Journal of Mathematical Psychology 56 (2012) 356–374

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Default Bayes factors for ANOVA designs

Jeffrey N. Rouder [a,*], Richard D. Morey [b], Paul L. Speckman [c], Jordan M. Province [a]

[a] Department of Psychological Sciences, University of Missouri, United States

[b] Faculty of Behavioural and Social Sciences, University of Groningen, The Netherlands

[c] Department of Statistics, University of Missouri, United States

ARTICLE INFO

Article history:
Received 14 December 2011
Received in revised form
3 July 2012
Available online 31 August 2012

Keywords:
Bayes factor
Model selection
Bayesian statistics
Linear models

ABSTRACT

Bayes factors have been advocated as superior to *p*-values for assessing statistical evidence in data. Despite the advantages of Bayes factors and the drawbacks of *p*-values, inference by *p*-values is still nearly ubiquitous. One impediment to the adoption of Bayes factors is a lack of practical development, particularly a lack of ready-to-use formulas and algorithms. In this paper, we discuss and expand a set of default Bayes factor tests for ANOVA designs. These tests are based on multivariate generalizations of Cauchy priors on standardized effects, and have the desirable properties of being invariant with respect to linear transformations of measurement units. Moreover, these Bayes factors are computationally convenient, and straightforward sampling algorithms are provided. We cover models with fixed, random, and mixed effects, including random interactions, and do so for within-subject, between-subject, and mixed designs. We extend the discussion to regression models with continuous covariates. We also discuss how these Bayes factors may be applied in nonlinear settings, and show how they are useful in differentiating between the power law and the exponential law of skill acquisition. In sum, the current development makes the computation of Bayes factors straightforward for the vast majority of designs in experimental psychology.

1. Introduction

Psychological scientists routinely use data to inform theory. It is common to report *p*-values from *t*-tests and *F*-tests as evidence favoring certain theoretical positions and disfavoring others. There are a number of critiques of the use of *p*-values as evidence, and we join a growing chorus of researchers who advocate the Bayes factor as a measure of evidence for competing positions (Edwards, Lindman, & Savage, 1963; Gallistel, 2009; Kass, 1993; Myung & Pitt, 1997; Raftery, 1995; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007). Even though many of us are convinced that Bayes factor is intellectually more appealing that inference by *p*-values, there is a pronounced lack of detailed development of Bayes factors for real-world experimental designs common in psychological science. Perhaps the problem can be illustrated by a recent experience of the first author. After giving a colloquium talk comparing Bayes factors to *p*-values, he was approached by an excited colleague asking for help computing a Bayes factor for a run-of-the-mill three-way ANOVA design. At the time, the first author did not know how to compute this Bayes factor. After all, there were no books that covered it, and the computation was not built into any commonly-used software.

Although the Bayes factor is conceptually straightforward, the computation requires a specification of priors over all parameters and an integration of the likelihood with respect to these priors. Useful priors should exhibit two general properties. First, they should be judiciously chosen because the resulting Bayes factors depends to some degree on the prior. Second, they should be computationally convenient so that the integration of the likelihood is stable and relatively fast. Showing that the priors are judicious and convenient entails much development. Substantive researchers typically have neither the skills nor the time to develop Bayes factors for their own choice of priors. To help mitigate this problem, we provide *default priors* and associated Bayes factors for common research designs. These default priors are general, broadly applicable, computationally convenient, and lead to Bayes factors that have desirable theoretical properties. The defaults priors may not be the best choice in all circumstances, but they are reasonable in most.

The topic in this paper is the development of default Bayes factors for the linear model underlying ANOVA and regression. In experimental psychology there is a distinction between linear models, which are used to assess the effects of manipulations, and domain-specific models of psychological processes. Linear models are simple and broadly applicable, whereas process models are typically nonlinear, complex, and targeted to explore specific phenomena, processes, or paradigms. In many cases, an ultimate

* Correspondence to: 210 McAlester Hall, Columbia, MO 65211, United States.
E-mail address: rouderj@missouri.edu (J.N. Rouder).

0022-2496/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2012.08.001

goal is the development of Bayes factor methods for comparing competing process models. Given this distinction and the appeal of process models, it may seem strange that the majority of the development here is for linear models. There are three advantages to this development. First, ANOVA and regression are still the most popular tests in experimental psychology. Developing Bayes factors for these models is a necessary precursor for widespread adoption of the method. In this paper we provide development for many ANOVA designs, including within-subject, between-subject and mixed designs. Second, many nonlinear models have linear subcomponents. Linear subcomponents may be used to account for nuisance variation in the sampling of participants or items. For example, Pratte and Rouder (2011) fit Yonelinas' dual process recognition-memory model (Yonelinas, 1999) to real-world recognition-memory data where each observation comes from a unique cross of people and items. To fit the model, Pratte and Rouder placed additive linear models on critical mnemonic parameters that incorporated people and items as additive random effects. In cases such as this, development of Bayes factors for inference with linear models is a natural precursor to development for nonlinear models. Third, the priors suggested here may transfer well to nonlinear cases. We provide an example of this transfer by developing Bayes factors to test between the power law and the exponential law of skill acquisition.

This paper is organized as follows. In the next section, we review common critiques of null hypothesis significance testing, which lead naturally to consideration of the Bayes factor. In Section 3, the Bayes factor is presented, along with a discussion of how it should be interpreted when assessing the evidence from data for competing positions. Following this discussion, we discuss the properties of good default priors, and provide default priors for the one-sample case. These existing default priors are then generalized for several effects in Sections 5 and 6. In Sections 7 and 8, we present Bayes factors for one-way and multi-way ANOVA, respectively, for both random and fixed effects. In Section 9, we discuss how within-subject, between-subject and mixed designs may be analyzed. In Section 10 we provide an example from linguistics that is known to be particularly problematic. In linguistic designs, both items and participants should be treated simultaneously as random effects, and failure to do so substantially affects the quality of inference (Clark, 1973). We show how this treatment may be accomplished in a straightforward fashion with the developed Bayes factor methodology. Sections 11–14 provide discussion about the large-sample properties of the Bayes factors, alternative choices for priors, solutions for regression designs, and a discussion of computational issues, respectively. In Section 15, we discuss how the developed priors may be extended for nonlinear cases, and provide an example in assessing learning curves.

2. Critiques of significance testing

It has often been noted that there is a fundamental tension between null hypothesis significance testing and the goals of science. On the one hand, researchers seek simplicity or parsimony to explain target phenomena. An example of such simplicity comes from the work of Gilovich, Vallone, and Tversky (1985), who assessed whether basketball shooters display hot and cold streaks in which the outcome of one shot attempt affects the outcome of subsequent ones. They concluded that there was no such dependency, which is a conclusion in favor of simplicity over complexity. In null hypothesis significance tests, the simpler model which serve as nulls may only be rejected and never affirmed. Hence, researchers using significance testing find themselves on the "wrong side" of the null hypothesis whenever they argue for the null hypothesis. If the null is true, the best case outcome of a significance test is a statement about a lack of evidence for an effect. It would be desirable to state positive evidence for a lack of an effect.

Being on the wrong side of the null is not rare. Other examples include tests of subliminal perception (perception must be shown to be at chance levels, e.g., Dehaene et al., 1998; Murphy & Zajonc, 1993), expectancies of an equivalence of performance across group membership (such as gender, e.g., Shibley Hyde, 2005), or assessment of a lack of interaction between factors (e.g., Sternberg, 1969). Additionally, models that predict stable relationships, such as the Fechner–Weber Law,[1] serve as null hypotheses. Researchers who test strong theoretical positions that predict specified invariances or regularities in data are typically on the wrong side of the null. From a theoretical point of view, being on the wrong side of the null is an enviable position: the goal of scientific theory is often to model or explain observed invariances. Testing strong invariances often indicates a high level of theoretical sophistication. From a practical point of view, however, being on the wrong side of the null presents statistical difficulties. This tension, that null hypotheses are theoretically desirable yet are impossible to support by significance testing, has been noted repeatedly (Gallistel, 2009; Kass, 1993; Raftery, 1995; Rouder et al., 2009).

