it remains bounded as long as not all $n_j \to \infty$ for $j \in [K]$ (Ly, 2018, Ch. 6). A Bayes factor is across-sample consistent if the limit of the K-sample Bayes factor as a function of the fixed observations of the groups $i \in [K-1]$ results in a $K-1$ sample Bayes factor (Peña, 2018, Ch. 4). Note that we can consider without loss of generality the situation where the first $K-1$ samples are fixed as $n_K \to \infty$ because of labelling invariance. For the following, we assume that S_K^2 is a $\sqrt{n_K}$-consistent estimator for the data-governing variance σ_0^2 of the Kth group, which by Chebyshev's inequality is certainly the case when $Y_{Ki} \sim \mathcal{N}(\mu_K, \sigma_0^2)$.

We call the K-sample Bayes factor $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2)$ *across-sample consistent* if, as $n_K \to \infty$, it converges in probability under σ_0^{-2} to a $K-1$ Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(y^{[K-1]})$, comparing the hypotheses

$$\mathcal{H}_{0\,;\,\sigma_0^2}^{[K-1]} : \tau_j = \sigma_0^{-2} \text{ for all } j \in [K-1], \tag{3.6}$$

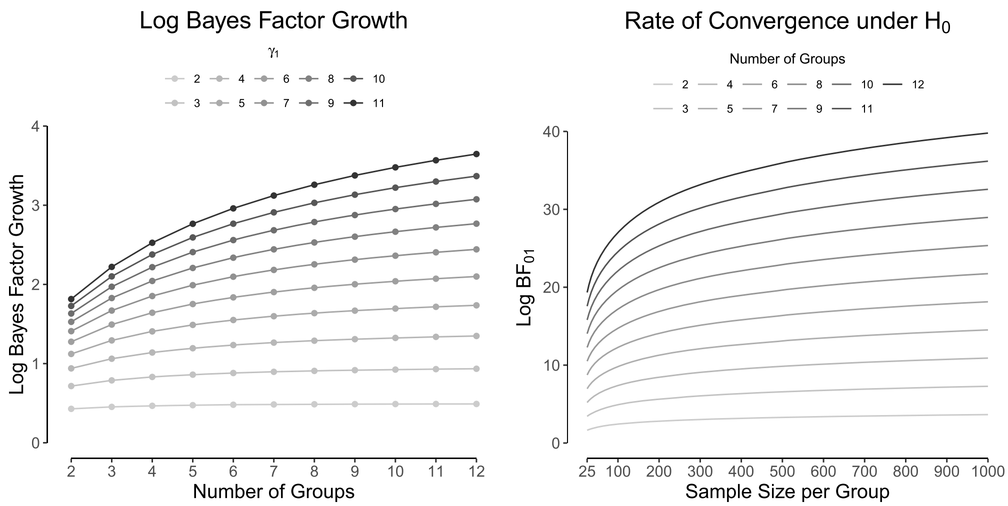

Figure 1: Left: Shows the rate of the linear growth of the log Bayes factor under $\mathcal{H}_1$ for increasing γ_1 and number of groups. Right: Shows how $\log(\mathrm{BF}_{01})$ grows as a function of n when $\mathcal{H}_0$ is true for different number of groups K. All Bayes factors were computed with the default value $u = 1/2$.

$$\mathcal{H}_{1\,;\,\sigma_0^2}^{[K-1]} : \tau_j \neq \sigma_0^{-2} \text{ for some } j \in [K-1]. \tag{3.7}$$

Here the null hypothesis states that the $K-1$ precisions are all equal to the known constant σ_0^{-2}, whereas the alternative states that at least one precision is unequal to σ_0^{-2}.

The theorem below implies that the proposed Bayes factor converges in probability to a lower dimensional Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ that is based on uniform priors on the nuisance parameters $\vec{\mu} \in \mathbb{R}^{K-1}$, and an inverse Dirichlet distribution on the precisions $\vec{\tau} = (\tau_1, \ldots, \tau_{K-1}) \in \mathbb{R}^{K-1}$ scaled by $1/\sigma_0^{-2}$, that is,

$$\pi_{\sigma_0^2}(\vec{\tau} \,|\, \mathcal{M}_1^{[K-1]}) = \frac{(\sigma_0^2)^{K-1} \prod_{j=1}^{K-1} (\sigma_0^2 \tau_j)^{u_j - 1}}{\mathcal{B}(\vec{u}, w)(1 + \sigma_0^2 \vec{\tau}_+)^{\vec{u}_+ + w}}, \tag{3.8}$$

where we wrote $w = u_K$ so the statement only involves vectors of length $K-1$. The integral representation of the multivariable generalisation of Tricomi's confluent hypergeometric function of the second kind $\mathcal{U}$, see for instance (Ng et al., 2011; Phillips, 1988), shows that the resulting $K-1$ sample Bayes factor is given by

$$\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2}) = \frac{\int \left(\prod_{j=1}^{K-1} \tau_j^{\frac{\nu_j}{2}} \right) \exp(-\frac{1}{2} \sum_{j=1}^{K-1} \nu_j s_j^2 \tau_j) \pi_{\sigma_0^2}(\vec{\tau} \,|\, \mathcal{M}_1^{[K-1]}) \mathrm{d}\vec{\tau}}{(\sigma_0^2)^{-\frac{\vec{\nu}_+}{2}} \exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})},$$

$$= \frac{\left(\prod_{j=1}^{K-1}\Gamma(\frac{\nu_j}{2}+u_j)\right)\mathcal{U}\left(\frac{\vec{\nu}}{2}+\vec{u}\,;\,\frac{\vec{\nu}_+}{2}-u_K+1\,;\,\frac{\overrightarrow{\nu s^2}}{2\sigma_0^2}\right)}{\mathcal{B}(\vec{u},w)\exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})}, \tag{3.9}$$

where $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1}s_{K-1}^2)$ denotes the vector of sums of squares, $(\overrightarrow{\nu s^2})_+ = \sum_{j=1}^{K-1}\nu_j s_j^2$, and $\vec{\nu}_+ := \sum_{j=1}^{K-1}\nu_j$, as before.

Theorem 3.5 (Limit and Across-Sample $\sqrt{n}_K$-consistency)**.** *If S_K^2 is an $\sqrt{n_K}$-consistent estimator for σ_0^2, then the Bayes factor $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2)$ is a $\sqrt{n}_K$-consistent estimator of the $K-1$-sample Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ given in Equation (3.9). Furthermore, if $Y_{Ki} \sim \mathcal{N}(\mu_K, \sigma_0^2)$, then $\sqrt{n}_K(S_K^2 - \sigma_0^2)$ is asymptotically normal, and consequently so is the K-sample Bayes factor, that is,*

$$\sqrt{n_K}\left(\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2) - \mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 2\sigma_0^4\breve{T}_1^2\right), \tag{3.10}$$

where $\breve{T}_1$ is given in the supplementary materials. ◇

4 Special Cases, Deviations from the Default, and Multiple Comparisons

The comparison of $K = 2$ groups occurs frequently in practice and we discuss the Bayes factor for this special case in the following section. We also consider three modifications of the default choice in order to incorporate a subject assessment of the test-relevant parameter, and to accommodate directed tests and interval Bayes factors. Lastly, we also consider the problem of testing all possible (in)equalities, that is, the multiple comparisons problem.

