Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 43: 1162–1177, 2016
doi: 10.1111/sjos.12235

Restricted most powerful Bayesian tests for linear models

SCOTT D. GODDARD
Department of Mathematics and Statistics, University of Alaska Fairbanks
VALEN E. JOHNSON
Department of Statistics, Texas A & M University

ABSTRACT. Uniformly most powerful Bayesian tests (UMPBTs) are a new class of Bayesian tests in which null hypotheses are rejected if their Bayes factor exceeds a specified threshold. The alternative hypotheses in UMPBTs are defined to maximize the probability that the null hypothesis is rejected. Here, we generalize the notion of UMPBTs by restricting the class of alternative hypotheses over which this maximization is performed, resulting in restricted most powerful Bayesian tests (RMPBTs). We then derive RMPBTs for linear models by restricting alternative hypotheses to g priors. For linear models, the rejection regions of RMPBTs coincide with those of usual frequentist F-tests, provided that the evidence thresholds for the RMPBTs are appropriately matched to the size of the classical tests. This correspondence supplies default Bayes factors for many common tests of linear hypotheses. We illustrate the use of RMPBTs for ANOVA tests and t-tests and compare their performance in numerical studies.

Key words: ANOVA, Bayesian testing, g prior, linear models, t-test, UMPBT

1. Introduction

Tests of hypotheses in linear models are among the most commonly used statistical procedures (ASA, 1980). However, Bayesian versions of these tests are seldom reported in the scientific literature because they require the specification of an alternative hypothesis, and in many settings a clearly defined objective alternative hypothesis is not available. In this article, we address this problem by defining restricted most powerful Bayesian tests (RMPBTs). Special cases of these tests are applied to tests of regression coefficients in linear models by restricting the class of alternative hypotheses to take the form of Zellner's g prior Zellner (1986). By adjusting the evidence threshold of the RMPBT to match type I error rates, we obtain tests that possess the same rejection regions as classical tests. As a consequence, RMPBTs provide default Bayes factors that can be reported along with frequentist p-values. Under an assumption of equipoise (i.e. equal prior probabilities are assigned to the null and alternative hypothesis), these Bayes factors can be used to define the posterior probabilities that the null and alternative hypotheses are true, quantities that are often of primary interest to practitioners.

To make these notions more precise, we introduce the following notation. Let $\mathbf{y}$ denote an observation drawn from a family of densities $f_{\boldsymbol{\theta}}$ indexed by a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We define distinct statistical hypotheses by assuming distinct prior densities on $\boldsymbol{\theta}$; hence, let H_i denote the hypothesis defined by the specification of prior density π_i on $\boldsymbol{\theta}$. Let BF_{ij} denote the Bayes factor in favor of hypothesis H_i against H_j. The Bayes factor $BF_{ij}(\mathbf{y})$ is defined as

$$BF_{ij}(\mathbf{y}) = \frac{m_i(\mathbf{y})}{m_j(\mathbf{y})}, \qquad \text{where} \qquad m_k(\mathbf{y}) = \int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\mathbf{y})\pi_k(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

With this notation, Johnson (2013a) defined a uniformly most powerful Bayesian test for evidence threshold δ [UMPBT(δ)] against a fixed null hypothesis H_0 to be the hypothesis test in

favor of an alternative hypothesis H_1 that maximized the probability that its Bayes factor in favor of H_1 exceeded its evidence threshold δ. That is, the UMPBT(δ) test satisfies

$$\mathbf{P}_{\boldsymbol{\theta}_t}[BF_{10}(\mathbf{y}) > \delta] \geq \mathbf{P}_{\boldsymbol{\theta}_t}[BF_{20}(\mathbf{y}) > \delta],$$

for all possible values of $\boldsymbol{\theta}_t$ and all alternative hypotheses H_2.

An important property of UMPBTs is that the rejection regions for these tests (i.e. the values of $\mathbf{y}$ for which $BF_{10} > \delta$) can often be made to coincide with the rejection regions for classical UMPTs by choosing δ to match the size of the classical test. Unfortunately, UMPBTs exist in a relatively limited number of testing scenarios (e.g. one parameter exponential families), and in particular they cannot be defined for tests of parameters in the normal general linear model when variance parameters are not known *a priori*. This is because the alternative hypothesis that maximizes the probability that the resulting Bayes factor exceeds a given threshold depends on the unknown variance parameter, so that different UMPBTs are obtained for different values of the variance parameters. Thus, a unique UMPBT cannot be defined for all data-generating parameters. To remedy this situation, we define an extension of UMPBTs that we call restricted most powerful Bayesian tests. The extension is obtained by restricting the class of prior densities on $\boldsymbol{\theta}$ that define the hypotheses to a parametric class, say $\pi(\boldsymbol{\theta} \mid \boldsymbol{\psi})$.

Definition 1. A *π-restricted most powerful Bayesian test* for evidence threshold $\delta > 0$ in favor of the alternative hypothesis $H_1 : \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{\psi}_1)$ against a fixed null hypothesis H_0, denoted as π-RMPBT(δ), is a Bayesian hypothesis test in which the Bayes factor for the test satisfies

$$\mathbf{P}_{\boldsymbol{\theta}_t}[BF_{10}(\mathbf{y}) > \delta] \geq \mathbf{P}_{\boldsymbol{\theta}_t}[BF_{20}(\mathbf{y}) > \delta],$$

for any $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}$ and for all alternative hypotheses $H_2 : \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{\psi}_2)$, where π is a density function parameterized by $\boldsymbol{\psi}$, and $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \boldsymbol{\Psi}$. A RMPBT($\delta$) refers to a π-RMPBT(δ), where the dependence on the parametric class of prior densities π has been suppressed.

In essence, we obtain RMPBTs by narrowing the search of alternative hypotheses to a class of prior densities on $\boldsymbol{\theta}$. We assume that this class either incorporates prior knowledge or provides computational convenience. The optimization within this class produces a value for one or more hyperparameters $\boldsymbol{\psi}$ which maximize the probability that the Bayes factor exceeds δ over all possible values of $\boldsymbol{\psi}$ and over all $\boldsymbol{\theta}_t$.

The criterion used to define RMPBTs, like the criterion used to define UMPBTs, has implications for the type of consistency that Bayes factors based on these tests achieve. RMPBTs are defined to maximize the probability that their Bayes factor in favor of an alternative hypothesis exceeds a threshold, and in this sense they resemble classical most powerful tests. For a fixed evidence threshold, this also means that they incur a non-zero Type 1 error rate under a true null hypothesis even as the sample size tends to infinity. Furthermore, for a fixed sample size n, they generally do not produce Bayes factors that tend to infinity as the evidence in the data becomes more extreme in favor of the alternative hypothesis. Instead, their Bayes factors often satisfy $P(BF_{10} > \delta) \to 1$ as the evidence against the null increases (for a fixed sample size n). This aspect of RMPBTs for linear models is apparent from results presented later and is discussed further in the Conclusions and Discussion section.