The asymmetry in significance testing in which the null may be rejected but not supported is a staple of introductory statistics courses. Yet, it has a subtle but pervasive implication that is often overlooked: significance tests overstate the case against the null (Berger & Berry, 1988; Edwards et al., 1963; Wagenmakers, 2007). This bias is highly problematic because it means that researchers may reject the null without substantial evidence against it. The following argument, adapted from Sellke, Bayarri, and Berger (2001), demonstrates this bias. Consider the distributions of p-values under competing hypotheses (Fig. 1(A)). If the null hypothesis is false, then p-values tend to be small, and decrease (in distribution) as sample size increases. The dashed line colored green shows the distribution of p-values when the underlying effect size is 0.2 and the sample size is 50; the dashed–dotted line colored red shows the same when the sample size is increased to 500. The distribution of p-values under the null, however, is quite different. Under the null, all p-values are equally likely (solid line colored blue in Fig. 1(A)). This uniform distribution under the null hypothesis holds regardless of sample size.

If the null is rejected by significance testing, then, presumably, the observed data are more improbable under the null than under some other point alternative. A reasonable measure of evidence is the factor by which the data are more probable under this alternative than under the null. Suppose a data set with sample size of 50 yields a p-value in the interval between 0.04 and 0.05, which is sufficiently small by convention to reject the null hypothesis. Fig. 1(B) shows the distributions of p-values around this interval for the null and the alternative that the effect size $= 0.2$. The probabilities that the p-value will fall in the interval are represented by the shaded areas under the curves, which are 0.01 and 0.04 under the null and alternative hypotheses, respectively. The ratio is $0.04/0.01 = 4$: the probability of the observed p-value is four times more likely under the alternative than under the null. Although such a ratio constitutes evidence for the alternative, it is not as substantial as might be mistakenly inferred by the fact that the p-value is less than 0.05.

Fig. 1(C) shows a similar plot for the null and alternative (effect size $= 0.2$) for a large sample size of 500. For this effect size

[1] The Fechner–Weber Law (Fechner, 1966; Masin, Zudini, & Antonelli, 2009) describes how bright a flash must be to be detected against a background. If the background has intensity I, the flash must be of intensity $I(1+\theta)$ to be detected. The parameter θ, the Weber fraction, is posited to remain invariant across different background intensities.

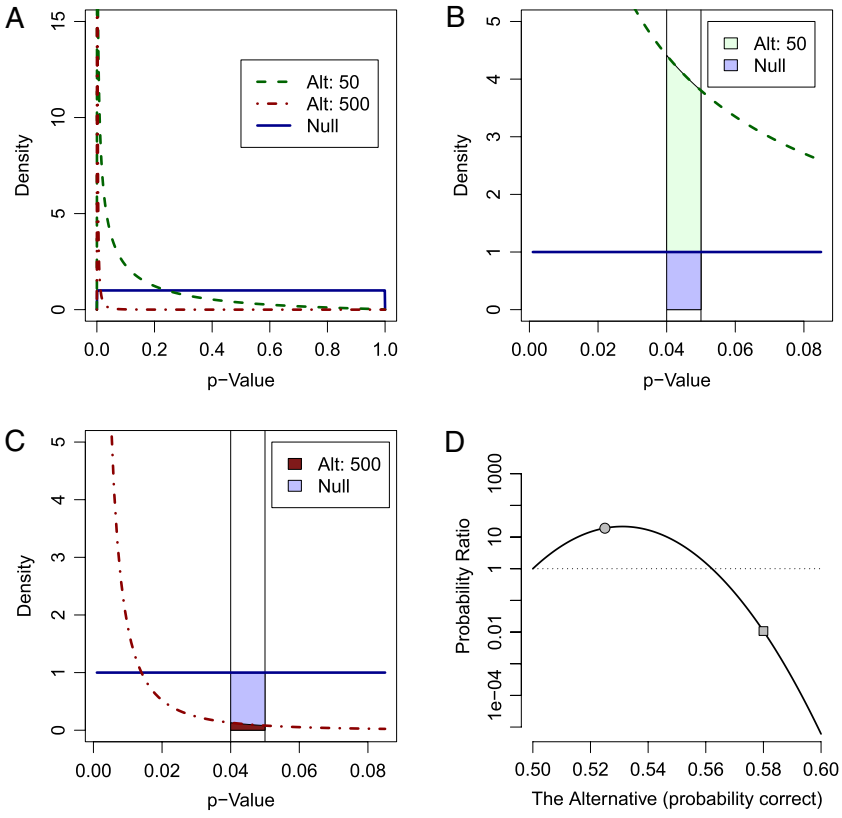

Fig. 1. Significance tests overstate the evidence against the null hypothesis. A. The distribution of p-values for an alternative with effect-size of 0.2 (dashed and dashed–dotted lines are for sample sizes of 50 and 500, respectively) and the null (solid line). B. Probability of observing a p-value between 0.04 and 0.05 for the alternative (effect size = 0.2) and null for $N = 50$. The probability favors the alternative by a ratio of about 4 to 1. C. Probability of observing a p-value between 0.04 and 0.05 for the alternative (effect size = 0.2) and null for $N = 500$. The probability favors the null by a factor of 10. D. The probability ratio as a function of alternative. The probability ratio is the probability of observing a t-value of 2.51 and $N = 100$ given an alternative divided by the probability of observing this t-value for $N = 100$ given the null. The circle and square points highlight alternatives for which the ratios favor the alternative and null, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and sample size, very small p-values are the norm. Let's again suppose we observe a p-value between 0.04 and 0.05, which leads conventionally to a rejection of the null hypothesis. The probability of observing this p-value under the null remains at 0.01. But the probability of observing it under the alternative with such a large sample size is close to 0.001. Therefore, observing a p-value between 0.04 and 0.05 is about ten times more likely under the null than under the alternative.[2] This behavior of significance testing in which researchers reject the null even though the evidence overwhelmingly favors it is known as *Lindley's paradox* (Lindley, 1957), and is a primary critique of inference by p-values in the statistical literature.

In Fig. 1(B) and (C), we compared the evidence for the null against an alternative in which the effect size under the alternative was a specific value (0.2). One could ask about these probability ratios for other effect sizes. Consider a recent study of Bem (2011), who claims that people may feel or sense future events that could not be known without psychic powers. In his Experiment 1, Bem asks 100 participants to guess which of two erotic pictures will be shown at random, and finds participants have an accuracy of 0.531, which is significantly above the chance baseline value of 0.50 ($t(99) = 2.51$; $p < 0.007$). Such small p-values are conventionally interpreted as sufficient evidence to reject the null. Fig. 1(D), solid line, shows the probability that the p-value falls between 0.0065 and 0.0075 under a specific alternative relative to that under the null. These ratios vary greatly with the choice of alternative. Alternatives that are very near the null hypothesis of 0.5 – say, 0.525 – are preferred over the null (filled circle in Fig. 1(D)). Alternatives further from 0.5, say 0.58 (filled square) are definitely not preferred over the null. Note that even though the null is rejected at $p = 0.007$, there is only a small range of alternatives where the probability ratio exceeds 10, and for no alternative does it exceed 25, much less 100 (as might naively be inferred from a p-value less than 0.01). We see that the null may be rejected by p-values even when the evidence for every specific point alternative is more modest. Note that the critique that p-values overstate the evidence is not dependent on a Bayesian perspective, and that the probabilities and probability ratios in Fig. 1 are used as measures of evidence within the frequentist paradigm, where they are called likelihood ratios (Hacking, 1965; Royall, 1997).

3. The Bayes factor

The probability ratio in Fig. 1(D) can be generalized to the *Bayes factor* as follows. Let B_{01} denote the Bayes factor between Models $\mathcal{M}_0$ and $\mathcal{M}_1$. For discretely distributed data,

$$B_{01} = \frac{\Pr(\text{Data}|\mathcal{M}_0)}{\Pr(\text{Data}|\mathcal{M}_1)}.$$

For continuously-distributed data, these probabilities are replaced with probability densities. We use the term probability loosely in the development to refer either to probability mass or to probability density, depending on whether the data are discrete

[2] More generally, a p-value at any nonzero point, say 0.05, constitutes increasing evidence for the null in the large sample-size limit.

or continuous. We use subscripts on Bayes factors to refer to the models begin compared, with the first and second subscript referring to the model in the numerator and denominator, respectively. Accordingly, the Bayes factor for the alternative relative to the null is denoted B_{10}, $B_{10} = 1/B_{01}$.