4.1 The Bayes Factor for $K = 2$ Groups

For the $K = 2$ group case, the null model of equal precisions has three parameters $(\mu_1, \mu_2, \bar{\tau})$ whereas the alternative has four $(\mu_1, \mu_2, \bar{\tau}, \vartheta)$. The comparison of interest is then between $\mathcal{H}_0 : \vartheta = \frac{1}{2}$ and $\mathcal{H}_1 : \vartheta \neq \frac{1}{2}$. In this case, the proposed Bayes factor simplifies to

$$\mathrm{BF}_{10}(\boldsymbol{s^2}) = \frac{\mathcal{B}(\frac{\nu_1}{2}+u_1, \frac{\nu_2}{2}+u_2)}{\mathcal{B}(u_1,u_2)}\left(1+\frac{\nu_1 s_1^2}{\nu_2 s_2^2}\right)^{\frac{\nu_1+\nu_2}{2}} {}_2F_1\left(\frac{\nu_1+\nu_2}{2}, \frac{\nu_1+2u_1}{2}; \frac{\nu_1+\nu_2+2(u_1+u_2)}{2}; \frac{\nu_2 s_2^2-\nu_1 s_1^2}{\nu_2 s_2^2}\right), \tag{4.1}$$

where ${}_2F_1$ refers to the Gaussian or ordinary hypergeometric function, which has the integral representation ${}_2F_1(a,b\,;\,c\,;\,z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\mathrm{d}t$, with $\mathrm{Re}(c) > \mathrm{Re}(b) > 0$ (Abramowitz and Stegun, 1972, eq. 15.3.1). Observe that across-sample consistency implies that for $Y_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_0^2)$ and $n_2 \to \infty$, the two-sample Bayes

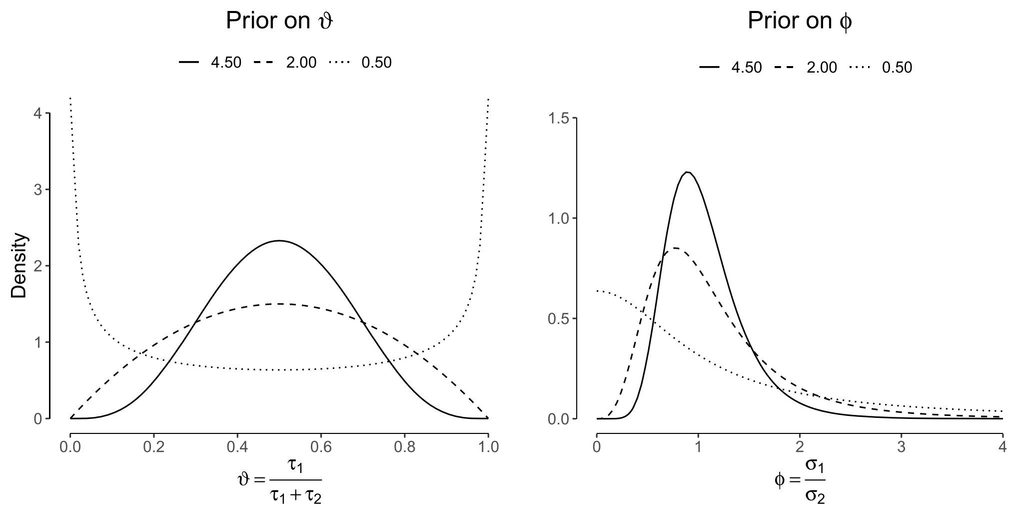

Figure 2: Prior on ϑ (left) and induced prior on ϕ (right) for $u := u_1 = u_2 \in \{4.50, 2.00, 0.50\}$; see Section 4.2 for the rationale behind these values.

factor is a $\sqrt{n}_2$-consistent estimator of the one-sample Bayes factor

$$\mathrm{BF}_{10\,;\,\sigma_0^2}^{[1]}(s_1^2) = \frac{\Gamma(\frac{\nu_1}{2} + u_1)\mathcal{U}\left(\frac{\nu_1}{2} + u_1\,;\,\frac{\nu_1}{2} - u_2 + 1\,;\,\frac{\nu_1 s_1^2}{2\sigma_0^2}\right)}{\mathcal{B}(u_1, u_2)\exp(-\frac{\nu_1 s_1^2}{2\sigma_0^2})}. \tag{4.2}$$

This Bayes factor compares the alternative hypothesis $\mathcal{H}_{1\,;\,\sigma_0^2}^{[1]} : \tau_1 \neq \sigma_0^{-2}$ to the null hypothesis $\mathcal{H}_{0\,;\,\sigma_0^2}^{[1])} : \tau_1 = \sigma_0^{-2}$ with σ_0^2 known. Here $\mathcal{U}(a\,;\,b\,;\,z) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-zt}t^{a-1}(1 + t)^{b-a-1}\mathrm{d}t$ is the (one-dimensional) Tricomi's confluent hypergeometric function of the second kind (Abramowitz and Stegun, 1972, Eq. 13.2.5).

4.2 Prior elicitation for $K = 2$ groups

For prior elicitation, it is arguably more intuitive to express the prior on the test-relevant parameter in terms of the ratio of the standard deviations, $\phi = \frac{\sigma_2}{\sigma_1} = \sqrt{\frac{\vartheta}{1-\vartheta}}$, thus, $\int_0^1 \mathrm{d}\vartheta = \int_0^\infty 2\phi(1 + \phi^2)^{-2}\mathrm{d}\phi$. The prior $\vartheta \sim \mathrm{Beta}(u_1, u_2)$ underlying Equation (4.1) induces a generalized beta prime distribution on ϕ with density

$$\pi(\phi\,;\,u_1, u_2) = \frac{2\phi^{2u_1-1}(1 + \phi^2)^{-(u_1+u_2)}}{\mathcal{B}(u_1, u_2)}. \tag{4.3}$$

Figure 2 visualizes the prior assigned to ϑ and ϕ for various values of $u := u_1 = u_2$. A statistician may now elicit a researcher's prior beliefs in terms of (a ratio of) standard deviations conditional on the alternative holding true. For example, if the researcher believes that the probability of one standard deviation being twice as large or twice as small as the other does not exceed 95%, then she should choose $u = 4.50$. Note that the resulting Bayes factor is not information consistent anymore. It is also interesting to note that on this scale ϕ the mth raw moment is given by $\frac{\Gamma(\frac{m}{2}+u_1)\Gamma(u_2-\frac{m}{2})}{\Gamma(u_1)\Gamma(u_2)}$. Hence, it has no finite mean whenever $u_2 \leq 1/2$. A change of variables shows that the posterior distribution in terms of ϕ is given by

$$\pi(\phi \mid \boldsymbol{y}^{(2)}) = \frac{2\phi^{\nu_1+2u_1-1}(1+\phi^2)^{-(u_1+u_2)}(1+\frac{\nu_1 s_1^2}{\nu_2 s_2^2}\phi^2)^{-\frac{\nu_1+\nu_2}{2}}}{\mathcal{B}(\frac{\nu_1}{2}+u_1, \frac{\nu_2}{2}+u_2)\,{}_2F_1\left(\frac{\nu_1+\nu_2}{2}, \frac{\nu_1}{2}+u_1\,;\,\frac{\nu_1+\nu_2}{2}+u_1+u_2\,;\,1-\frac{\nu_1 s_1^2}{\nu_2 s_2^2}\right)}. \tag{4.4}$$

4.3 Interval Bayes Factors

Researchers may wish to extend the sharp null hypothesis $\vartheta = 1/2$ to include a null-region around the point null value. If the null-region overlaps with the prior under the alternative, this leads to an (inconsistent) peri-null Bayes factor (e.g., Ly and Wagenmakers, 2021; Morey and Rouder, 2011). If the null-region does not overlap with the prior under the alternative, that is, if we compare the hypotheses

$$\mathcal{H}_0 : \phi \in [a, b] \tag{4.5}$$
$$\mathcal{H}_1 : \phi \notin [a, b], \tag{4.6}$$

then this yields a non-overlapping interval-null Bayes factor (e.g., Berger and Delampady, 1987; Rousseau, 2007). The null-region is usually informed by the problem at hand, as we will see later on an example. For a potential default approach to specifying the non-overlapping interval bounds, see the supplementary materials.