The remainder of this article is organized as follows. In Section 2, we show that, by restricting the class of prior densities to g priors in the general linear model, we are able to define an RMPBT, and that the value of g has a simple form when the test's rejection region is matched to a classical α-size test. We then specialize this result for ANOVA and t-testing scenarios.

In Section 3, we present two simulation studies to compare the g prior-RMPBT to other Bayesian methods for setting g, and in Section 4, we illustrate an application of our method to a real data set. We conclude in Section 5 with a discussion of the results.

2. g Prior-restricted most powerful Bayesian tests

We begin by considering the general linear model

$$\begin{aligned} \boldsymbol{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad\qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n) \\ &= \mathbb{1}_n \beta_0 + \mathbf{X}_1 \boldsymbol{\beta_1} + \mathbf{X}_2 \boldsymbol{\beta_2} + \boldsymbol{\epsilon}. \end{aligned}$$

We partition $\mathbf{X}$ and β so that tests of hypotheses on subsets of β are performed on the sub-vector $\boldsymbol{\beta}_1$. The prior density that we propose is based on Zellner's g prior Zellner (1986), which in the general linear model leads to

$$\boldsymbol{\beta_1} | g, \sigma^2 \sim \mathcal{N}(\mathbf{0}, g\sigma^2 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}), \quad \text{and} \quad \pi(\sigma^2,\ \beta_0, \boldsymbol{\beta_2}) \propto 1/\sigma^2.$$

If we restrict attention to prior densities of this form, and assume (without loss of generality) that the model has been parameterized in such a way that $\mathbb{1}_n^\top \left[\mathbf{X}_1 \ \mathbf{X}_2 \right] = \mathbf{0}$ and $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$, then the value of g that provides the RMPBT(δ) is given by the following theorem.

Theorem 1. *Suppose that* $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \mathbf{I})$*, and partition* $\boldsymbol{X}$ *and* β *according to* $\mathbf{X} = \left[\mathbb{1}_n \ \mathbf{X}_1 \ \mathbf{X}_2 \right]$ *and* $\beta = \left[\beta_0 \ \beta_1^\top \ \beta_2^\top \right]^\top$*, where* $\mathbf{X}_i$ *has* p_i *columns and* $p = p_1 + p_2$*. Assume* $n > p_2 - 1$*, and that the design matrix has been constructed so that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are of full-column rank,* $\mathbb{1}_n^\top \left[\mathbf{X}_1 \ \mathbf{X}_2 \right] = \mathbf{0}$*, and* $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$*. Assume further that the joint prior distribution on* σ^2*,* β_0*, and* β_2 *is proportional to* $1/\sigma^2$*. If the null hypothesis is* $H_0 : \beta_1 = \mathbf{0}$ *and the alternative hypothesis is restricted to take the form*

$$H_1 : \beta_1 | g, \sigma^2 \sim \mathcal{N}\left(\mathbf{0}, g\sigma^2 \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1}\right) \tag{1}$$

for some value of $g > 0$*, then the RMPBT for evidence threshold* δ *is obtained by setting* g *equal to*

$$\arg\max_G (G)^{-1} \left[\delta^{\frac{-2}{n-p_2-1}} (1+G)^{\frac{n-p-1}{n-p_2-1}} - 1 \right]. \tag{2}$$

This theorem is important because it provides a default alternative hypothesis for constructing a Bayesian test of regression coefficients in a linear model. Although other objective methods for setting g have been proposed (Liang *et al.*, 2008), the RMPBT computed from (2) provides greater probability that the Bayes factor exceeds the given threshold (i.e. has greater power) than any other alternative hypothesis taking the form (1). As we demonstrate in Section 3, the resulting difference in power can often be quite appreciable.

The Bayes factor for the RMPBT can be expressed in terms of g and $\hat{F}$, where $\hat{F}$ is the observed F statistic for the classical test:

$$\hat{F} = \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{X}_1} \mathbf{y} / p_1}{\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_{\mathbb{1}}) \mathbf{y} / (n - p - 1)},$$

where the projection matrices $\mathbf{P}$ are defined in Appendix A.

Given $\hat{F}$, the Bayes factor can be expressed as

$$BF_{10}(\mathbf{y}) = (1+g)^{(n-p-1)/2}\left[1 + g \cdot \frac{n-p-1}{\hat{F}\,p_1 + n - p - 1}\right]^{-(n-p_2-1)/2}. \tag{3}$$

The evidence threshold δ must be determined before g can be computed from (2). In classical terms, the evidence threshold plays a role that is similar to the size of a test; it specifies the value of the Bayes factor required to reject the null hypothesis in favor of the alternative. In the case of UMPBTs, Johnson (2013a) fixed evidence thresholds by equating the rejection regions of UMPBTs and frequentist tests possessing specified type-I error rates. We propose to extend this idea for application to RMPBTs; the next theorem provides a mechanism for doing this.

(Proofs of all theorems are provided in Appendix A.)

Theorem 2. *Under the conditions in Theorem 1, the value of g that produces a g prior-RMPBT that has the same rejection region as a size-α classical F-test is obtained by setting*

$$g = F_{1-\alpha} - 1, \tag{4}$$

where $F_{1-\alpha}$ is the $1-\alpha$ quantile from an F distribution with p_1 and $n-p-1$ degrees of freedom. Moreover, the evidence threshold δ for the RMPBT with this value of g is given by

$$\delta = \left[\frac{p_1 F_{1-\alpha} + n - p - 1}{F_{1-\alpha}^{p_1/(n-p_2-1)}(n-p_2-1)}\right]^{(n-p_2-1)/2}. \tag{5}$$

The Bayes factor for this test can be expressed in terms of $F_{1-\alpha}$ and $\hat{F}$ as

$$BF_{10}(\mathbf{y}) = F_{1-\alpha}^{(n-p-1)/2}\left[\frac{F_{1-\alpha} + \hat{F}\frac{p_1}{n-p-1}}{1 + \hat{F}\frac{p_1}{n-p-1}}\right]^{-(n-p_2-1)/2}. \tag{6}$$

There is an interesting similarity between the expression for g in Theorem 2 and the local empirical Bayes estimate for g described in Liang *et al.* (2008),

$$\hat{g}^{\mathrm{EBL}} = \max\{\hat{F} - 1, 0\}. \tag{7}$$

The RMPBT value for g in (4) is obtained from (7) by substituting $F_{1-\alpha}$ for $\hat{F}$. The implications of this difference are explored in the simulation study presented in Section 3.