When models are parameterized,

$$B_{01} = \frac{\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \Pr(\text{Data}|\mathcal{M}_0, \boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1} \Pr(\text{Data}|\mathcal{M}_1, \boldsymbol{\theta})\pi_1(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ are the parameter spaces for Models $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively, and π_0 and π_1 are the *prior* probability density functions of the parameters for the respective models. These priors describe the researcher's prior belief or uncertainty about the parameters. The specification of priors is critical to defining models, and is the point where subjective probability enters the computation of Bayes factor. The argument for subjective probability is made most elegantly in the psychological literature by Edwards et al. (1963), to whom we refer the interested reader. Readers interested in the axiomatic foundations of subjective probability are referred to Cox (1946), De Finetti (1992), and Jaynes (1986). The numerator and denominator are also called the *marginal likelihoods* as they are the integral of the likelihood functions with respect to the priors.

Bayes factors describe the relative probability of data under competing positions. In Bayesian statistics, it is possible to evaluate the relative odds of the positions themselves, conditional on the data:

$$\frac{\Pr(\mathcal{M}_0|\text{Data})}{\Pr(\mathcal{M}_1|\text{Data})} = B_{01} \times \frac{\Pr(\mathcal{M}_0)}{\Pr(\mathcal{M}_1)},$$

where the $\Pr(\mathcal{M}_0|\text{Data})/\Pr(\mathcal{M}_1|\text{Data})$ and $\Pr(\mathcal{M}_0)/\Pr(\mathcal{M}_1)$ are posterior and prior odds, respectively. The prior odds describe the beliefs about the models before observing the data. The Bayes factor, then, describes how the evidence from the data should change beliefs. For example, a Bayes factor of $B_{01} = 100$ indicates that posterior odds should be 100 times more favorable to the alternative than the prior odds.

The distinction between prior odds, posterior odds and Bayes factors provides an ideal mechanism for adding value to findings. Researchers should report the Bayes factor, and readers can update their own priors accordingly (Good, 1979; Jeffreys, 1961). Sophisticated researchers may add guidance and value to their analysis by suggesting prior odds, or ranges of prior odds. We use prior odds to add context to our Bayes factor analysis of Bem's (2011) claim of extrasensory perception of future events that cannot otherwise be known (Rouder & Morey, 2011). Our Bayes factor analysis of Bem's data yielded a Bayes factor of 40 in favor of an effect consistent with ESP. We cautioned readers, however, to hold substantially unfavorable prior odds toward ESP as there is no proposed mechanism, and its existence runs contrary to well-established principles in physics and biology. We believe that a Bayes factor of 40 is too small to sway readers who hold appropriately skeptical prior odds. Of course, a Bayes factor of 40 may be more consequential in less controversial domains where prior odds are less extreme.

Because Bayes factors measure the evidence for competing positions, they have been recommended for inference in psychological settings (an incomplete list includes Edwards et al., 1963; Gallistel, 2009; Lee & Wagenmakers, 2005; Mulder, Klugkist, van de Schoot, Meeus, & Hoijtink, 2009; Rouder et al., 2009; Vanpaemel, 2010; Wagenmakers, 2007). There are, however, other Bayesian approaches to inference including Aitkin's (1991, see Liu and Aitkin, 2008) posterior Bayes factors, Kruschke's (2011) use of posterior distributions on contrasts, and Gelman and colleagues' notion of model checking through predictive posterior p-values (e.g., Gelman, Carlin, Stern, & Rubin, 2004). The advantages and disadvantages of these methods remain an active and controversial topic in the statistical and social-science methodological literature. Covering this literature is outside the scope of this paper, and the interested reader is referred elsewhere: good reviews include Aitkin (1991, especially the subsequent comments), Berger and Sellke (1987, especially the subsequent comments), Raftery (1995), and, more recently, Gelman and Shalizi (in press). Our view is that none of these alternative approaches offers the ability to state evidence for invariances and effects in as convincing and as clear a manner as does Bayes factors. Additional discussion is provided in the conclusion as well as in Morey, Romeign, and Rouder (in press).

4. One-sample designs

4.1. Model and priors

In this section, we develop default priors for a one-sample design as an intermediate step toward developing Bayes factors for ANOVA designs. The development in this section will be directly relevant throughout. In a one-sample design, there is a single population, and the researcher's question of interest is whether the mean of that population is zero. An example of a one-sample design is a pretest-intervention-posttest design (Campbell & Stanley, 1963) in which the researcher tracks each individual's change between the pretest and posttest. The question of whether the mean intervention effect is zero is typically assessed via consideration of a p-value from a paired-sample t-test. The observed intervention effects are modeled as independent and identically distributed random variables:

$$y_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2), \quad i = 1, \ldots, N.$$

The null model, that there is no treatment effect, is given by $\mu = 0$. To compute a Bayes factor, we must also choose a prior distribution for μ under the alternative. It may seem desirable to make μ arbitrarily diffuse to approximate a state of minimal prior knowledge. This choice, however, is unwise. Diffuse priors imply that all values are equally plausible, including those that are obviously implausible. For instance, under a diffuse prior, an effect of 5% is as plausible as an effect of one million percent. When the likelihood under the alternative is averaged over large, implausible values, the average approaches zero. Hence, arbitrarily diffuse priors lead to the result that the null is more probable than the alternative regardless of the data (Lindley, 1957).

Jeffreys (1961) recommends reparameterizing the problem in terms of effect size, which is denoted by δ, where $\delta = \mu/\sigma$ is a dimensionless quantity. The model may then be rewritten:

$$y_i \sim \text{Normal}(\sigma\delta, \sigma^2).$$

Null and alternative models differ in the choice of priors on δ:

$$\mathcal{M}_0 : \delta = 0,$$

$$\mathcal{M}_1 : \delta \sim \text{Cauchy},$$

where the Cauchy is a distribution with probability density function

$$\pi(x) = \frac{1}{(1+x^2)\,\pi}, \tag{1}$$

and π in the denominator on the right-hand side is the common mathematical constant. Additional details about the Cauchy distribution are provided in Johnson, Kotz, and Balakrishnan (1994).

Priors must be specified for the remaining parameter in the model, σ^2. Fortunately, because this parameter plays an analogous role in both $\mathcal{M}_0$ and $\mathcal{M}_1$, it is possible and desirable to place a noninformative Jeffreys prior on σ^2:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

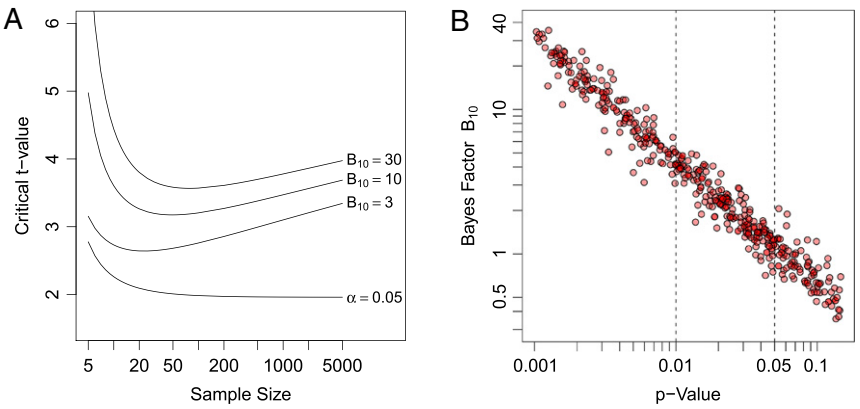

Fig. 2. A. Needed t-values for stating evidence for an effect as a function of sample size. The lower line shows the needed t-values for p-value of 0.05. The upper lines are the t-values corresponding to $B_{10} = 3, 10, 30$. B. Bayes factor evidence as a function of p-value for 855 t-tests reported in 2007.
Source: Adapted from Wetzels et al. (2011).

Bayarri and Garcia-Donato (2007) call this combination of priors the *JZS priors* in recognition of the contributions of Jeffreys (1961) as well as Zellner and Siow (1980), who generalized these priors for linear models. The resulting Bayes factor, called the *JZS Bayes factor*, is

$$B_{01}(t, N) = \frac{\left(1 + \frac{t^2}{N-1}\right)^{-N/2}}{\int_0^\infty (1+Ng)^{-1/2}\left(1 + \frac{t^2}{(1+Ng)(N-1)}\right)^{-N/2}\pi(g)dg}, \quad (2)$$

where $\pi(g)$ is the probability density function of the inverse χ^2 distribution with one degree of freedom:

$$\pi(g) = (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)}. \quad (3)$$

The expression is convenient because the data enter only though the test statistic $t = \bar{y}\sqrt{N}/s_y$, where $\bar{y}$ and s_y are the sample mean and sample standard deviation of the data, respectively. Fortunately, the expression is computationally convenient as the integration is across a single dimension and may be performed quickly and accurately using Gaussian quadrature (Press, Teukolsky, Vetterling, & Flannery, 1992). Rouder et al. (2009) provide a web applet for computing the JZS Bayes factor at http://pcl.missouri.edu/bayesfactor.