4.4 Directed Bayes Factors

Researchers sometimes desire to quantify evidence in favor of hypotheses such as $\mathcal{H}_- : \sigma_1^2 > \sigma_2^2$, or $\mathcal{H}_+ : \sigma_1^2 < \sigma_2^2$. More generally, let $\mathcal{H}_r$ denote such an order-constrained or directed hypothesis. Since $\sigma_1^2 = (2\vartheta\bar{\boldsymbol{\tau}})^{-1}$ and $\sigma_2^2 = (2(1-\vartheta)\bar{\boldsymbol{\tau}})^{-1}$, we have that $\sigma_1^2 > \sigma_2^2$ implies $\vartheta < 1/2$. We therefore restrict the beta prior on ϑ accordingly in the calculation of the marginal likelihood for $\mathcal{H}_r$ (see also Ly et al., 2016a), which can then be used to calculate directed Bayes factors.

In the more general $K > 2$ group case, we can similarly specify equality or inequality constraints by encoding them in the prior distribution on $\vec{\vartheta}$. An example of such a constrained hypotheses is given by

$$\mathcal{H}_r : \vartheta_1 = \vartheta_2 > (\vartheta_3, \vartheta_4, \vartheta_5 = \vartheta_6) > \vartheta_7 \ ,$$

which incorporates two equality constraints ($\vartheta_1 = \vartheta_2$ and $\vartheta_5 = \vartheta_6$), several order constraints (e.g., $\vartheta_1 > \vartheta_3$, $\vartheta_1 > \vartheta_4$, $\vartheta_3 > \vartheta_7$, $\vartheta_4 > \vartheta_7$), and no constraints between the ϑ_3, ϑ_4, $\vartheta_5 = \vartheta_6$ (and therefore also the standard deviations and variances). Note that while this hypothesis is formulated in terms of the parameter ϑ, it has immediate implications for the precisions and thus for the standard deviations and variances. We could also directly formulate the hypotheses on the variances or standard deviations, for example, with $(\sigma_1 = \sigma_2) > \sigma_3$ implying that $(\vartheta_1 = \vartheta_2) < \vartheta_3$. This flexibility allows researchers to translate substantive predictions directly into statistical hypotheses.

We compute Bayes factors including mixed hypotheses such as $\mathcal{H}_r$ as follows. First, we introduce a new auxiliary hypothesis $\mathcal{H}_a$ which does not include order-constraints. In our example, this yields

$$\mathcal{H}_a : \vartheta_1 = \vartheta_2, \vartheta_3, \vartheta_4, \vartheta_5 = \vartheta_6, \vartheta_7 \ .$$

We estimate the (auxiliary) Bayes factor BF_{ra} by dividing the proportion of samples ϑ that respect the order-constraints in $\mathcal{H}_r$ in the posterior by the proportion of samples that respect it in the prior (Klugkist et al., 2005). Separately, we then estimate the Bayes factor in favor of $\mathcal{H}_a$ over $\mathcal{H}_1$ (or $\mathcal{H}_0$) using bridge sampling (Meng and Wong, 1996; Gronau et al., 2017). Combining these two Bayes factors yields the desired Bayes factor in favor of $\mathcal{H}_r$ over $\mathcal{H}_1$ (or $\mathcal{H}_0$), that is, $\text{BF}_{r1} = \text{BF}_{ra} \times \text{BF}_{a1}$. The R package *bfvartest*, which is available from `https://github.com/fdabl/bfvartest`, implements this and all other procedures described above; see the supplementary materials for how to use the package.

4.5 Comparison to a Fractional Bayes Factor

One alternative to choosing the prior based on desiderata, as done in this paper, is to use the data to inform the prior. O'Hagan (1995) proposed the *fractional* Bayes factor, which uses a fraction $b = {}^{m_0}/_{n}$ of the entire likelihood to construct a prior, where m_0 is the size of the minimal training sample and n is the sample size. Böing-Messing and Mulder (2018) developed a fractional Bayes factor for testing the (in)equality of several population variances. Here, we compare our proposed default Bayes factor to their fractional Bayes factor.

Since the likelihood is the same, the key difference between the two Bayes factors is in their respective prior specification. As we are concerned with hypotheses that can feature both inequality and equality constrains, we need to introduce additional notation. Let $\mathcal{H}_r$ denote a hypothesis with q_r^E equality and q_r^I inequality constraints on K population variances, such that there are $J_r = K - q_r^E$ unique variances $\vec{\sigma}_r^2 = (\sigma_1^2, \ldots, \sigma_{J_r}^2)$. Further, let K_j be the number of populations sharing the unique variance σ_j^2, and n_{j_k} be the sample size of the k^{th} population sharing the unique variance σ_j^2. Böing-Messing and Mulder (2018) use population-specific fractions given by $b_{j_k} = {}^{2}/_{n_{j_k}}$, where $m_0 = 2$ is the minimal training sample size for the automatic prior to be proper; it is in this sense that their Bayes factor relies on minimal prior information. They calculate the marginal likelihood for hypothesis $\mathcal{H}_r$ as

$$p(y^{[K]} \mid \mathcal{H}_r) = \frac{\int_{\Omega_t} \int_{\mathbb{R}^K} f(y^{[K]}; \boldsymbol{\mu}, \vec{\sigma_r}^2) \pi(\boldsymbol{\mu}, \vec{\sigma_r}^2) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\vec{\sigma_r}^2}{\int_{\Omega_t^a} \int_{\mathbb{R}^K} f(y^{[K]}; \boldsymbol{\mu}, \vec{\sigma_r}^2)^{\boldsymbol{b}} \pi(\boldsymbol{\mu}, \vec{\sigma_r}^2) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\vec{\sigma_r}^2} \ , \tag{4.7}$$

where $\boldsymbol{b}$ is the vector of population-specific fractions, $\pi(\boldsymbol{\mu}, \vec{\sigma}_r^2) \propto \prod_{i=1}^{J_r} \sigma_i^{-2}$ is the Jeffreys prior, Ω_t specifies the region of integration depending on the inequality constraints in $\mathcal{H}_t$, and Ω_t^a is the adjusted integration region given by

$$\Omega_t^a = \left\{ \vec{\sigma_r}^2 : \boldsymbol{R}^I [a_1 \sigma_1^2 \ldots a_{J_r} \sigma_{J_r}^2] > \vec{0} \right\} \quad , \tag{4.8}$$

where $\boldsymbol{R}^I$ encodes the inequality constraints among the J_r unique variances, and where $a_j = {}^{K_j}\!/_2 \sum_{k=1}^{K_j} \left(1 - \frac{s_{j_k}^2}{n_{j_k}}\right)$. Böing-Messing and Mulder (2018) show that this setup leads to the following expression for the marginal likelihood of $\mathcal{H}_r$:

$$p(y^{[K]} \mid \mathcal{H}_r) = \frac{\int_{\Omega_r} \prod_{j=1}^{J_r} \text{Inv-Gamma}\left(\sigma_j^2; \frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}, \frac{\sum_{k=1}^{K_j} (n_{j_k} - 1) s_{j_k}^2}{2}\right) \mathrm{d}\sigma_j^2}{\int_{\Omega_r} \prod_{j=1}^{J_r} \text{Inv-Gamma}\left(\frac{K_j}{\sum_{k=1}^{K_j}\left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2} \sigma_j^2; \frac{K_j}{2}, \frac{K_j}{2}\right) \mathrm{d}\sigma_j^2}$$
$$\times \, \pi^{\frac{-\sum_{j=1}^{J_r} \sum_{k=1}^{K_j} (n_{j_k} - 2)}{2}}$$
$$\left(\prod_{j=1}^{J_r} \prod_{k=1}^{K_j} \left(\frac{n_{j_k}}{2}\right)^{\frac{1}{2}}\right) \prod_{j=1}^{J_r} \frac{\Gamma\left(\frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}\right) \left(\sum_{k=1}^{K_j} \left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2\right)^{\frac{K_j}{2}}}{\Gamma\left(\frac{K_j}{2}\right) \left(\sum_{k=1}^{K_j} (n_{j_k} - 1) s_{j_k}^2\right)^{\frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}}} \quad , \tag{4.9}$$