For the tests we have been discussing, equation (5) provides a direct method for setting the evidence threshold δ in order to set the test's rejection region equal to that of an α-size test. Doing this facilitates, the comparison between the strength of evidence summarized by a p-value and that summarized by the Bayes factor of an RMPBT. Figure 1 plots the δ threshold against the α level for three tests which differ only in dimension p. The figure illustrates how standard classical test sizes such as $\alpha = 0.05$ correspond to RMPBTs with weak evidence thresholds. For some tests, moderate levels of α even correspond to evidence thresholds lower than $\delta = 1$ (the horizontal gray line), meaning that evidence that on balance *supports* the null hypothesis can induce the classical test to *reject* H_0. This discrepancy between Bayes factors and p-values has been noted previously by many authors, including Berger & Sellke (1987), Edwards *et al.* (1963), Berger & Delampady (1987), Kass & Raftery (1995), and Johnson (2013b), and argues in favor of requiring more stringent criteria for rejecting tested null hypotheses in frequentist testing.

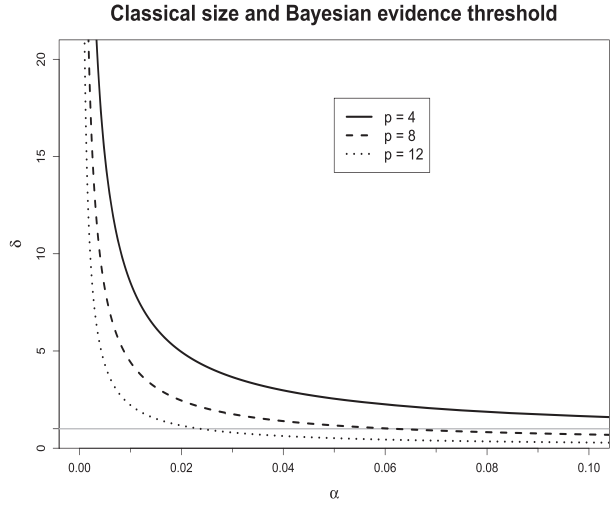

Fig. 1. Plot depicting the relationship, as given in Theorem 2, between classical test size α and evidence threshold δ when the rejection regions of the classical test and RMPBT are set equal to one another. In each case, $n = 60$ and $p_1 = p_2 = 0.5p$. A horizontal line is drawn at $\delta = 1$.

We next consider g prior-RMPBTs for two special cases of the general linear model: the one-way analysis of variance (ANOVA) model, and the two-sample t-test. In each case, the simplest parameterization of the model uses a design matrix of the form

$$\mathbf{X} = \left[\mathbb{1}_n \;\; \mathbf{X}_1 \right] = \begin{bmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbb{1}_{n_2} & \mathbf{0} & \mathbb{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}_{n_J} & \mathbf{0} & \mathbf{0} & \dots & \mathbb{1}_{n_J} \end{bmatrix},$$

where $J = 2$ for the two-sample t-test. To make the corresponding model

$$\mathbf{y} = \mathbb{1}_n \beta_0 + \mathbf{X}_1 \beta_1 + \boldsymbol{\epsilon}$$

identifiable, various constraints can be used. One option is to eliminate one column of $\mathbf{X}$ (equivalent to setting one component of β_1 equal to 0). The use of such constraints in the Bayesian setting has generated discussion in Gelman (2005) and Rouder *et al.* (2012). Gelman recommends an alternative constraint $\mathbb{1}^\top \beta_1 = 0$, whereas Rouder *et al.* employ this constraint only for fixed factors. In the following corollaries, we assume that an identifiable parameterization of the design matrix has been specified, although the particular parameterization used is not important as long as the following conditions are satisfied:

(1) the design matrix can be written as $\mathbf{X}^* = \left[\mathbb{1}_n \;\; \mathbf{X}_1^* \right]$ for some $n \times (J-1)$ matrix $\mathbf{X}_1^*$, and
(2) the column space of $\mathbf{X}^*$ is the same as the column space of $\mathbf{X}$ (i.e. the column space of $\mathbf{X}_1^*$ coincides with the column space of $\mathbf{X}_1$).

The parameter vector constraints described by the functions `contr.treatment`, `contr.SAS`, `contr.sum`, `contr.helmert`, and `contr.poly` in R (R Core Team, 2015) are all examples of parameterizations that satisfy these conditions. We define β_0^* and β_1^* as the corresponding regression parameters.

The principal problem in applying Theorem 1 to the one-way ANOVA setting is that the condition $\mathbb{1}^\top \mathbf{X}_1^* = \mathbf{0}$ is not, in general, satisfied by $\mathbf{X}_1^*$. Wetzels *et al.* (2012) resolved this problem by centering the columns in $\mathbf{X}_1^*$ so that the resulting model is

$$\mathbf{y} = \mathbb{1}_n \beta_0^* + (\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\, \mathbf{X}_1^* \beta_1^* + \boldsymbol{\epsilon}, \tag{8}$$

where $\mathbf{P}_{\mathbb{1}} = \frac{1}{n}\mathbb{1}_n \mathbb{1}_n^\top$. It can be shown that

$$(\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^* \beta_1^* = \mathbf{0} \iff \mathbf{X}_1^* \beta_1^* = \mathbf{0} \quad \text{or} \quad \mathbf{X}_1^* \beta_1^* \propto \mathbb{1}_n,$$

and as a result, the test of whether $\beta_1^* = \mathbf{0}$ in model (8) and the classical one-way ANOVA test have the same null hypothesis. For concreteness, we use the Wetzels *et al.* parameterization to state the g prior-RMPBT for one-way ANOVA tests.