4.2. Properties of the Bayes factor

Some of the characteristic differences between inference by Bayes factor and p-values are shown in Fig. 2. Fig. 2(A) shows the needed t-value for stating particular levels of evidence for an effect. Consider the line for a Bayes factor of $B_{10} = 3$, which indicates that the data are three times more likely under the alternative than under the null. First, note that larger t-values are needed to maintain a $B_{01} = 3$ than are needed to maintain a $p = 0.05$ criterion. Second, note that as the sample size becomes large, increasingly larger t-values are needed to maintain the same level of evidence. The need for increasing t-values contrasts with inference by p-values. Fig. 2(B) shows the practical consequences of these different characteristics. The figure summarizes the findings of Wetzels et al. (2011), who provided p-values and JZS Bayes factors for all 855 t-tests reported in the *Journal of Experimental Psychology: Learning, memory, and Cognition* and *Psychonomic Bulletin and Review* in 2007. We have plotted the results for the 440 tests that have p-values between 0.001 and 0.15. The plot shows that although Bayes factors and p-values rely on the same information in the data, they are calibrated differently. In particular, the tendency of p-values to overstate the evidence in data against the null hypothesis is apparent. For example, a p-value of 0.05 may correspond to as much evidence for the alternative as for the null, and even a p-value of 0.005 hardly confers a strong advantage for the alternative.

4.3. Desirable theoretical properties of default priors

Our goal in this paper is to develop default priors that may be used broadly and easily. One criteria for choosing these priors is to consider the theoretical properties of the resulting Bayes factors. The one-sample Bayes factor in Eq. (2) has the following desirable properties:

- *Scale invariance*. The value of the Bayes factor is unaffected by multiplicative changes in the unit of measure of the observations. For instance, if observations are in a unit of length, the Bayes factor is the same whether the measurement is in nanometers or light-years. This invariance comes about because of the scale-invariant nature of the prior on σ^2 and the placing of a prior on effect size rather than on mean (Jeffreys, 1961).
- *Consistency*. In the large sample limit, the Bayes factor approaches the appropriate bound (Liang, Paulo, Molina, Clyde, & Berger, 2008):

  $$\delta = 0 \Longrightarrow \lim_{N\to\infty} B_{10}(t(N), N) = 0,$$
  $$\delta \neq 0 \Longrightarrow \lim_{N\to\infty} B_{10}(t(N), N) = \infty,$$

  where $t(N) = \bar{y}\sqrt{N}/s_y$ is the t-statistic.
- *Consistent in information*. The Bayes factor approaches the correct limit as t increases, e.g., $\lim_{t\to\infty} B_{10}(t, N) = \infty$ for all N. This last property is called consistency in information, and it holds for the Cauchy prior on effect size, but not for a normal prior on effect size (Jeffreys, 1961; Zellner & Siow, 1980). The property holds when the prior has slowly-diminishing tails, and serves as additional motivation for the Cauchy prior on δ.

5. Multivariate generalizations of the Cauchy

The focus of this paper is the development of default-prior Bayes factor for ANOVA settings. In the previous development, there was a single effect parameter, δ, on which the prior is a Cauchy distribution. In ANOVA and regression designs, we will posit several effect parameters, and a suitable prior for each. There are two possible extensions of the Cauchy, and the contrast between them is informative. The first is a straightforward

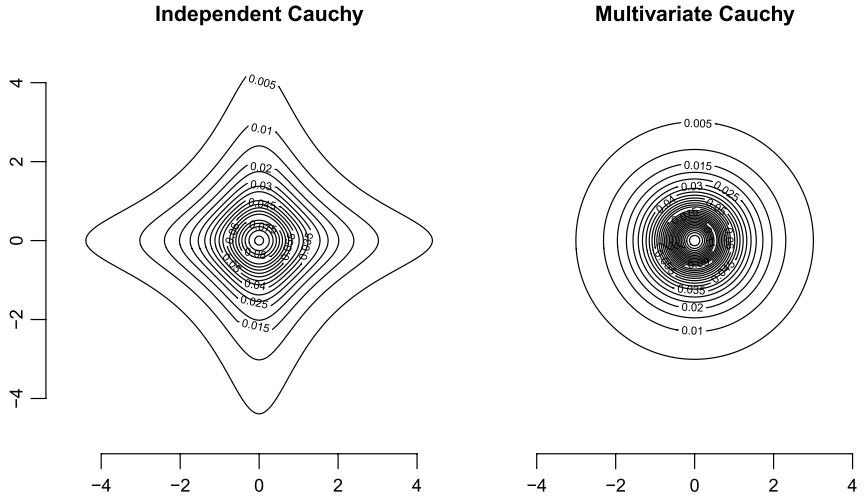

Fig. 3. Two bivariate Cauchy distributions. For both distributions, the marginal distributions are univariate Cauchy. In the independent Cauchy distribution (left), the joint is the product of the marginals. In the multivariate Cauchy distribution (right), there is a dependence with more joint density on effects that are similar in magnitude.

independent Cauchy prior in which the multivariate prior density on p effects is simply the product of p univariate prior densities. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ be a vector of p effects. The independent Cauchy has a density function,

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \frac{1}{(1+\theta_i^2)\pi}. \tag{4}$$

A plot of a bivariate independent Cauchy prior is shown on the left side of Fig. 3, and it is characterized by a lasso shape. The lack of symmetry, in which there is sizable mass for large values of one effect and small values of the other, is a natural consequence of the fat tails of the Cauchy.

The second generalization, conventionally termed the *multivariate Cauchy* (Kotz & Nadarajah, 2004), is given by the joint probability density function

$$\pi(\boldsymbol{\theta}) = \frac{\Gamma\left[(1+p)/2\right]}{\Gamma(1/2)\pi^{p/2}\left[1+\sum_{i=1}^{p}\theta_i^2\right]^{(1+p)/2}}. \tag{5}$$

The marginal distribution for any one of the p dimensions is a univariate Cauchy with density given in (1). A plot of the bivariate case is shown on the right side of Fig. 3, and the defining characteristic of this generalization is a specified dependence among the effects such that they are similarly sized in magnitude. When compared to the independent Cauchy, the multivariate Cauchy places less mass on the combinations of large values for one effect and small values on the other. The result of this dependence is a symmetric bivariate distribution, but this symmetry should not be confused for independence.

The motivation for the multivariate Cauchy comes from the relationship between the normal and Cauchy distributions. The Cauchy results from a mixture of normals with different variances. Consider the following conditional model on an effect θ,

$$\theta|g \sim \text{Normal}(0, g)$$

where g is the variance. Zellner and Siow (1980) note that if g follows an inverse-χ^2 distribution with one degree of freedom, then the marginal distribution of θ is a univariate Cauchy. For the multivariate case, let $\boldsymbol{\theta}$ be a vector of p effects,

$$\boldsymbol{\theta}|g \sim \text{Normal}(\mathbf{0}, g\boldsymbol{I}_p),$$

where $\boldsymbol{I}_p$ is the identity matrix of size p. If g follows an inverse-χ^2 distribution with one degree of freedom, then the marginal distribution of $\boldsymbol{\theta}$ is the multivariate Cauchy given in (5). The independent Cauchy may also be expressed as a mixture of normals. Let $\boldsymbol{G}$ be a $p \times p$ diagonal matrix with values $g_1, g_2, \ldots, g_p$ on the diagonal. Let

$$\boldsymbol{\theta}|\boldsymbol{G} \sim \text{Normal}(\mathbf{0}, \boldsymbol{G}),$$

and let each g_i be distributed independently as an inverse-χ^2 with a single degree of freedom. Then the marginal distribution of $\boldsymbol{\theta}$ is the independent Cauchy in (4). Both the independent and multivariate Cauchy generalizations will be useful, and, as is discussed, each is used to encode different sets of relations among effects.