where Inv-Gamma$(x; \alpha, \beta)$ is the density of the inverse Gamma distribution, and the ratio of the two integrals gives the probability that the constraints hold in the posterior divided by the probability that they hold in the prior. This ratio equals 1 when testing hypotheses without order-constraints, i.e., $\Omega_t^\alpha = \Omega_t$. From Equation (4.9) it follows that the prior distribution assigned to σ_j^2 under hypothesis $\mathcal{H}_r$ is given by

$$\sigma_j^2 \sim \text{Inv-Gamma}\left(\frac{K_j}{2}, \frac{\sum_{k=1}^{K_j} \left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2}{2}\right) \quad ,$$

where n_{j_k} and $s_{j_k}^2$ are the sample size and the sum of squares of the k^{th} group sharing population variance σ_j^2. Note that, in contrast to our proposed default prior, the prior for the fractional Bayes factor proposed by Böing-Messing and Mulder (2018) depends on the data. Similarly, our prior specification results in a joint distribution on $\boldsymbol{\sigma^2}$ that cannot be factorized, that is, it results in a dependent prior, where the dependence is created through the weights $\vec{\vartheta}$. The prior specification by Böing-Messing and Mulder (2018) induces a Dirichlet prior on $\vec{\vartheta}$ with $u = {}^{K_j}\!/_2$ and a non-standard prior on $\bar{\tau}$ (it follows a Gamma distribution if and only if all sample sizes and sum of squares are equal). Figure 3 shows our default Bayes factor and the fractional Bayes factor for $K = 2$, sample sizes $n := n_1 = n_2 \in [5, \ldots, 200]$, and different values of $\phi = \{1, 1.2, 1.3, 1.4, 1.5\}$. While our proposed default Bayes factor and the fractional Bayes factor differ, they show very similar results for $u = {}^1\!/_2$.

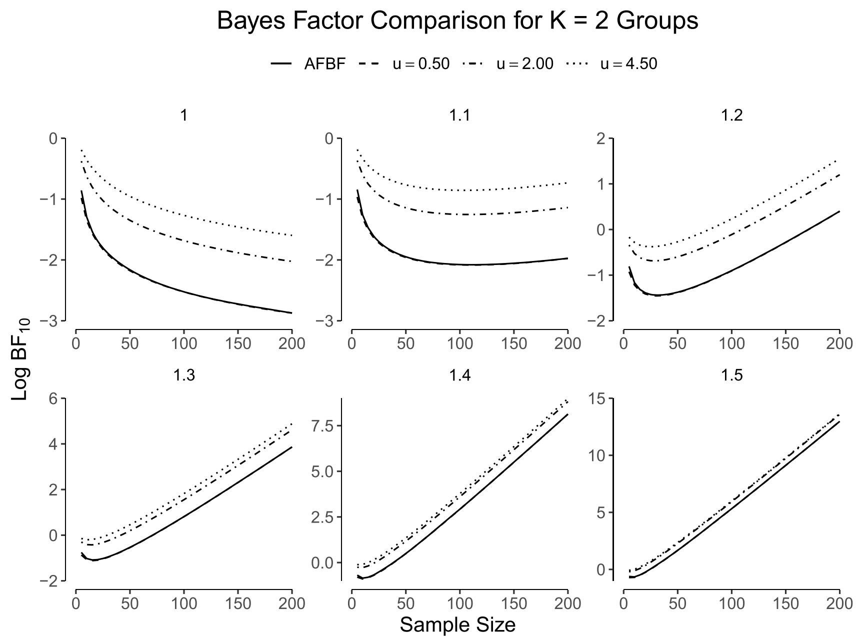

Figure 3: Comparison of the Bayes factor proposed by Böing-Messing and Mulder (2018) and our Bayes factor for $K = 2$ groups as a function of $n := n_1 = n_2$, prior specification $u := u_1 = u_2$, and effect size $\phi = \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$.

There an interesting discrepancy between the two Bayes factors when testing directed hypotheses. In case there is overwhelming evidence for the hypothesis that $\mathcal{H}_r : \sigma_1^2 > \ldots > \sigma_K^2$, the Bayes factor in favor of it over $\mathcal{H}_1 : \sigma_1^2 \neq \ldots \neq \sigma_K^2$ reaches the bound $K!$. However, in case there are the same J equalities in both hypotheses, the fractional Bayes factor does not reach the bound of $(K - J)!$, while our proposed default Bayes factor does. This is because Böing-Messing and Mulder (2018) set $b_{j_k} = {}^2\!/_{n_{j_k}}$ for all groups. While this is desirable in the sense that one thus uses the same 'minimal' amount of information under each hypothesis, this results in a different shape parameter of the inverse gamma prior distribution, and the bound is therefore not reached, which can be considered a shortcoming of the fractional Bayes factor.

4.6 Multiple Comparisons

So far, we have focused on comparing the null hypothesis $\mathcal{H}_0$ in which all variances are equal against the alternative hypothesis $\mathcal{H}_1$ in which all variances were free to vary or against mixed hypotheses $\mathcal{H}_r$ which allow for inequalities, equalities, and order-constraints. However, researchers are sometimes also interested in assessing all possible (in)equalities. Statistically, all possible configurations of equality and inequality constraints can be uniquely represented as partitions of the groups, where any number of groups are equal if they are in the same partition. Given K groups, the number of partitions of size j is given by the Stirling numbers of the second kind, denoted $\left\{ {K \atop j} \right\}$. The total number of partitions is given by the K^{th}-Bell number, which is defined as a sum over the Stirling numbers:

$$B_K = \sum_{j=0}^{K} \left\{ {K \atop j} \right\} . \tag{4.10}$$

The Bell numbers grow quickly, with $K = 10$ already yielding $115,975$ models. This results in a multiple comparisons problem, which in a Bayesian framework can be addressed by suitable adjusting the prior model odds (e.g., Jeffreys, 1961; Westfall et al., 1997). Inspired by the work on variable selection in regression (Scott and Berger, 2006, 2010), van den Bergh and Dablander (2022) recently proposed a beta-binomial prior for this problem, comparing it to a Dirichlet process prior proposed by Gopalan and Berry (1998) as well as to other methods to multiple comparisons that do not require specifying a prior over all models (Westfall et al., 1997; de Jong, 2019; Jeffreys, 1961). For a small number of groups, one can directly calculate the marginal likelihood of each model and use the posterior model probabilities for inference:

$$p(\mathcal{H}_j \mid y^{[K]}) = \frac{p(y^{[K]} \mid \mathcal{H}_j)\pi(\mathcal{H}_j)}{\sum_{i=0}^{B_K} p(y^{[K]} \mid \mathcal{H}_i)\pi(\mathcal{H}_i)} = \frac{\mathrm{BF}_{j0}\pi(\mathcal{H}_j)}{\sum_{i=0}^{B_K} \mathrm{BF}_{i0}\pi(\mathcal{H}_i)} , \tag{4.11}$$

where B_K is the K^{th} Bell number and the prior models probabilities $\pi(\mathcal{H}_j)$ are suitable adjusted, as detailed in van den Bergh and Dablander (2022). Table 1 shows the results of an analysis detailed in Section 5.6 for a $K = 4$ group case under different model priors. For details, we refer the interested reader to van den Bergh and Dablander (2022), who also develop a stochastic search method to deal with larger K.