Corollary 1. *Assume that*

$$y_{ij} \beta_0 + \beta_j + \epsilon_{ij},$$

where y_{ij} is observation i under treatment j for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ and ϵ_{ij} are independent, mean-zero normally-distributed observational errors with constant variance σ^2. Under the parameterization in (8), assume that the prior density for (σ^2, β_0^*) is given by

$$\pi(\sigma^2, \beta_0^*) \propto 1/\sigma^2.$$

Then the g prior-RMPBT for evidence level δ for testing hypotheses

$$H_0 : \beta_1^* = \mathbf{0}, \qquad \textit{versus} \qquad H_1 : \beta_1^* \mid g, \sigma^2 \sim \mathcal{N}\left(0, g\sigma^2 \left(\mathbf{X}_1^{*\top}(\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\right)^{-1}\right)$$

is obtained by setting g equal to

$$\arg\max_G (G)^{-1}\left[\delta^{\frac{-2}{n-1}}(1+G)^{\frac{n-J}{n-1}} - 1\right].$$

The value of g that produces a g prior-RMPBT that has the same rejection region as a size-α classical F-test is obtained by setting

$$g = F_{1-\alpha} - 1, \tag{9}$$

where $F_{1-\alpha}$ is the $1-\alpha$ from an F distribution with $J-1$ and $n-J$ degrees of freedom, and

$$\delta = \left[\frac{(J-1)F_{1-\alpha} + n - J}{F_{1-\alpha}^{(J-1)/(n-1)}(n-1)}\right]^{(n-1)/2}. \tag{10}$$

The values of g and δ in this corollary do not depend on the particular form of the parameterization of the design matrix because they are not functions of $\mathbf{X}_1^*$. Similarly, the value of the Bayes factor obtained for the g prior-RMPBT is invariant to the choice of design matrix, even though the prior on the regression coefficient β_1^* does depend on the parameterization of the design matrix. The invariance of the Bayes factor to the parameterization of the design matrix follows from its expression as

$$BF_{10}(\mathbf{y}) = F_{1-\alpha}^{(n-J)/2}\left[\frac{F_{1-\alpha} + \hat{F}\frac{J-1}{n-J}}{1 + \hat{F}\frac{J-1}{n-J}}\right]^{-(n-1)/2}, \tag{11}$$

which does not depend on $\mathbf{X}_1^*$.

In addition to the UMPBTs developed in Johnson (2013b), approximate UMPBTs are given therein for one- and two-sample t-tests. However, these approximations fail for large values of the sample mean $\bar{y}$. As an alternative, Corollary 1 can be applied to obtain a g prior-RMPBT for the two-sample t-test as follows.

Corollary 2. *Assume that $y_{ij}|\beta_0, \beta_1, \beta_2, \sigma^2$ are conditionally independent normally distributed random variables with mean $\beta_0 + \beta_j$ and variance σ^2 for $i = 1, \ldots, n_j$ and $j = 1, 2$. Under model (8), let $\pi(\sigma^2, \beta_0^*) \propto 1/\sigma^2$ and suppose that the design matrix satisfies conditions 1 and 2 stated above. For the test of*

$$H_0 : \beta_1^* = \mathbf{0} \quad \textit{versus} \quad H_1 : \beta_1^* | g, \sigma^2 \sim \mathcal{N}\left(0, g\sigma^2 \left(\mathbf{X}_1^{*\top}(\mathbf{I} - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\right)^{-1}\right),$$

the g prior-RMPBT for evidence level δ is obtained by setting g equal to

$$\arg\max_G (G)^{-1}\left[\delta^{\frac{-2}{n-1}}(1+G)^{\frac{n-2}{n-1}} - 1\right].$$

Furthermore, the value of g that produces a g prior-RMPBT that has the same rejection region as a size-α classical t-test is obtained by setting g equal to

$$g = t_{1-\alpha/2}^2 - 1,$$

where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a t distribution with $n-2$ degrees of freedom. Moreover, the evidence threshold δ is given by

$$\delta = \left[\frac{t_{1-\alpha/2}^2 + n - 2}{t_{1-\alpha/2}^{2/(n-1)}(n-1)}\right]^{(n-1)/2}.$$

In the test of Corollary 2, the Bayes factor can be written as a function of the classical t statistic,

$$\hat{t} = \sqrt{\frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{X}_1}\mathbf{y}}{\mathbf{y}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbb{1}})\mathbf{y}/(n-2)}},$$

and a quantile from the t distribution as

$$BF_{10}(\mathbf{y}) = t_{1-\alpha/2}^{n-2}\left[\frac{t_{1-\alpha/2}^2 + \hat{t}^2\frac{1}{n-2}}{1 + \hat{t}^2\frac{1}{n-2}}\right]^{-(n-1)/2}. \tag{12}$$

The previous corollaries describe RMPBTs for the one-way ANOVA and two-sample t-tests. These corollaries follow directly from Theorem 1. However, one-sample t-tests are not a special case of Theorem 1 because these tests are tests of the intercept term (rather than the effect term) in that theorem. Instead, the following theorem describes the g prior-RMPBT for a one-sample t-test. Without loss of generality, we consider only the case of testing $H_0 : \beta_0 = 0$.

Theorem 3. *Assume that $y_i|\beta_0, \sigma^2$ are independent normally distributed random variables with mean β_0 and variance σ^2 for $i = 1, \ldots, n$. Under the priors $\pi(\sigma^2) \propto 1/\sigma^2$ and $\beta_0|g, \sigma^2 \sim \mathcal{N}\left(0, g\sigma^2/n\right)$, the g prior-RMPBT for testing $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ for evidence threshold δ is obtained by setting g equal to*

$$\arg\max_G (G)^{-1}\left[(1+G)^{(n-1)/n}\delta^{-2/n} - 1\right]. \tag{13}$$

The value of g that produces a g prior-RMPBT that has the same rejection region as a size-α classical t-test is obtained by setting g equal to

$$t^2_{1-\alpha/2} - 1,$$

where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a t distribution with $n-1$ degrees of freedom. Moreover, the evidence threshold δ is given by

$$\delta = \left[\frac{t^2_{1-\alpha/2} + n - 1}{t^{2/n}_{1-\alpha/2} n}\right]^{n/2}.$$

The Bayes factor of the one-sample t-test, expressed as a function of the classical t statistic $\hat{t}$ and the corresponding quantile of the t distribution, is

$$BF_{10}(\mathbf{y}) = t^{n-1}_{1-\alpha/2}\left[\frac{t^2_{1-\alpha/2} + \hat{t}^2\frac{1}{n-1}}{1+\hat{t}^2\frac{1}{n-1}}\right]^{-n/2}. \tag{14}$$

3. Experimental results

3.1. Simulation studies

Numerous criteria have been proposed to set the value of g in g prior-based Bayesian hypothesis tests and variable selection methods. The Jeffrey-Zellner-Siow (JZS) prior, which is a multivariate Cauchy prior distribution on β, is discussed in Zellner & Siow (1980). The unit information prior (UIP) described by Kass & Wasserman (1995) sets $g = n$. The risk inflation criterion (RIC) prior (Foster & George, 1994) sets $g = p^2$. The local empirical-Bayes (EBL) prior obtains an estimate of g by maximizing $m(\mathbf{y}|\mathcal{M}_{\boldsymbol{\gamma}})$, the marginal likelihood of $\mathbf{y}$ given model $\mathcal{M}_{\boldsymbol{\gamma}}$. The hyper-$g$ prior described in Liang *et al.* (2008) places a particular prior distribution on g.