6. Bayes factor for ANOVA models

In this section, we provide general development for ANOVA models. It is convenient to use matrix notation. Let $\boldsymbol{y}$ be a vector of N observations. It is convenient to start with a linear model with p effects:

$$\boldsymbol{y} = \mu\mathbf{1} + \sigma\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{6}$$

where μ is a grand mean parameter, $\mathbf{1}$ is a column vector of length N with entries of 1, $\boldsymbol{\theta}$ is a column vector of standardized effect parameters of length p, and $\boldsymbol{X}$ is a $N \times p$ design matrix. The vector $\boldsymbol{\epsilon}$ containing the error terms is a column vector of length N:

$$\boldsymbol{\epsilon}|\sigma^2 \sim \text{Normal}(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

Note that the parameterization of the linear model in (6) differs from more conventional presentations (e.g., McCullagh & Nelder, 1989). In (6) the effects are standardized relative to the standard deviation of the error, and, consequently, $\sigma\boldsymbol{X}\boldsymbol{\theta}$ explicitly includes a scale factor σ.

ANOVA models have two constraints. First, the covariates are categorical and indicate group membership. Membership may be indicated by setting design-matrix entries to be 1 or 0, denoting whether an observation is from a specific level of a specific factor or not, respectively. Second, factors provide a natural hierarchy or grouping (Gelman, 2005). It is reasonable to think *a priori* that levels within a factor are exchangeable whereas levels across factors are not. We will implement this notion of exchangeability in our development.

Priors are needed on parameters μ, σ^2, and the vector of standardized effects $\boldsymbol{\theta}$. As previously, we place a Jeffreys prior on μ and σ^2:

$$\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}.$$

For ANOVA models with categorical covariates, we assume the following g-prior structure:

$$\boldsymbol{\theta}|\boldsymbol{G} \sim \text{Normal}(\mathbf{0}, \boldsymbol{G}), \tag{7}$$

where $\boldsymbol{G}$ is a $p \times p$ diagonal matrix. A different prior, discussed subsequently, is used when the covariate is continuous rather than categorical.

To complete the specification of the prior, the analyst needs to choose the diagonal of $\boldsymbol{G}$. One possible choice of priors is to use a separate g parameter for each element of $\boldsymbol{\theta}$. In this case, the diagonal of $\boldsymbol{G}$ consists of $g_1, \ldots, g_p$. The priors on these parameters are

$$g_i \overset{\text{i.i.d.}}{\sim} \text{Inverse-}\chi^2(1), \quad i = 1, \ldots, p.$$

The corresponding marginal prior on $\boldsymbol{\theta}$ is the independent Cauchy distribution. The independent Cauchy prior is useful when there is no *a priori* relationship among effects. Yet, in some cases, it is more appropriate to assume that effects vary on a similar scale, and are not arbitrarily different from one another. In this case, the multivariate Cauchy may be more appropriate. The multivariate Cauchy prior is implemented by setting $\boldsymbol{G} = g\boldsymbol{I}$, and $g \sim \text{Inverse-}\chi^2(1)$. The development of Bayes factors for this single-g model is discussed in Bayarri and Garcia-Donato (2007).

Gelman (2005) comments that ANOVA should be viewed as a hierarchical grouping of effects into factors where levels within but not across factors are *exchangeable*. When effects share a common g parameter, they are indeed exchangeable in that they are random deviates from a common parent distribution in a hierarchical structure. Hence, effects within a factor should share a common g parameter while those across should not. For example, suppose there are four effects, $\theta_1, \ldots, \theta_4$ with θ_1 and θ_2 describing the effect of one factor and θ_3 and θ_4 describing the effect of another. Because the first two levels are exchangeable within one factor and the second in a different factor, we may specify that the scales of the first two effects may be more similar to each other, but may be dissimilar to those for the last two effects. In this case, a separate g-parameter for each factor is appropriate, e.g.,

$$\boldsymbol{G} = \begin{pmatrix} g_1 & 0 & 0 & 0 \\ 0 & g_1 & 0 & 0 \\ 0 & 0 & g_2 & 0 \\ 0 & 0 & 0 & g_2 \end{pmatrix}.$$

In this case, the priors on g_1 and g_2 would be independent inverse chi-square with one degree-of-freedom. The marginal prior on $\boldsymbol{\theta}$ in this case is two multivariate Cauchy priors, where each is a bivariate distribution across two levels of a factor. These two multivariate Cauchy distributions are independent of one another. We develop Bayes factors for any combination of independent and multivariate Cauchy distributions. Let r denote the number of unique g parameters in $\boldsymbol{G}$, and let $\boldsymbol{g} = (g_1, \ldots, g_r),\ 1 \le r \le p$.

The marginal likelihood, m, for the ANOVA model is obtained by integrating the likelihood against the joint prior for μ, σ^2, $\boldsymbol{\theta}$, and $\boldsymbol{g}$. It is not possible to express this integral across all parameters as a closed-form expression. Fortunately, it is possible to derive a closed-form expression for the integral across μ, σ^2, and $\boldsymbol{\theta}$,

$$m = \int_{g_1} \cdots \int_{g_r} T_m(\boldsymbol{g})\pi(g_1)\cdots\pi(g_r)\, dg_1 \cdots dg_r, \tag{8}$$

where $T_m(\boldsymbol{g})$ is the likelihood integrated with respect to the joint priors on μ, σ^2 and $\boldsymbol{\theta}$, and where $\pi(g)$ is the probability density function of an inverse-χ^2 distribution with one degree of freedom given in (3). To define $T_m(\boldsymbol{g})$, let

$$\begin{aligned} \boldsymbol{P}_0 &= \frac{1}{N}\mathbf{1}\mathbf{1}', \\ \tilde{\boldsymbol{y}} &= (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{y}, \\ \tilde{\boldsymbol{X}} &= (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{X}, \\ \boldsymbol{V}_{\boldsymbol{g}} &= \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{G}^{-1}. \end{aligned}$$

Then the integrated likelihood is

$$T_m(\boldsymbol{g}) = \frac{\Gamma((N-1)/2)}{\pi^{(N-1)/2}|\boldsymbol{G}|^{1/2}|\boldsymbol{V}_{\boldsymbol{g}}|^{1/2}\sqrt{N}(\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{g}}^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}})^{(N-1)/2}}.$$

The derivation of $T_m(\boldsymbol{g})$ is provided in the Appendix.

Bayes factors for the model in (6) may be constructed with reference to the null model, $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{\epsilon}$.

Using the same argument as in the Appendix, the corresponding marginal likelihood, denoted m_0, is

$$m_0 = \frac{\Gamma((N-1)/2)}{\pi^{(N-1)/2}\sqrt{N}(\boldsymbol{y}'\boldsymbol{y} - N\bar{\boldsymbol{y}}^2)^{(N-1)/2}},$$

where $\bar{\boldsymbol{y}} = \mathbf{1}'\boldsymbol{y}/N$. The Bayes factor between the model in (6) and the null model is

$$B_{10} = \int_{g_1} \cdots \int_{g_r} S(\boldsymbol{g})\pi(g_1)\cdots\pi(g_r)\, dg_1 \cdots dg_r \tag{9}$$

where

$$S(\boldsymbol{g}) = \frac{1}{|\boldsymbol{G}|^{1/2}|\boldsymbol{V}_{\boldsymbol{g}}|^{1/2}}\left(\frac{\boldsymbol{y}'\boldsymbol{y} - N\bar{\boldsymbol{y}}^2}{\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{g}}^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}}\right)^{(N-1)/2}.$$

Eq. (9) is used throughout for computing Bayes factors. Appropriate choices for $\boldsymbol{G}$ and $\boldsymbol{X}$ in various ANOVA designs are discussed in the following sections. Computational issues in evaluating (9) are discussed in Section 14.

The proposed default prior is similar to those proposed by Zellner and Siow (1980) and recommended for ANOVA by Wetzels, Grasman, and Wagenmakers (2012). Yet, there are two critical differences. The Zellner–Siow prior is based on a single g parameter whereas our prior is more flexible and allows for different g parameters across different factors. A second critical difference is that the Zellner–Siow prior on effect sizes has an additional scaling term: $\boldsymbol{\theta}|g \sim \text{Normal}(\mathbf{0}, g(\boldsymbol{X}'\boldsymbol{X}/N)^{-1}\boldsymbol{I}_p)$, where $(\boldsymbol{X}'\boldsymbol{X}/N)^{-1}$ is this new term. In Section 13 we discuss the meaning of this additional term, and argue that such scaling is appropriate for continuous covariates (regression) but inappropriate for categorical covariates (ANOVA).