5 Practical Examples

In the following sections we apply our proposed Bayes factor test on a number of examples.

5.1 Sex Differences in Personality

There is a rich history of research and theory about differences in variability between men and women, going back at least to Charles Darwin (Darwin, 1871). Borkenau et al. (2013) studied whether men and women differ in the variability of personality traits. Here, we focus on peer-rated conscientiousness in Estonian women and men ($s_f^2 = 15.6$, $s_m^2 = 19.9$, $n_f = 969$, $n_m = 716$). The left panel in Figure 4 visualizes the raw data, and the middle panel shows the prior (using $u = 1/2$) and the posterior distribution for the effect size ϕ. The default Bayes factor yields $\mathrm{BF}_{10} = 12.98$ in favor of a difference in variance, and the right panel shows a sensitivity analysis to the specification of u in the default Bayes factor (note that the x-axis scale is $1/u$); as expected, a smaller value of u corresponds to a wider prior of ϕ under $\mathcal{H}_1$ and decreases the predictive performance

	Beta-binomial Prior		Dirichlet Process Prior	
Hypothesis	$\alpha = 1, \beta = 1$	$\alpha = 1, \beta = 4$	$\alpha = 1$	$\alpha = 1.817$
{Flemish, German, Estonian, Czech}	0.250 (0.446)	0.571 (0.739)	0.250 (0.368)	0.116 (0.192)
{Flemish}, {German, Czech}, {Estonian}	0.042 (0.029)	0.019 (0.007)	0.042 (0.016)	0.064 (0.034)
{Flemish, Estonian}, {German}, {Czech}	0.042 (0.005)	0.019 (0.001)	0.042 (0.003)	0.064 (0.006)
{Flemish, Czech}, {German}, {Estonian}	0.042 (0.000)	0.019 (0.000)	0.042 (0.000)	0.064 (0.000)
{Flemish}, {German, Estonian}, {Czech}	0.042 (0.083)	0.019 (0.018)	0.042 (0.053)	0.064 (0.118)
{Flemish, German}, {Estonian}, {Czech}	0.042 (0.015)	0.019 (0.004)	0.042 (0.009)	0.064 (0.023)
{Flemish}, {German}, {Estonian, Czech}	0.042 (0.018)	0.019 (0.004)	0.042 (0.015)	0.064 (0.029)
{Flemish, Estonian}, {German, Czech}	0.036 (0.030)	0.041 (0.017)	0.042 (0.014)	0.035 (0.019)
{Flemish, German}, {Estonian, Czech}	0.036 (0.060)	0.041 (0.038)	0.042 (0.056)	0.035 (0.049)
{Flemish, Czech}, {German, Estonian}	0.036 (0.004)	0.041 (0.002)	0.042 (0.003)	0.035 (0.004)
{Flemish, Estonian, Czech}, {German}	0.036 (0.005)	0.041 (0.004)	0.083 (0.009)	0.070 (0.007)
{Flemish, German, Estonian}, {Czech}	0.036 (0.061)	0.041 (0.041)	0.083 (0.105)	0.070 (0.111)
{Flemish, German, Czech}, {Estonian}	0.036 (0.003)	0.041 (0.002)	0.083 (0.005)	0.070 (0.005)
{Flemish}, {German, Estonian, Czech}	0.036 (0.211)	0.041 (0.120)	0.083 (0.339)	0.070 (0.390)
{Flemish}, {German}, {Estonian}, {Czech}	0.250 (0.029)	0.029 (0.001)	0.042 (0.003)	0.116 (0.012)

Table 1: Prior (and posterior) probabilities of the different hypotheses under different model priors illustrated on the example discussed in Section 5.6. Groups with the same population variance are put into the same set, e.g. $\sigma_1 = \sigma_2 \neq \sigma_3 = \sigma_4$ corresponds to $\{\{\sigma_1, \sigma_2\}, \{\sigma_3, \sigma_4\}\}$.

of $\mathcal{H}_1$ compared to $\mathcal{H}_0$. Nevertheless, across the range of u visualized in Figure 4, there is strong evidence that Estonian men show larger variability in conscientiousness than Estonian women. For comparison, a frequentist analysis using Bartlett's test (Bartlett, 1937) yields $\chi^2(1) = 12.54$, $p = 0.0004$. The Vovk-Sellke bound $1/(-e \cdot p \log(p))$ (Vovk, 1993; Sellke et al., 2001) gives the maximum possible odds in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$ based on the p-value, and yields 118.11.

5.2 Testing Against a Single Value

Polychlorinated biphenyls (PCB), which are used in the manufacture of large electrical transformers and capacitors, are hazardous contaminants when released into the environment. Suppose that the Environmental Protection Agency is testing a new device for measuring PCB concentration (in parts per million) in fish, requiring that the instrument yields a variance of less than 0.10 (a standard deviation $\sigma_0 \leq 0.32$), thus $\phi > 1$. This suggests the use of a directed Bayes factor. Seven PCB readings on the same sample of fish are subsequently performed, yielding a sample standard deviation of $s = 0.22$ and a sample effect size of $\hat{\phi} = \frac{\sigma_0}{s} = 1.42$ (see Mendenhall and Sincich, 2016, p. 420). We compare the following hypotheses

$$\mathcal{H}_0 : \phi = 1$$
$$\mathcal{H}_+ : \phi > 1,$$

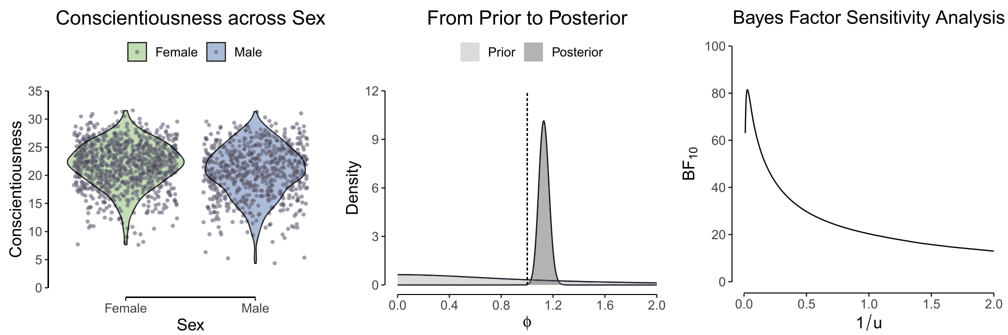

Figure 4: Left: Peer-rated conscientiousness of Estonian men and women. Middle: Prior and posterior of ϕ (with $u = 1/2$). Right: Bayes factor sensitivity analysis for $u \in [1/2, 100]$.

which yields $\text{BF}_{+0} = 0.51$ for the default value $u = 1/2$, a value slightly higher than for an undirected test, $\text{BF}_{10} = 0.41$. To illustrate prior elicitation, assume that the makers of the new device are highly confident, assigning 50% probability to the outcome that the new device reduces the required standard deviation at least by half. Defining $\phi = \frac{\sigma_0}{\sigma_{\text{device}}}$, this formally translates into $\pi(\phi \in [2, \infty]) = 1/2$, which is fulfilled by a (truncated) prior with $u = 2.16$. Using this prior specification results in $\text{BF}_{+0} = 0.83$.