In this section, we compare the performance of UIP, RIC, and EBL priors to that of the g prior-RMPBT in a numerical study. To evaluate performance in a basic hypothesis testing problem, the average power at a selected alternative hypothesis is estimated for each method in a simulated testing problem. This study also provides a sensitivity analysis for δ. A second simulation study compares the power functions for the two-sample t-test under the g prior-RMPBT and the approximate UMPBT.

Although the g prior-RMPBT is, by definition, guaranteed to provide the highest probability of rejection at the given evidence level within the class of g prior alternatives, it is interesting to examine the relative power achieved by the other methods and to compare the actual values of g used under each proposal. We emphasize, in making these comparisons, that the expected values of the Bayes factors under various alternatives will often be much higher than it is under the RMPBT; the RMPBT only provides the maximum probability of exceeding a specified evidence threshold.

For simplicity, we restrict attention to a balanced one-way ANOVA test in which the true model for the random effects β^*_1 is given by

$$\beta^*_1 \sim \mathcal{N}\left(\mathbf{0}, g_t\sigma^2(\mathbf{X}^{*\top}_1(\mathbf{I}-\mathbf{P}_1)\mathbf{X}^*_1)\right). \tag{15}$$

Here, g_t is a fixed 'true' value of g, $\sigma^2 = 5$ is the error variance, and $J = 3$ and $n = 15$. The elements of β^*_1 are generated from a centered model such that the values of g_t are on the same scale as the RMPBT value of g. Values of g_t were selected to be equal to the RIC

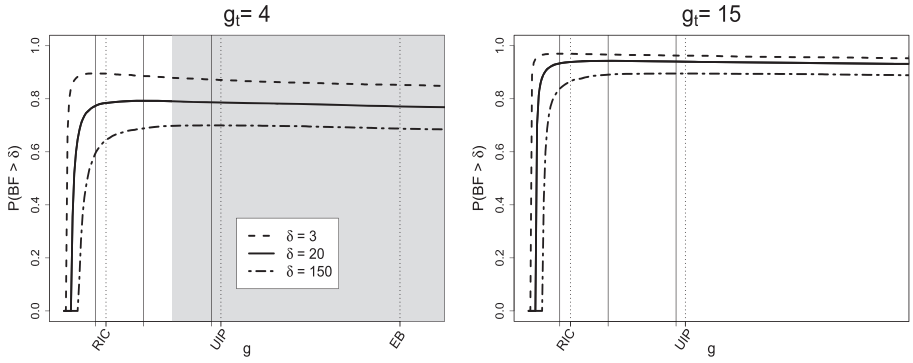

Fig. 2. Numerical simulation results comparing the probability of exceeding a threshold $(\delta = 3, 20, 150)$ for a grid of values of g between 0 and 35. Vertical dotted lines indicate the values of the RIC value of g, the UIP value of g, and the median EBL value of g (missing in the right panel due to being greater than 35). Each vertical solid line indicates an RMPBT value of g, thereby crossing its respective power curve at the peak. The shaded region represents the middle 50 per cent of EBL values of g. In this simulation, $J = 3, n = 15$, and $\sigma^2 = 5$.

value $[(J-1)^2 = 4]$ and the UIC value (n = 15). Figure 2 displays $\mathbf{P}_{g_t}[BF_{10}(\mathbf{y}) > \delta]$ as a function of g from these two experiments, where data were simulated using model (15). The evidence thresholds used in each plot were $\delta = 3, 20, 150$, which are the minimum thresholds for 'positive', 'strong', and 'very strong' evidence according to the modified schedule in Kass & Raftery (1995). Vertical dotted lines are drawn to indicate values of g corresponding to the RIC prior, the UIP prior, and the median EBL prior. Vertical solid lines indicate RMPBT values of g for the three power curves. Also shown is a shaded region which represents the center 50 per cent of $\hat{g}^{\text{EBL}}$ values from the simulation.

Each RMPBT value of g, by definition, corresponds to the peak of its respective curve in each plot. Because of the flatness of the curve, values of g approximately equal to the RMPBT appear to provide comparable power to the RMPBT, as do values moderately greater than it. But small values of g experience a precipitous drop in power. The RIC and UIP tests, defining values of g close to the RMPBT in this simulation, possess nearly as much power for small to moderate δ, and the difference decreases as g_t grows. The variability in results obtained with the local empirical value of g increases quickly as the true value of g_t becomes large.

Figure 2 also provides some sensitivity analysis to the specification of δ. As one would expect, the power curve is uniformly higher for smaller values of δ. These differences diminish as g_t grows because the random effects become easier to detect, no matter what δ is used. The values of g for the RMPBTs vary roughly between those of RIC and UIP, demonstrating that any specification of δ between 3 and 150 amounts to no more sensitivity in the final result to the subjective choices made in running the test than that encountered where researchers choose between using RIC and UIP for g.

Results from this simulation study were also used to compare Bayes factors from the RMPBT to classical p-values through the relation specified in (10) with $\alpha = 0.05$. To this end, g was set according to (9). As a follow-up to Figure 1, the left panel of Figure 3 displays the resulting correspondence between p-values and Bayes factors for this experiment. This plot directly illustrates the tendency for the magnitude of p-values to exaggerate evidence against the null hypothesis, because, for example, evidence against H_0 that crosses the Kass & Raftery threshold for 'strong' ($BF > 20$) requires a p-value of around 0.003. Similar findings with more details are described in Johnson (2013b).

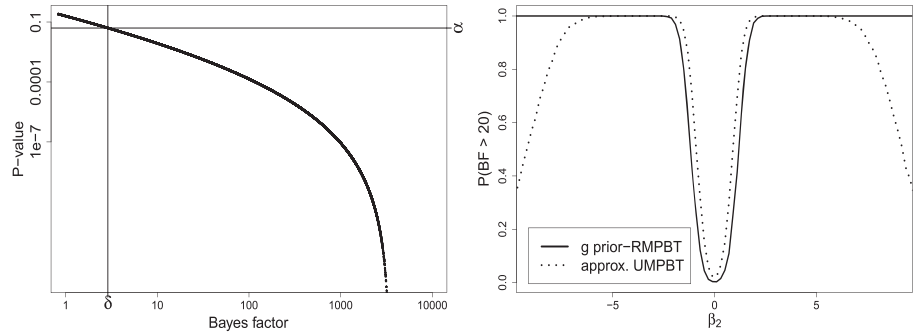

Fig. 3. Left panel: p-values plotted against Bayes factors using the RMPBT in a one-way ANOVA. As in Figure 2, $J = 3$, $n = 15$, $\sigma^2 = 5$, and $g_t = 4$. Now, δ is set to correspond to $\alpha = 0.05$ using (10). Right panel: A two-sided two-sample t-test power curve for the g prior-RMPBT and the approximate UMPBT from Johnson 2013. In this simulation, $n_1 = n_2 = 15$, $\delta = 20$, $\sigma^2 = 1$, and $\beta_0 = \beta_1 = 0$. The power of each test is plotted against a range of β_2 values.