7. One-way ANOVA designs

In this section, we develop the default Bayes factor for the case where observations are classified into one of a groups. Let $\boldsymbol{\alpha}$ be a vector of a effects, $\alpha = (\alpha_1, \ldots, \alpha_a)'$. The corresponding model is

$$\boldsymbol{y} = \mu\mathbf{1} + \sigma\boldsymbol{X}_\alpha\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{10}$$

The design matrix, denoted $\boldsymbol{X}_\alpha$, has N rows and a columns, and is populated by entries of one or zero that indicate group membership. For instance, if 7 observations came from 3 groups, with the first two observations in the first group, the next two observations in the second group, and the last three observations in the third group, the design matrix would be

$$\boldsymbol{X}_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The model in (10) is not identifiable without additional constraint as there are a total of $a + 1$ parameters that determine the a cell means. In classical statistics, the additional constraint reflects whether effects are treated as *fixed* or *random*. For fixed effects, additional linear constraints are imposed, e.g., $\sum_i \alpha_i = 0$.

For random effects, the constraint comes from considering each effect as a sample from a common distribution, or, as discussed previously, as exchangeable. Gelman (2005) recommends this hierarchical approach for both fixed and random effects, and we follow this recommendation here. Gelman also recommends that analysts impose the usual sum-to-zero linear constraints as well, and the difference between fixed and random effects is a matter of interpretation but not computation. We do not take this last recommendation. Instead, we make a sharp distinction between treating factors as fixed and random. When factors are treated as fixed, the usual sum-to-zero constraints are imposed. When they are treated as random, these constraints are not imposed. As a rule of thumb, it is appropriate to treat a factor as fixed when they are manipulated through a few levels, and the focus is on the difference between levels. Likewise, it is appropriate to treat a factor as random when levels are sampled, such as the sampling of participants from a participant pool or the sampling of words from a language, and the focus is on generalization to all possible levels of the factor. We consider the random effects model first as it is more straightforward.

7.1. Random effects model

A natural specification for the random effects one-way ANOVA model is

$$\boldsymbol{\alpha} \mid g \sim \text{Normal}(\mathbf{0}, g\boldsymbol{I}),$$

where g is the variance of the random effects. The prior on g is $g \sim \text{Inverse-}\chi^2(1)$, and the resulting marginal prior on $\boldsymbol{\alpha}$ is the multivariate Cauchy in (5).

The marginal likelihood of this random-effects model is given in (8) by setting $\boldsymbol{X} = \boldsymbol{X}_\alpha$ and $\boldsymbol{G} = g\boldsymbol{I}$. The Bayes factor in (9) may be expressed as follows. Let y_{ij} be the jth observation in the ith group, $i = 1, \ldots, a, j = 1, \ldots, n_j$; let $\bar{y}_{i\cdot}$ be the sample mean for the ith group; and let $\bar{y}_{\cdot\cdot}$ be the grand sample mean. Box I, Eq. (11) provides the Bayes factor between the model in (10) and the null given by $\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{\epsilon}$.

If the design is balanced, then (11) reduces to

$$B_{10} = \int_g (1+gn)^{-(a-1)/2} \times \left(1 - \frac{R^2}{(1+gn)/gn}\right)^{-(N-1)/2} \pi(g)\, dg, \tag{12}$$

where R^2 is the unadjusted proportion of variance accounted for by the model[3] and $n = n_1 = \cdots = n_a$. The one-dimensional integral in (11) and (12) may be conveniently and accurately evaluated with Gaussian quadrature.

7.2. Fixed effects models

In one-way ANOVA, the fixed effect constraint is $\sum_i \alpha_i = 0$. One approach is to consider only the first $a-1$ effects and set the last one to $\alpha_a = -\sum_{i=1}^{a-1} \alpha_i$. A drawback of this approach, however, is that the choice of eliminated effect is arbitrary. Moreover, the marginal prior on the eliminated effect cell mean is more diffuse than on the others.

A better approach to implementing the sum-to-zero constraint is to project the space of a dimensions into a space of dimension $a-1$ with the property that the marginal prior on all a effects is identical. The constraint that $\sum \alpha_i = 0$ may be implemented by placing a prior with negative correlation across the effects. A suitable choice for the covariance matrix across the effects is

$$\boldsymbol{\Sigma}_a = \boldsymbol{I}_a - \boldsymbol{J}_a/a$$

where $\boldsymbol{I}_a$ is the identity matrix (of size a) and $\boldsymbol{J}_a$ is a square matrix of size a with entries 1.0. For example, if $a = 3$, the resulting covariance matrix is

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}.$$

The above covariance matrix is not full rank, as it captures the side condition on $\boldsymbol{\alpha}$. Consequently, $\boldsymbol{\Sigma}_a$ may be decomposed as

$$\boldsymbol{\Sigma}_a = \boldsymbol{Q}_a \boldsymbol{I}_{a-1} \boldsymbol{Q}_a'$$

where $\boldsymbol{Q}_a$ is an $a \times (a-1)$ matrix of the $a-1$ eigenvectors of unit length corresponding to the nonzero eigenvalues of $\boldsymbol{\Sigma}_a$, and $\boldsymbol{I}_{a-1}$ is an identity matrix of size $a-1$. The new parameter vector of $a-1$ effects, $\boldsymbol{\alpha}^*$, is defined by

$$\boldsymbol{\alpha}^* = \boldsymbol{Q}_a' \boldsymbol{\alpha}.$$

Inspection of these matrices is helpful in understanding the nature of parameter constraint. For two groups,

$$\boldsymbol{Q}_2' = \left(\sqrt{2}/2, \quad -\sqrt{2}/2\right).$$

For five groups,

$$\boldsymbol{Q}_5' = \begin{pmatrix} 0.89 & -0.22 & -0.22 & -0.22 & -0.22 \\ 0 & 0.87 & -0.29 & -0.29 & -0.29 \\ 0 & 0 & 0.82 & -0.41 & -0.41 \\ 0 & 0 & 0 & 0.71 & -0.71 \end{pmatrix}.$$

Note that $\boldsymbol{Q}_a$ defines an orthonormal set of contrasts that identify the $a-1$ parameters.

Let $\boldsymbol{X}_\alpha^*$ denote the $N \times (a-1)$ design matrix that maps $\boldsymbol{\alpha}^*$ into observations:

$$\boldsymbol{X}_\alpha^* = \boldsymbol{X}_\alpha \boldsymbol{Q}_a. \tag{13}$$

With this full-rank parameterization, the fixed-effect model is

$$\boldsymbol{y} = \mu\mathbf{1} + \sigma \boldsymbol{X}_\alpha^* \boldsymbol{\alpha}^* + \boldsymbol{\epsilon}. \tag{14}$$

A prior is needed on $\boldsymbol{\alpha}^*$, and we use a multivariate Cauchy:

$$\boldsymbol{\alpha}^* | g \sim \text{Normal}(\mathbf{0}_{a-1}, g\boldsymbol{I}_{a-1}), \qquad g \sim \text{Inverse-}\chi^2(1)$$

where the $\mathbf{0}$ column vector is of length $a-1$. This prior maintains a notion of exchangeability, though the exchangeability is on the differences between effects rather than the effects themselves.

The Bayes factor is calculated from (9) by setting $\boldsymbol{X} = \boldsymbol{X}_\alpha^*$ and setting $\boldsymbol{G} = g\boldsymbol{I}_{a-1}$. This Bayes factor will, in general, be different from the random-effects Bayes factor in (11). If the design is balanced, however, it can be shown that the Bayes factor reduces to the same expression as that for the random-effects in (12). This equivalence of random-effect and fixed-effect Bayes factors in balanced one-way designs is analogous to the equivalence of F-tests for one-way, balanced designs. Whereas most researchers use balanced designs, consideration of fixed or random effects is not critical in this case. There are, however, important differences for multiple factor designs.

In ANOVA designs, researchers are sometimes concerned about additional contrasts, such as whether any two levels differ. For instance suppose a factor has three levels and the main-effect Bayes factor indicates that the full model is preferred to the null model. Then, three intermediate models may be proposed where any two levels equal each other. Each of these models can be implemented with a simple two-column design matrix and tested with the above methodology. The resulting pattern of Bayes factors across these models, as well as that across the full model, may be compared in analysis.

[3] The R^2 statistic is

$$R^2 = \frac{\sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2}.$$

$$B_{10} = \int_g K(\boldsymbol{n}, g) \left(\frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2 + \frac{1}{g}\left(\sum_i c_i \bar{y}_i^2 - \left(\sum_i c_i \bar{y}_i\right)^2 \Big/ \left(\sum_i c_i\right)\right)}{\sum_i \sum_j \left(y_{ij} - \bar{y}\right)^2} \right)^{-(N-1)/2} \pi(g)\, dg \tag{11}$$

where $\boldsymbol{n} = (n_1, \ldots, n_a)'$,

$N = \sum_i n_i$,

$c_i = \dfrac{n_i}{n_i + 1/g}$,

and $K(\boldsymbol{n}, g) = \sqrt{N} \left(\dfrac{\prod_i 1/(1+gn_i)}{\sum_i n_i/(1+gn_i)} \right)^{1/2}$.