5.3 Comparing Measurement Precision

In paleoanthropology, researchers study the anatomical development of modern humans. An important problem in this area is to adequately reconstruct excavated skulls. Sholts et al. (2011) compared the precision of coordinate measurements of different landmark types on human crania using a 3D laser scanner and a 3D digitizer. They reconstructed five excavated skulls and found — for landmarks of Type III, that is, the smooth part of the forehead above and between the eyebrows — an average (across skulls) standard deviation of 0.98 for the Digitizer ($n_1 = 990$) and an average standard deviation of 0.89 for the Laser ($n_2 = 990$). We define $\phi = \frac{\sigma_{\text{Digitizer}}}{\sigma_{\text{Laser}}}$ and observe that the sample effect size is 1.10. We demonstrate two tests. First, we test whether the Laser has a lower standard deviation than the Digitizer, writing

$$\begin{aligned}\mathcal{H}_0 &: \phi = 1 \\ \mathcal{H}_+ &: \phi > 1 \ .\end{aligned}$$

The default Bayes factor in favor of $\mathcal{H}_1$ is $\text{BF}_{+0} = 4.93$ — about double the undirected Bayes factor $\text{BF}_{+0} = 2.47$ — indicating moderate evidence for the hypothesis that a 3D Laser is a more precise tool for measuring Type III landmarks on the excavated human scull compared to a 3D Digitizer. Second, in this specific scenario, a researcher might treat the Digitizer as being equally as precise as the Laser when its standard deviation differs by a maximum of 10%. She might then choose to compare the following non-overlapping hypotheses:

$$\mathcal{H}_0^{'} : \phi \in [0.90, 1.10]$$
$$\mathcal{H}_+^{'} : \phi > 1.10 \ .$$

The Bayes factor with $u = 1/2$ in favor of $\mathcal{H}_0'$ is $\mathrm{BF}_{0+}^{'} = 7.03$, indicating moderate support for the hypothesis that the Laser and the Digitizer have about equal performance. In general, we recommend researchers use the default Bayes factor unless substantive prior knowledge or particular circumstances justify a different test. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 9.16$, $p = 0.0025$, with a Vovk-Sellke bound of 24.76.

5.4 The "Standardization" Hypothesis in Archeology

Economic growth encourages increased specialization in the production of goods, which leads to the "standardization" hypothesis: increased production of an item would lead to it becoming more uniform. Kvamme et al. (1996) sought to test this hypothesis by studying chupa-pots, a type of earthenware produced by three different Philippine communities: the *Dangtalan*, where ceramics are primarily made for household use; the *Dalupa*, where ceramics are traded in a non-market based barter economy; and the *Paradijon*, which houses full-time pottery specialists that sell their ceramics to shopkeepers for sale to the general public. Thus, there is an increased specialization across these three communities. Kvamme et al. (1996) use circumference, height, and aperture as measures for the chupa-pots; here, we focus on the latter two. The authors test whether the standard deviations across these three groups are different, comparing

$$\mathcal{H}_0 : \sigma_1 = \sigma_2 = \sigma_3$$
$$\mathcal{H}_1 : \sigma_1 \neq \sigma_2 \neq \sigma_3 \ ,$$

where σ_1, σ_2, and σ_3 correspond to the standard deviations of chupa-pots in the Dangtalan, Dalupa, and Paradijon communities, respectively. Since our Bayes factor test only requires summary statistics, we can test these hypotheses using the data from Table 4 in Kvamme et al. (1996). The authors observed $n = 55$ pots from the Dangtalan community with a standard deviation in aperture of 12.74; $n = 171$ pots from the Dalupa community with a standard deviation of 8.13; and $n = 117$ pots from the Paradijon community with a standard deviation of 5.83. Using our default prior choice of $u = 1/2$, we find overwhelming evidence for a difference in the standard deviations of the aperture measurements, $\log(\mathrm{BF}_{10}) = 20$. Note that we can formulate a stronger statistical hypothesis based on the substantive "standardization" hypothesis, namely that the standard deviations in aperture *increase* from the Paradijon to the Dangtalan community, $\mathcal{H}_r : \sigma_1 > \sigma_2 > \sigma_3$. This yields even stronger evidence, $\log(\mathrm{BF}_{r0}) = 21.80$, such that the Bayes factor in favor of $\mathcal{H}_r$ compared to $\mathcal{H}_1$ is very close to its theoretical maximum, $\mathrm{BF}_{r1} = 5.98 \approx 3!$. If we were to use height instead of aperture measurements of the pots, which yield standard deviations of 9.60, 7.23, and 7.81, respectively, the evidence in favor of $\mathcal{H}_1$ and $\mathcal{H}_r$ compared to $\mathcal{H}_0$ would be much weaker, $\mathrm{BF}_{10} = 2.27$ and $\mathrm{BF}_{r0} = 2.87$, respectively. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 49.94$, $p < 0.00001$ with a (log) Vovk-Sellke bound of 20.75 for the aperture measurements and $\chi^2(1) = 7.18$, $p = 0.0277$ with a Vovk-Sellke bound of 3.71 for the height measurements.

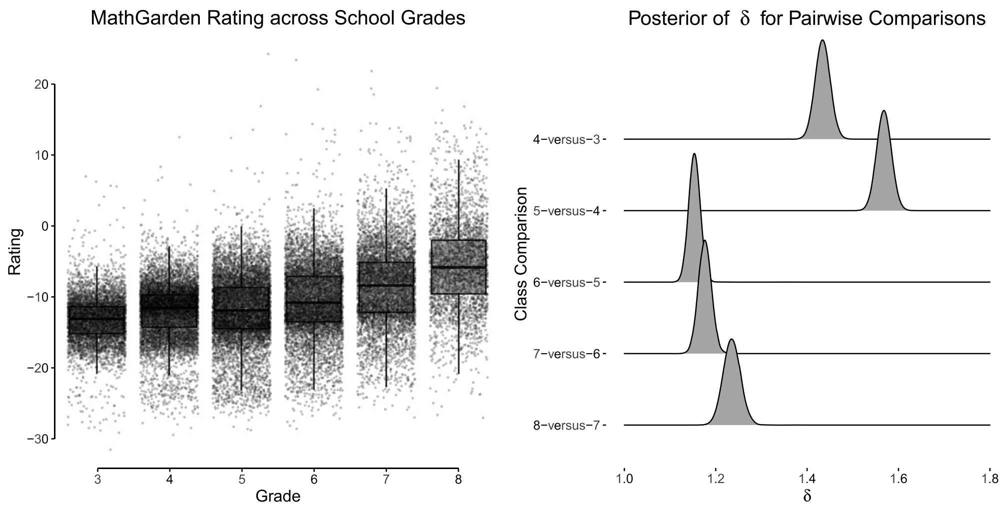

Figure 5: Left: Shows Math Garden rating scores across school grades. Right: Shows posterior of ϕ for pairwise consecutive class comparisons. Virtually all probability mass is assigned to $\phi > 1$, implying that, indeed, the variance increases with every school grades.

5.5 Increased Variability in Mathematical Ability

Aunola et al. (2004) find that the variance in mathematical ability increases across school grades. Using large-scale data from Math Garden, an online learning platform in the Netherlands (Brinkhuis et al., 2018), we assess the evidence for this hypothesis using our Bayes factor test. Math Garden assigns each pupil a rating, similar to an Elo score used in chess, and which increases if the pupil solves problems correctly. We have data from $n = 41,801$ different pupils across school grades $3 - 8$, which is visualized in the left panel of Figure 5. From grade 3 upwards, the standard deviations of the Math Garden ratings are $3.08, 3.69, 4.62, 4.97, 5.39$, and 5.99, for respective sample sizes of $6,410$, $9,395$, $9,160$, $7,549$, $6,007$, and $3,280$. Following Aunola et al. (2004), we wish to compare the following three hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \sigma_i = \sigma_j \quad \forall(i,j) \\ \mathcal{H}_1 &: \sigma_i \neq \sigma_j \quad \forall(i,j) \\ \mathcal{H}_r &: \sigma_i > \sigma_j \quad \forall(i > j) \ . \end{aligned}$$

Using the default choice $u = 1/2$, we find overwhelming support in favor of a difference in the standard deviations, $\log(\text{BF}_{10}) = 1660.53$. As is suggested by the raw data visualized in the left panel of Figure 5, we also find overwhelming support for an increase in variability with increased school grade, $\log(\text{BF}_{r0}) = 1667.11$. The order-constrained hypothesis again strongly outperforms the unrestricted hypothesis, yielding evidence close to its theoretical maximum, $\text{BF}_{r1} = 719.69 \approx 6!$. The right panel in Figure 5 shows the posterior distribution of ϕ for pairwise comparisons across school grades. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 3366.70$, $p < 0.00001$ with a (log) Vovk-Sellke bound of 1664.07.