A separate simulation study was used to compare the power curves of the g prior-RMPBT and the approximate UMPBT from Johnson (2013b) in two-sample t-tests. The right panel of Figure 3 displays the resulting power curves. The two-sided tests were simulated 5,000 times at an increasing sequence of β_2 values, where β_1 and β_0 were held fixed at 0. In this experiment, $n_1 = n_2 = 15$ and $\delta = 20$. As expected, the approximate UMPBT outperforms the g prior-RMPBT in terms of power for small to moderate values of β_2; this occurs because the UMPBT alternatives are not restricted to the class of g priors. However, as evidence against the null hypothesis becomes strong, the quality of the approximation to the UMPBT decays and its power declines. The g prior-RMPBT does not suffer from this problem and actually provides higher power for large values of β_2.

3.2. Real life example

To illustrate the use of g prior-RMPBTs on real data, we re-analyzed the seaweed grazer data previously analyzed by Qian & Shen (2007). The experimental design in this study was a randomized complete block design with six treatments (grazers) in eight blocks (intertidal locations) with two replications. The response variable y_{ijk} was the logit of the percentage seaweed recovery in the kth experimental plot (k = 1,2) in block j (j = 1,...,8) under treatment i (i = 1,...,6). An ANOVA model for the experiment can be written as

$$y_{ijk} = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3ij} + \epsilon_{ijk},$$

where β_1 is the vector of treatment effects, β_2 is the vector of block effects, and β_3 is the vector of interactions. The elements in the vector $\boldsymbol{\epsilon}$ are assumed to be i.i.d. mean-zero normal random variables.

We begin by testing the interaction effect, β_3. In the notation of Theorem 1, we have $n = 96$ and $p = p_1 + p_2 = 35 + (5 + 7)$, so that with the intercept β_0 there are 48 parameters in the model. If we set g and δ so that the RMPBT corresponds to a 5 per cent classical test, then $g = 0.67$ and $\delta = 3.83$. The Bayes factor for the resulting RMPBT test is 1.76 and the p-value is 0.12. Under the added assumption of equipoise, the posterior probability in favor of the null hypothesis is 0.36.

The main effects are tested next. The Bayes factor of the treatment effect is 1.7×10^7, and that of the blocking effect is 1.4×10^6. The corresponding p-values are 4.5×10^{-20} and 5.4×10^{-17},

Table 1. *Results of the RMPBT and frequentist test of the seaweed grazers data.*

Effect	g	BF_{10}	δ	p-value	$\mathbf{P}(H_0\|\mathbf{y})$
Treatment	1.41	1.7×10^7	3.0	4.5×10^{-20}	6.0×10^{-8}
Block	1.21	1.4×10^6	3.2	5.4×10^{-17}	7.2×10^{-7}
Interaction	0.67	1.76	3.8	0.1209	0.3627

respectively, with null hypothesis posterior probabilities of 6.0×10^{-8} and 7.2×10^{-7}. These results are summarized in Table 1.

We emphasize that the Bayes factors cited in this example were obtained through straightforward calculations that were based only on the F statistics reported from standard ANOVA software. Indeed, an R function to compute these values is described in Appendix B. RMPBT methodology thus provides a simple mechanism for converting classical test statistics and p-values into Bayes factors. This methodology also makes explicit the alternative hypothesis that is implicitly being tested in a significance test, and provides practitioners with an estimate of the posterior probability that both the null and alternative hypotheses are true, given the prior probabilities they assign to the truth of each hypothesis.

4. Conclusion and discussion

The conditions in Theorem 1 encompass many of the ANOVA, ANCOVA, and linear regression tests performed in practice. By calibrating δ to provide an α-size test, the g prior-RMPBT provides an alternative quantification of the evidence against the null hypothesis, as well as a description of the weight of evidence in favor of it. For this reason, we view the g prior-RMPBT as a supplement to the classical F-test. Under an assumption of equipoise, the g prior-RMPBT provides an objective estimate of the posterior probabilities of the null and alternative hypotheses, quantities that in many cases are of primary interest to practitioners. A simple R function, `aov_mpb`, that provides these estimates is described in Appendix B.

There remain some additional considerations concerning g prior-RMPBT's properties and behavior. For one, RMPBTs and UMPBTs optimize over all possible values of $\boldsymbol{\theta}_t$, the data-generating value of the parameter being tested. This optimization results in a maximization of the probability that the Bayes factor exceeds δ for values of $\boldsymbol{\theta}_t$ that satisfy H_0. Philosophical objections to this facet of these tests should be balanced against the fact that the probability of a false rejection is still controllable through the specification of δ. These tests' principal virtue lies in the objective and default alternative hypotheses they define, together with the supplemental information they provide to frequentist test results.

Also, the tests described herein are susceptible to the 'information paradox' described in Liang *et al.* (2008), wherein, for some test statistic T_n, the Bayes factor $BF(T_n)$ in favor of H_1 fails to diverge as $T_n \to \infty$ for a fixed value of n. In other words, as evidence mounts against H_0, the Bayes factor 'tops out' at some finite value. Of course, the Bayes factors expressed in (3), (6), (11), (12), and (14) converge to expressions that grow exponentially with n as their test statistics increase, which, for moderate sample sizes and moderate sizes of δ, may lessen our concern over this information paradox. It is also important to note the RMPBTs are defined to maximize the probability their Bayes factors exceed a specified threshold, and this probability does converge to 1 for finite n as T_n grows in the Bayes factors provided in Section 2.

Likewise, as evidence against H_0 diminishes, i.e. $T_n \to 0$, the Bayes factors in (6), (11), (12), and (14) approach a non-zero constant. In the worst cases, the one-sample and two-sample t-test Bayes factors in (12) and (14) converge to the inverse of a tail-area quantile; in the other cases, the lower bound is a positive power of such. These consistency problems stem, in part, from the fact that the rejection regions of these tests have been set equal to those of classical F

tests and t tests, which are known to possess the same issues (Berger & Sellke, 1987). This is evident from the fact that the Bayes factor in (3), which does not equate rejection regions to a classical test, converges to 0 as $\hat{F} \to 0$, for fixed n. Further discussion of this issue is provided in Johnson (2013a).