Box I.

8. Multi-way ANOVA

In many applications, researchers employ factorial designs in which they seek to assess main effects and interactions. In this section, we develop the Bayes factor for multiple factors. Although the following developments generalize seamlessly to any number of factors, we will focus on the two-factor case for concreteness. Let a and b denote the number of levels for the first and second factors, respectively. Let $\boldsymbol{\alpha}$ be a vector of a standardized effects for the first factor, let $\boldsymbol{\beta}$ be a vector of b standardized effects for the second factor, and let $\boldsymbol{\gamma}$ be a vector of $a \times b$ standardized interaction effects. A full model may be given by

$$\mathcal{M}_f : \boldsymbol{y} = \mu \mathbf{1} + \sigma \left(\boldsymbol{X}_\alpha \boldsymbol{\alpha} + \boldsymbol{X}_\beta \boldsymbol{\beta} + \boldsymbol{X}_\gamma \boldsymbol{\gamma} \right) + \boldsymbol{\epsilon}. \tag{15}$$

Design matrices $\boldsymbol{X}_\alpha, \boldsymbol{X}_\beta$ and $\boldsymbol{X}_\gamma$ describe how effect parameters map onto observations. For example, if $a = 2, b = 2$, and there is one replicate per cell, the design matrices are

$$\boldsymbol{X}_\alpha = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{X}_\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{X}_\gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For balanced designs with n replicates per cell, these design matrices are given compactly by

$$\boldsymbol{X}_\alpha = \boldsymbol{I}_a \otimes \mathbf{1}_{b \times n}, \qquad \boldsymbol{X}_\beta = \mathbf{1}_a \otimes \boldsymbol{I}_b \otimes \mathbf{1}_n, \qquad \boldsymbol{X}_\gamma = \boldsymbol{I}_{a \times b} \otimes \mathbf{1}_n,$$

where subscripts on $\mathbf{1}$ and $\boldsymbol{I}$ denote the sizes, and $\otimes$ denotes a Kronecker product (Eves, 1980).

In factorial designs, researchers are interested in an array of models that encode constraints on main effects and interactions. In addition to the full model, $\mathcal{M}_f$, there are seven submodels of the full model for the two-way design:

$$\mathcal{M}_{\alpha+\beta} : \boldsymbol{y} = \mu \mathbf{1} + \sigma \left(\boldsymbol{X}_\alpha \boldsymbol{\alpha} + \boldsymbol{X}_\beta \boldsymbol{\beta} \right) + \boldsymbol{\epsilon}.$$

$$\mathcal{M}_{\alpha+\gamma} : \boldsymbol{y} = \mu \mathbf{1} + \sigma \left(\boldsymbol{X}_\alpha \boldsymbol{\alpha} + \boldsymbol{X}_\gamma \boldsymbol{\gamma} \right) + \boldsymbol{\epsilon}.$$

$$\mathcal{M}_{\beta+\gamma} : \boldsymbol{y} = \mu \mathbf{1} + \sigma \left(\boldsymbol{X}_\beta \boldsymbol{\beta} + \boldsymbol{X}_\gamma \boldsymbol{\gamma} \right) + \boldsymbol{\epsilon}.$$

$$\mathcal{M}_\alpha : \boldsymbol{y} = \mu \mathbf{1} + \sigma \boldsymbol{X}_\alpha \boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

$$\mathcal{M}_\beta : \boldsymbol{y} = \mu \mathbf{1} + \sigma \boldsymbol{X}_\beta \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$$\mathcal{M}_\gamma : \boldsymbol{y} = \mu \mathbf{1} + \sigma \boldsymbol{X}_\gamma \boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

as well as the null model,

$$\mathcal{M}_0 : \boldsymbol{y} = \mu \mathbf{1} + \boldsymbol{\epsilon}.$$

8.1. Fixed, random, and mixed effects

Different models of effects may be implemented through the design matrices, as we discuss in the following sections.

8.1.1. Random effects

Consider first the case in which both factors are treated as random effects, and consequently, the interaction terms are random effects as well. We recommend the following prior structure with three separate g parameters for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$:

$$\begin{aligned} \boldsymbol{\alpha} \mid g_\alpha &\sim \text{Normal}(\mathbf{0}, g_\alpha \boldsymbol{I}_a), \\ \boldsymbol{\beta} \mid g_\beta &\sim \text{Normal}(\mathbf{0}, g_\beta \boldsymbol{I}_b), \\ \boldsymbol{\gamma} \mid g_\gamma &\sim \text{Normal}(\mathbf{0}, g_\gamma \boldsymbol{I}_{a \times b}), \end{aligned} \tag{16}$$

with $g_k \overset{\text{i.i.d.}}{\sim} \text{Inverse-}\chi^2(1)$ for $k = \alpha, \beta, \gamma$. Note here that the prior on standardized effects is the product of three independent, multivariate Cauchy distributions. Within a factor, the levels are related through a common g parameter. Yet, there are separate g parameters across factors (and their interactions), and this indicates that the factors themselves are unrelated. The Bayes factor for the full model relative to the null model, denoted $B_{f,0}$, is given in (9) with $\boldsymbol{X} = (\boldsymbol{X}_\alpha, \boldsymbol{X}_\beta, \boldsymbol{X}_\gamma)$ and $\boldsymbol{G} = \text{diag}(g_\alpha \mathbf{1}_a', g_\beta \mathbf{1}_b', g_\gamma \mathbf{1}_{ab}')$. Computational approaches to performing the resulting three-dimensional integral are discussed in Section 14. Bayes factors for the submodels are given analogously.

8.1.2. Fixed effect models

Consider the case where both factors are treated as fixed effects, and, consequently, the interaction is fixed as well. The usual side conditions on fixed effects are

$$\begin{aligned} \sum_i \alpha_i &= 0, \\ \sum_j \beta_j &= 0, \\ \sum_i \gamma_{ij} &= 0, \\ \sum_j \gamma_{ij} &= 0. \end{aligned} \tag{17}$$

The side conditions on main effects each impose one linear constraint; the side condition on interactions imposes $I + J - 1$ linear constraints.

To capture these side conditions, it is helpful to specify a matrix operation for the construction of interaction design matrices from

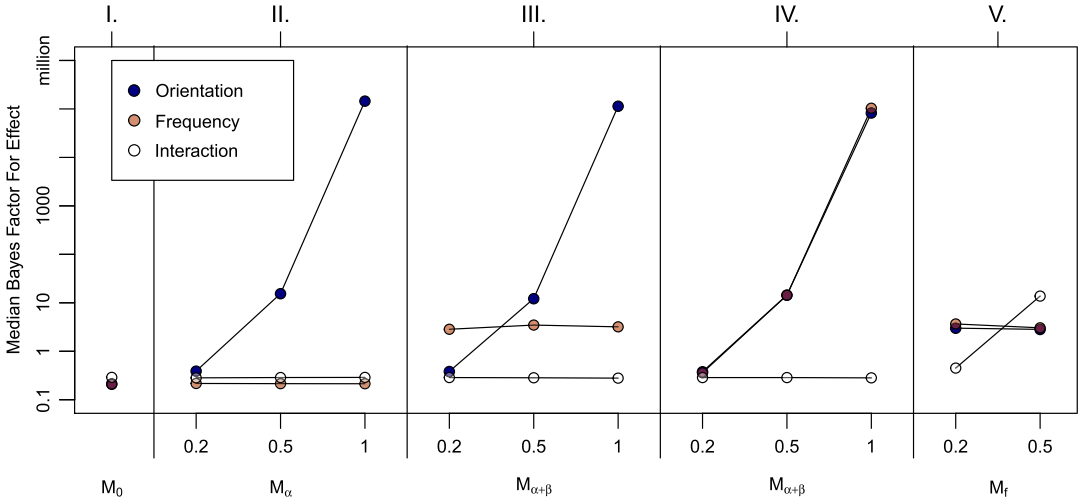

Fig. 4. Median Bayes factor from simulated data. I. Data generated from the null model. II. Data generated with main effects in orientation. True effect-size values for orientation were 0.2, 0.5, and 1. III. Same as previous simulation, except there was a true main effect of frequency as well (true orientation effect-size values of 0.2, 0.5, and 1; true frequency effect-size value of 0.4). IV. Data generated with equal-sized true main effects in orientation and frequency. V. Data generated with main effects of both factors (true effect-size values of 0.4) and an interaction (true effect size values of 0.2 and 0.5). Orientation and frequency are modeled as fixed effects.