5.6 Country Differences in Conscientiousness

As our last example, we illustrate how researchers could use our default Bayes factor combined with the work by van den Bergh and Dablander (2022) to test all possible (in)equalities between variances. We utilize the data set by Borkenau et al. (2013) again, but now test whether the Czech ($s_C^2 = 20, n = 714$), Estonian ($s_E^2 = 17.7, n = 1685$), German ($s_G^2 = 17.3, n = 303$), and Flemish ($s_F^2 = 14.2, n = 291$) population differ in their variances of peer-rated conscientiousness. The posterior probability for each hypothesis under a different prior model specification can be found in Table 1. We find that the null hypothesis of no differences generally yields the highest posterior probability, followed by the hypothesis which states that the Flemish population variance differs from the rest. The left panels in Figure 6 show the posterior distributions for each variance under the full model (top) and when model-averaging across all models (bottom) using the beta-binomial($\alpha = 1, \beta = 4$), which is recommended by van den Bergh and Dablander (2022). We see that there is pronounced shrinkage towards the average variance, which is an indication that the model in which all variances are equal is strongly supported (see also Table 1). The right panel shows the probability that any two populations show the same variance in their peer-rated conscientiousness. We find that the German and Estonian population are most likely and the Flemish and Czech population least likely to have the same variance. This is also reflected in the unconstrained variance estimates shown in the left panel. For comparison, a Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 11.51$, $p = 0.0093$ with a Vovk-Sellke bound of 8.48.

6 Conclusion

In this paper, we proposed a default Bayes factor test for assessing the (in)equality of several population variances and showed that it fulfills a number of desiderata for Bayesian model comparison (e.g., Bayarri et al., 2012; Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a; Ly, 2018; Peña, 2018). In addition, we extended the Bayes factor test to cover the $K-1$-sample case, non-overlapping interval nulls, and mixed restrictions for the $K > 2$ case. The proposed procedure allows researchers to inform their statistical tests with prior knowledge. It also generalizes Jeffreys's test for the agreement of two standard errors (Jeffreys, 1939, pp. 222-224); see the supplementary materials. We have also illustrated how our method — combined with specifying suitable model priors — can be used to test all possible (in)equalities between variances while adjusting for multiplicity (van den Bergh and Dablander, 2022)

A limitation of the proposed methodology is that it assumes that the data follow a Gaussian distribution, which might not always be adequate in practical applications. A potential extension would be to use a t-distributions with a small number of degrees of freedom $\nu \geq 3$, so as to better accommodate outliers, and then test whether the scales

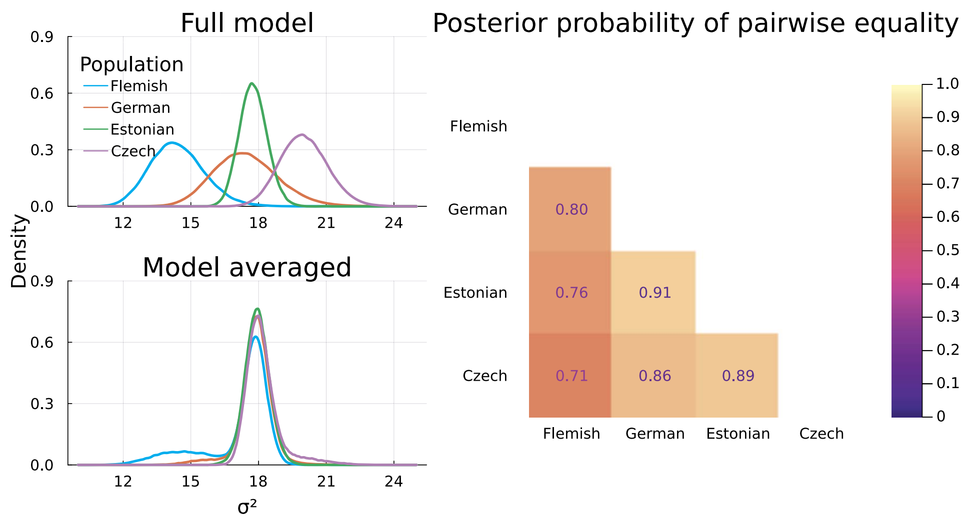

Figure 6: Left: Posterior means of the full model where all variances are assumed to be different (top) and posterior means when averaging across all models using a beta-binomial($\alpha = 1$, $\beta = 4$) prior (bottom). Right: Posterior probabilities for pairwise equality across all populations.

of these t-distributions differ. Another future avenue is to allow for data from the same unit, that is, allow for correlated observations or dependent groups. For the present, we believe that our work provides an elegant Bayesian complement to popular classical tests for assessing the (in)equality of several independent population variances, ready for routine applications.

Author Contributions FD and DvB proposed the study. They both worked out the initial derivations and proofs for the deterministic $K = 2$ case with the help of AL. FD wrote the first draft of the manuscript and, together with DvB, analyzed the data. FD developed the software package with the help of DvB. AL extended the results to the $K \geq 2$ case and provided the proofs shown in the supplementary materials. FD, DvB, and AL wrote the manuscript. EJW provided detailed feedback on the manuscript and guidance throughout. All authors read and approved the submitted version of the paper. They also declare that there were no conflicts of interest.

Supplementary Material

Supplementary Material for "Default Bayes Factors for Testing the (In)equality of Several Population Variances" (DOI: 10.1214/23-BA1369SUPP; .pdf). Includes detailed derivations and proofs of all theorems mentioned in the main manuscript as well as code to reproduce the analysis of the empirical examples.

References

Abramowitz, M. and Stegun (1972). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*, volume 55. New York, United States: Dover publications. MR0415956. 707, 708

Aunola, K., Leskinen, E., Lerkkanen, M.-K., and Nurmi, J.-E. (2004). "Developmental Dynamics of Math Performance From Preschool to Grade 2." *Journal of Educational Psychology*, 96(4): 699–713. 700, 717

Bahadur, R. R. and Bickel, P. J. (2009). "An optimality property of Bayes' test statistics." *Lecture Notes-Monograph Series*, 57: 18–30. MR2681656. doi: https://doi.org/10.1214/09-LNMS5704. 705

Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901): 268–282. MR0024103. 714

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 703, 718

Berger, J. O. and Delampady, M. (1987). "Testing precise hypotheses." *Statistical Science*, 317–335. MR0920141. 709

Böing-Messing, F. and Mulder, J. (2018). "Automatic Bayes factors for testing equality and inequality-constrained hypotheses on variances." *Psychometrika*, 83(3): 1–32. MR3851948. doi: https://doi.org/10.1007/s11336-018-9615-z. 700, 710, 711, 712

Borkenau, P., Hřebíčková, M., Kuppens, P., Realo, A., and Allik, J. (2013). "Sex differences in variability in personality: A study in four samples." *Journal of Personality*, 81(1): 49–60. 713, 718

Brinkhuis, M. J., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L., and Maris, G. (2018). "Learning as It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System." *Journal of Learning Analytics*, 5(2): 29–46. 717

Brown, M. B. and Forsythe, A. B. (1974). "Robust tests for the equality of variances." *Journal of the American Statistical Association*, 69(346): 364–367. MR0005402. 700