Finally, the results described in this article depend on the use of the Normal-Gamma g prior on model coefficients, which restricts their applicability. Although the g prior has found wide and extensive application in Bayesian model averaging (Feldkircher & Zeugner, 2009) and variable selection methods, RMPBTs may be sought for other classes of priors, including non-local priors (Johnson & Rossell, 2010). Finally, the extension of RMPBTs to non-linear models is currently under investigation.

Acknowledgements

The authors wish to thank two anonymous reviewers for valuable comments and suggestions that have improved this article's content and presentation.

References

American Statistical Association Committee on Training of Statisticians for Industry. (1980). Preparing statisticians for careers in industry. *Amer. Statist.* **34**, 65–80.

Berger, J. O. & Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–335.

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: Irreconcilability of p-values and evidence. *J. Amer. Statist. Assoc.* **82**, 112–122.

Edwards, W., Lindman, H. & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70**, 193–242.

Feldkircher, M. & Zeugner, S. (2009). Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging. IMF working paper WP/09/202, International Monetary Fund, Washington DC, USA.

Foster, D. & George, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

Gelman, A. (2005). Analysis of variance–why it is more important than ever. *Ann. Statist.* **33**, 1–53.

Hicks, C. & Turner, K. (1999). *Fundamental concepts in the design of experiments*, (5th ed.)., Oxford University Press, New York City, NY.

Johnson, V. E. (2013a). Uniformly most powerful Bayesian tests. *Ann. Statist.* **41**, 1716–1741.

Johnson, V. E. (2013b). Revised standards for statistical evidence. In *Proc. Natl. Acad. Sci.* USA **110**, 19313–19317.

Johnson, V. E. & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72**, 143–170.

Kass, R. & Raftery, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Kass, R. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *J. Amer. Statist. Assoc.* **90**, 928–934.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410–423.

Qian, S. & Shen, Z. (2007). Ecological applications of multilevel analysis of variance. *Ecology* **88**, 2489–2495.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.

Rouder, J., Morey, R., Speckman, P. & Province, J. (2012). Default Bayes factors for ANOVA designs. *J. Math. Psych.* **56**, 356–374.

Wetzels, R., Grasman, R. & Wagenmakers, E. (2012). A default Bayesian hypothesis test for ANOVA designs. *Amer. Statist.* **66**, 104–111.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g prior distributions. In *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (eds Goes, P. K. & Zellner, A.), 233–243. North-Holland/Elsevier, Amsterdam.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian statistics: Proceedings of the first international meeting held in Valencia* (eds Bernardo, J. M., DeGroot, M. H., Lindley, D. V. & Smith, A. F. M.), 585–603. University of Valencia Press, Valencia.

Received January 2015, in final form March 2016

Scott D. Goddard, Department Mathematics and Statistics, University of Alaska Fairbanks, P.O. Box 756660, Fairbanks, AK 99775-6660, USA.
E-mail: sdgoddard@alaska.edu

Appendix A: Proofs of theorems

Proof 1. Under the alternative hypothesis, $m_1(\mathbf{y})$, the marginal density is given by

$$m_1(\mathbf{y}) = (2\pi)^{-(n-p_2-1)/2}\frac{(1+g)^{-p_1/2}}{\sqrt{n}}|\mathbf{X}_2^\top\mathbf{X}_2|^{-1/2}\frac{\Gamma((n-p_2-1)/2)}{\left\{\frac{1}{2}\mathbf{y}^\top(\mathbf{I}-\frac{g}{1+g}\mathbf{P}_{\mathbf{X}_1}-\mathbf{P}_{\mathbf{X}_2}-\mathbf{P}_{\mathbb{1}})\mathbf{y}\right\}^{(n-p_2-1)/2}},$$

where $\mathbf{P}_{\mathbf{X}_i} = \mathbf{X}_i(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\mathbf{X}_i^\top$ and $\mathbf{P}_{\mathbb{1}} = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$. Under the null hypothesis, the marginal density is

$$m_0(\mathbf{y}) = (2\pi)^{-(n-p_2-1)/2}\frac{1}{\sqrt{n}}|\mathbf{X}_2^\top\mathbf{X}_2|^{-1/2}\frac{\Gamma((n-p_2-1)/2)}{\left\{\frac{1}{2}\mathbf{y}^\top(\mathbf{I}-\mathbf{P}_{\mathbf{X}_2}-\mathbf{P}_{\mathbb{1}})\mathbf{y}\right\}^{(n-p_2-1)/2}}.$$

Therefore, the Bayes factor in favor of the alternative is

$$BF_{10}(\mathbf{y}) = (1+g)^{(n-p-1)/2}\left[1+g\frac{1-R_1^2}{1-R_0^2}\right]^{-(n-p_2-1)/2},$$

where R_i^2 is the coefficient of determination for the model in hypothesis i. The probability of the Bayes factor exceeding a threshold can be expressed as

$$\mathbf{P}_{\boldsymbol{\beta},\sigma^2}\left(\frac{1-R_1^2}{1-R_0^2} < \frac{\delta^{\frac{-2}{n-p_2-1}}(1+g)^{\frac{n-p-1}{n-p_2-1}}-1}{g}\right).$$

This probability is maximized by maximizing the right-hand side of the inequality in g, regardless of the distribution of the left-hand side. □

Proof 2. The rejection region for the frequentist test is

$$\left\{\mathbf{y} : \hat{F} > F_{1-\alpha}\right\},$$

where $\hat{F}$ is the test statistic and the constant $F_{1-\alpha}$ is the $1-\alpha$ quantile of an F distribution with p_1 and $n-p-1$ degrees of freedom. For the Bayesian test using a g prior, the rejection region is

$$\left\{\mathbf{y} : \frac{1-R_1^2}{1-R_0^2} < g^{-1}\left[\delta^{\frac{-2}{n-p_2-1}}(1+g)^{\frac{n-p-1}{n-p_2-1}}-1\right]\right\},$$

which can be expressed as

$$\left\{\mathbf{y} : \hat{F} > c\right\},$$

where the constant c equals

$$\left(\frac{n-p-1}{p_1}\right)\cdot\left(\frac{1+g-\delta^{-2/(n-p_2-1)}(1+g)^{(n-p-1)/(n-p_2-1)}}{\delta^{-2/(n-p_2-1)}(1+g)^{(n-p-1)/(n-p_2-1)}-1}\right).$$

The rejection region for the Bayesian test can therefore be made equivalent to that of the frequentist test by setting $F_{1-\alpha}=c$. Solving for δ, we obtain

$$\delta^{2/(n-p_2-1)}=\frac{(1+g)^{(n-p-1)/(n-p_2-1)}(p_1F_{1-\alpha}+n-p-1)}{(n-p-1)(1+g)+p_1F_{1-\alpha}}.$$

This is the value of δ which gives a size-α test, given g.