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I., et al. (2018). "Prior distributions for objective Bayesian analysis." *Bayesian Analysis*, 13(2): 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 700, 703, 718

Dablander, F., van den Bergh, D., Wagenmakers, E.-J., and Ly, A. (2023). "Supplementary Material for "Default Bayes Factors for Testing the (In)equality of Several Population Variances"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1369SUPP. 700

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London, UK: John Murray. 713

de Jong, T. (2019). “A Bayesian approach to the correction for multiplicity.” 713

Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). “The impact of Levene’s test of equality of variances on statistical theory and practice.” *Statistical Science*, 24(3): 343–360. MR2757435. doi: https://doi.org/10.1214/09-STS301. 700

Gopalan, R. and Berry, D. A. (1998). “Bayesian multiple comparisons using Dirichlet process priors.” *Journal of the American Statistical Association*, 93(443): 1130–1139. MR1649207. doi: https://doi.org/10.2307/2669856. 713

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). “A tutorial on bridge sampling.” *Journal of Mathematical Psychology*, 81: 80–97. MR3722819. doi: https://doi.org/10.1016/j.jmp.2017.09.005. 710

Hendriksen, A., de Heide, R., and Grünwald, P. (2021). “Optional Stopping with Bayes Factors: A categorization and extension of folklore results, with an application to invariant situations.” *Bayesian Analysis*, 16(3): 961–989. MR4303875. doi: https://doi.org/10.1214/20-BA1234. 702

Hoijtink, H., Klugkist, I., and Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, United States: Springer. 700

Jeffreys, H. (1939). *Theory of Probability (1rd Ed.)*. Oxford, UK: Oxford University Press. MR0000924. 700, 701, 703, 718

Jeffreys, H. (1961). *Theory of Probability (3rd Ed.)*. Oxford, UK: Oxford University Press. 705, 713

Johnson, V. E. and Rossell, D. (2010). “On the use of non-local prior densities in Bayesian hypothesis tests.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 705

Kass, R. E. and Raftery, A. E. (1995). “Bayes factors.” *Journal of the american statistical association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 701

Klugkist, I., Kato, B., and Hoijtink, H. (2005). “Bayesian model selection using encompassing priors.” *Statistica Neerlandica*, 59(1): 57–69. MR2137381. doi: https://doi.org/10.1111/j.1467-9574.2005.00279.x. 710

Kvamme, K. L., Stark, M. T., and Longacre, W. A. (1996). “Alternative procedures for assessing standardization in ceramic assemblages.” *American Antiquity*, 61(1): 116–126. 699, 716

Levene, H. (1961). “Robust tests for equality of variances.” In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B. (eds.), *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, 279–292. Stanford, California: Stanford University Press. MR0120693. 700

Ly, A. (2018). “Bayes Factors for Research Workers.” Ph.D. thesis, University of Amsterdam. Retrieved from: https://hdl.handle.net/11245.1/e601b852-1b29-407b-a276-1ccd2a2ed37b. 705, 718

Ly, A., Verhagen, A. J., and Wagenmakers, E.-J. (2016a). "Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology." *Journal of Mathematical Psychology*, 72: 19–32. MR3506022. doi: `https://doi.org/10.1016/j.jmp.2015.06.004`. 700, 703, 709, 718

Ly, A., Verhagen, A. J., and Wagenmakers, E.-J. (2016b). "An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys." *Journal of Mathematical Psychology*, 72: 43–55. MR3506025. doi: `https://doi.org/10.1016/j.jmp.2016.01.003`. 700, 702, 703

Ly, A. and Wagenmakers, E.-J. (2021). "Bayes factors for peri-null hypotheses." *arXiv preprint arXiv:2102.07162*. MR4517127. doi: `https://doi.org/10.1007/s11749-022-00819-w`. 709

Mendenhall, W. M. and Sincich, T. L. (2016). *Statistics for Engineering and the Sciences (6th Edition)*. Chapman and Hall/CRC. 714

Meng, X.-L. and Wong, W. H. (1996). "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration." *Statistica Sinica*, 831–860. MR1422406. 710

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2016). "The philosophy of Bayes factors and the quantification of statistical evidence." *Journal of Mathematical Psychology*, 72: 6–18. MR3506021. doi: `https://doi.org/10.1016/j.jmp.2015.11.001`. 700

Morey, R. D. and Rouder, J. N. (2011). "Bayes factor approaches for testing interval null hypotheses." *Psychological Methods*, 16(4): 406–419. 709

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). "Dirichlet and related distributions: Theory, methods and applications." MR2830563. doi: `https://doi.org/10.1002/9781119995784`. 706

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 99–118. MR1325379. 710

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: Eliciting Experts' Probabilities*. John Wiley & Sons. 700

Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). "On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the Women's Genome Health Study." *PLoS Genetics*, 6(6): e1000981. 699

Peña, V. (2018). "Bayesian Model Uncertainty and Foundations." Ph.D. thesis, Duke University. Retrieved from https://hdl.handle.net/10161/17494. 705, 718

Phillips, P. C. B. (1988). "The characteristic function of the Dirichlet and multivariate F distributions." *Cowles Foudation for Research in Econonmics*, 1–17. 706

Robert, C. P. (2016). "The expected demise of the Bayes Factor." *Journal of Mathematical Psychology*, 72: 33–37. MR3506023. doi: `https://doi.org/10.1016/j.jmp.2015.08.002`. 702

Rousseau, J. (2007). "Approximating interval hypothesis: p-values and Bayes factors." In Bernardo, J., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting June 2–6, 2006*, volume 8, 417–452. Oxford University Press. MR2433202. 709

Saks, M. J., Hollinger, L. A., Wissler, R. L., Evans, D. L., and Hart, A. J. (1997). "Reducing variability in civil jury awards." *Law and Human Behavior*, 21(3): 243–256. 699

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of statistical planning and inference*, 136(7): 2144–2162. MR2235051. doi: https://doi.org/10.1016/j.jspi.2005.08.031. 713

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 713

Sellke, T., Bayarri, M., and Berger, J. O. (2001). "Calibration of ρ values for testing precise null hypotheses." *The American Statistician*, 55(1): 62–71. MR1818723. doi: https://doi.org/10.1198/000313001300339950. 714

Sholts, S. B., Flores, L., Walker, P. L., and Wärmländer, S. K. (2011). "Comparison of coordinate measurement precision of different landmark types on human crania using a 3D laser scanner and a 3D digitiser: implications for applications of digital morphometrics." *International Journal of Osteoarchaeology*, 21(5): 535–543. 699, 715

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., and Wagenmakers, E.-J. (2019). "A tutorial on Bayes Factor Design Analysis using an informed prior." *Behavior Research Methods*, 51(3): 1042–1058. 700

van den Bergh, D. and Dablander, F. (2022). "Flexible Bayesian Multiple Comparison Adjustment Using Dirichlet Process and Beta-Binomial Model Priors." *arXiv preprint arXiv:2208.07086*. 713, 718

Vovk, V. G. (1993). "A logic of probability, with application to the foundations of statistics." *Journal of the Royal statistical society: series B (Methodological)*, 55(2): 317–341. MR1224399. 714

Wasserstein, R. L. and Lazar, N. A. (2016). "The ASA's Statement on p-values: Context, process, and purpose." *The American Statistician*, 70(2): 129–133. MR3511040. doi: https://doi.org/10.1080/00031305.2016.1154108. 700

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997). "A Bayesian perspective on the Bonferroni adjustment." *Biometrika*, 84(2): 419–427. MR1467057. doi: https://doi.org/10.1093/biomet/84.2.419. 713

Acknowledgments

The authors would like to thank Victor Peña for inspiring discussions on across-sample consistency and the editor Michele Guindani and two anonymous reviewers for their remarks on a previous version of the manuscript.