By differentiating and equating to 0 the expression in (2), we obtain another expression for δ in terms of g:

$$\delta^{2/(n-p_2-1)}=(1+g)^{(n-p-1)/(n-p_2-1)}-\frac{g(1-\frac{p_1}{n-p_2-1})}{(1+g)^{p_1/(n-p_2-1)}}.$$

Solving for g and δ completes the proof. □

Proof 3. We will show that the conditions given in this corollary satisfy the conditions in Theorem 1. Letting $\mathbf{X}_2=\mathbf{0}$, $\boldsymbol{\beta}_2=\mathbf{0}$, and $p_2=0$, define $p=p_1=J-1$. It is easily seen that $\mathbb{1}^\top(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*=\mathbf{0}$. It only remains to show that $(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*$ is of full-column rank, or rank $J-1$. A rearrangement of the rank-nullity theorem gives

$$\operatorname{rank}\left((\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\right)=J-1-\operatorname{nullity}\left((\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\right).$$

We must show that the dimension of the null space of $(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*$ is 0, or equivalently that, for any vector $\mathbf{a}\in\mathbb{R}^{J-1}$,

$$(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\mathbf{a}=\mathbf{0}\quad\Longrightarrow\quad\mathbf{a}=\mathbf{0}.$$

Fix a vector $\mathbf{a}$ such that $(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*\mathbf{a}=\mathbf{0}$. Because the null space of $(\mathbf{I}_n-\mathbf{P}_{\mathbb{1}})$ is spanned by $\mathbb{1}_n$, and $\mathbf{X}_1^*\mathbf{a}$ is in that null space, there must be some constant b such that

$$\mathbb{1}_nb=\mathbf{X}_1^*\mathbf{a}.$$

But the reparameterization of the model ensured that $\mathbb{1}_n$ was not linearly dependent on the columns of $\mathbf{X}_1^*$, so it must be that $b=0$ and $\mathbf{a}=\mathbf{0}$.

The proof to the corollary follows from Theorems 1 and 2. □

Proof 4. The proof follows from corollary 1 using $J=2$ and the fact that the $1-\alpha$ quantile from an F distribution with 1 and $n-2$ degrees of freedom is equivalent to the square of the $1-\alpha/2$ quantile of a t distribution with $n-2$ degrees of freedom. □

Proof 5. Under the alternative hypothesis, the marginal density is given by

$$m_1(\mathbf{y})=(2\pi)^{-n/2}(1+g)^{-1/2}\frac{\Gamma(n/2)}{\left\{\frac{1}{2}\left[\sum_{i=1}^n y_i^2-\frac{g}{1+g}n\bar{y}^2\right]\right\}^{n/2}}.$$

Under the null hypothesis, the marginal density is

$$m_0(\mathbf{y}) = (2\pi)^{-n/2} \frac{\Gamma(n/2)}{\left\{\frac{1}{2}\sum_{i=1}^n y_i^2\right\}^{n/2}}.$$

Therefore, the Bayes factor in favor of the alternative is

$$(1+g)^{(n-1)/2}\left[1 + g\frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2}\right]^{-n/2},$$

and the probability of the Bayes factor exceeding a threshold can be expressed as

$$\mathbf{P}_{\beta_0,\sigma^2}\left\{\frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2} < g^{-1}\left[(1+g)^{(n-1)/n}\delta^{-2/n} - 1\right]\right\}.$$

This probability can be maximized by maximizing the expression on the left side. The rejection region of the frequentist test is

$$\left\{\mathbf{y} : |\hat{t}| > t_{1-\alpha/2}\right\},$$

where $\hat{t}$ is the test statistic and the constant $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a t distribution with $n-1$ degrees of freedom. For the Bayesian test, the rejection region is

$$\left\{\mathbf{y} : (1+g)^{(n-1)/2}\left[1 + g\frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2}\right]^{-n/2} > \delta\right\},$$

which is equivalent to

$$\left\{\mathbf{y} : |\hat{t}| > c\right\},$$

where the constant c equals

$$\left[\frac{g}{(1+g)^{(n-1)/n}\delta^{-2/n}} - 1\right]^{1/2}(n-1)^{1/2}.$$

Letting $t_{1-\alpha/2} = c$ and solving for δ yields

$$\delta^{2/n} = (1+g)^{(n-1)/n}\frac{t_{1-\alpha/2}^2 + n - 1}{t_{1-\alpha/2}^2 + n - 1 + g(n-1)}.$$

Differentiating (13) and setting the result to zero leads to

$$\delta^{2/n} = \left[(1+g)^{(n-1)/n} - \frac{g}{(1+g)^{1/n}}\cdot\frac{n-1}{n}\right]^{n/2}.$$

Solving for g and δ completes the proof. □

Appendix B: R code

The `UMPBT` package in R has functions for performing tests described here and in Johnson (2013a, 2013b). The g prior-RMPBT for ANOVA testing is performed using the `aov_mpb` function:

```
aov_mpb(formula, data=NULL, alpha=0.05, gamma=NULL, contrasts=NULL)
```

For an example, we use the `batteries` data set from Hicks & Turner (1999), which is also included in the `UMPBT` package. It consists of the response variable `life` and an explanatory variable `temp`.

```
> aov_mpb(life~temp, data=batteries)
           BF    gamma      p.value
temp 46083.43 3.064042 3.146372e-13
```

The function produces the Bayes factor in favor of the alternative hypothesis as well as the p-value, which matches the p-value given by the `aov` function. Specification of the formula argument follows the syntax of the generic `formula` function in R.

We illustrate the analysis of a $3 \times 3 \times 4$ crossed factorial experiment with the data set `rubber`, also from Hicks & Turner (1999) and also included in the `UMPBT` package.

```
> aov_mpb(time~temp*lab*mix,data=rubber)
                        BF    gamma      p.value
temp         1.643114e+09 2.594327 0.000000e+00
lab          1.409900e+06 2.818816 6.217249e-15
mix          9.583649e+04 2.594327 3.569072e-10
temp:lab     3.649044e+03 3.180965 4.722377e-08
temp:mix     1.357029e+03 2.972724 1.150316e-06
lab:mix      1.164397e+02 3.180965 1.394283e-04
temp:lab:mix 6.779397e+01 3.510330 3.675729e-04
```