main effect ones. Box II shows the definition of this matrix operator, denoted $\odot$. The design matrices of interactions in factorial designs are given by

$$\boldsymbol{X}_{\gamma} = \boldsymbol{X}_{\alpha} \odot \boldsymbol{X}_{\beta}.$$

The following full model captures the side conditions in (17):

$$\boldsymbol{y} = \mu\mathbf{1} + \sigma\left(\boldsymbol{X}^{*}_{\alpha}\boldsymbol{\alpha}^{*} + \boldsymbol{X}^{*}_{\beta}\boldsymbol{\beta}^{*} + \boldsymbol{X}^{**}_{\gamma}\boldsymbol{\gamma}^{**}\right) + \boldsymbol{\epsilon}.$$

Main-effects parameter vectors $\boldsymbol{\alpha}^{*}$ and $\boldsymbol{\beta}^{*}$ are of length $a-1$ and $b-1$, respectively, and the corresponding respective design matrices $\boldsymbol{X}^{*}_{\alpha}$ and $\boldsymbol{X}^{*}_{\beta}$ are derived from the centering projection analogously to (13). The interaction parameter vector $\boldsymbol{\gamma}^{**}$ is of length $(a-1)(b-1)$, and the corresponding design matrix is given by

$$\boldsymbol{X}^{**}_{\gamma} = \boldsymbol{X}^{*}_{\alpha} \odot \boldsymbol{X}^{*}_{\beta}.$$

The use of two asterisks in the superscript on interaction parameters and design matrices indicates that there are separate sum-to-zero constraints on both rows and columns in the matrix representation of interaction parameters. Prior specification of $\boldsymbol{\alpha}^{*}$, $\boldsymbol{\beta}^{*}$, and $\boldsymbol{\gamma}^{**}$ is analogous to (16). Moreover, all submodels are defined as the appropriate restriction on this full model.

8.1.3. Mixed interactions

The development extends in a straightforward manner to mixed interactions. For example, suppose the first factor is fixed and the second is random. The model is given by

$$\boldsymbol{y} = \mu\mathbf{1} + \sigma\left(\boldsymbol{X}^{*}_{\alpha}\boldsymbol{\alpha}^{*} + \boldsymbol{X}_{\beta}\boldsymbol{\beta} + \boldsymbol{X}^{*\cdot}_{\gamma}\boldsymbol{\gamma}^{*\cdot}\right) + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}^{*\cdot}_{\gamma} = \boldsymbol{X}^{*}_{\alpha} \odot \boldsymbol{X}_{\beta}$ is a design matrix with $(a-1)b$ columns and $\boldsymbol{\gamma}^{*\cdot}$ is an interaction vector of $(a-1)b$ effects which obeys the side constraint on row sums of interactions but not on column sums. This parameterization of mixed interactions is the same as in the classical Cornfield–Tukey mixed model (Cornfield & Tukey, 1956; Neter, Kutner, Wasserman, & Nachtschiem, 1996). Submodels are defined by various restrictions of this full model.

8.2. Assessment of main effects and interactions

Conventional ANOVA is a top-down approach in which the total variability is partitioned into main effects and interactions, and that which is residual. Each main effect and interaction is separately assessed through a comparison of the accounted variation relative to an appropriate error term. In the two-way case, researchers are interested in three comparisons: the two main-effect comparisons and the interaction. Here, we recommend several useful Bayes factor model comparisons.

Assessing interactions is the most straightforward, and a top-down approach that contrasts the performance of the full model to one without interactions is appropriate. We denote the corresponding Bayes factor by $B_{f,\alpha+\beta}$.[4] If the restriction without the target interaction is preferred to the full model with it, the interaction term is unnecessary to account for the data. Then, the appropriate Bayes factor to test the main effects of Factor 1 and Factor 2 are $B_{f,\beta+\gamma}$ and $B_{f,\alpha+\gamma}$, respectively, and the effect in question is preferred if the full model has higher marginal likelihood than the restriction without it. In some contexts, the analyst may be interested whether there is any effect of a factor rather than just a main effect. In this case, corresponding Bayes factor comparisons $B_{f,\beta}$ and $B_{f,\alpha}$ are appropriate for assessing Factor 1 and Factor 2, respectively.

We ran a small-scale set of simulations to assess the performance of these three Bayes factor contrasts. To make the situation concrete, we assumed that participants responded to the onset of Gabor patches that varied in orientation and frequency, modeled as fixed effects. There were 2 levels per factor and 10 replicates per cell in a simulated data set. We simulated data from 12 different true models, which comprised select combinations of main effects and interactions, and for each of these true models, 1000 simulated data sets were analyzed. Median Bayes factors across these 1000 sets for main effects and interactions are shown in Fig. 4. In Simulation I, far left panel, the null model serves as the generating model, and the Bayes factors for main effects and interaction correctly favor the null. In Simulation II, next panel, there is a main effect of orientation, that is, $\mathcal{M}_{\alpha}$ serves as the generating model. The three different effect size values[5] of orientation are shown (0.2, 0.5, and 1). Median Bayes factor for the main effect of orientation tracks with effect size, and the median Bayes factors for the interaction

[4] The Bayes factor may be computed by noting that $B_{f,\alpha+\beta} = B_{f,0}/B_{\alpha+\beta,0}$. Both $B_{f,0}$ and $B_{\alpha+\beta,0}$ are given in (9) with appropriate choices for $\boldsymbol{X}$ and $\boldsymbol{G}$.

[5] An effect size of 0.2 for a fixed factor with two levels means that the effect for both levels is 0.2 standardized units from the mean.

Let S and T be defined as

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \vdots & \vdots \\ s_{r1} & \cdots & s_{rm} \end{pmatrix}, \qquad \boldsymbol{T} = \begin{pmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \vdots & \vdots \\ t_{r1} & \cdots & t_{rn} \end{pmatrix},$$

The matrix operator $\odot$ is defined as

$$\boldsymbol{S} \odot \boldsymbol{T} = \begin{pmatrix} s_{11}t_{11} & s_{11}t_{12} & \cdots & s_{11}t_{1n} & s_{12}t_{11} & \cdots & s_{12}t_{1n} & \cdots & s_{1m}t_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{r1}t_{r1} & s_{r1}t_{r2} & \cdots & s_{r1}t_{rn} & s_{r2}t_{r1} & \cdots & s_{r2}t_{rn} & \cdots & s_{rm}t_{rn} \end{pmatrix}.$$

Box II.

	Vertical		Horizontal	
Freq.	Low	High	Low	High
	0.7	0.53	0.87	0.97
	0.71	0.52	0.72	0.9
	0.68	0.82	0.85	0.67
	0.57	0.78	0.82	0.66
	0.6	0.6	0.81	0.6
	0.68	0.67	0.88	0.61
	0.57	0.76	0.84	1.07
	0.67	0.65	0.94	0.89
	0.62	0.49	0.94	0.71
	0.75	0.63	0.68	0.6
Avg.	.66	.65	.84	.77

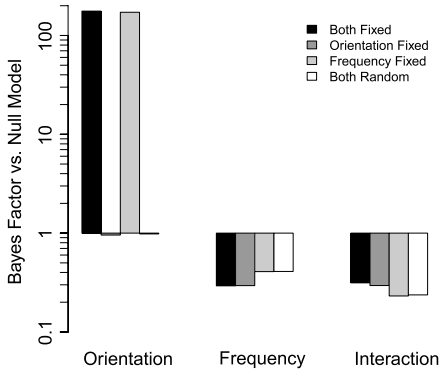

Fig. 5. *Left*: hypothetical response times (sec) to Gabor gratings that vary in orientation (vertical vs. horizontal) and frequency (low vs. high) in a 2×2 design. *Right*: resulting Bayes factor for seven models when effects are modeled as fixed, mixed, or random.

and main effect of frequency favor a null effect. In Simulation III there are main effects of both orientation and frequency, with the main effect of orientation manipulated (0.2, 0.5, 1) and the main effect of frequency held constant at 0.4. As can be seen, the Bayes factors track the true effect sizes well. Simulations IV and V show the case that there are two main effects of the same size, and when there are main effects and interactions, respectively. In all cases, the Bayes factor performs as expected. One desirable property that is evident is an independence or orthogonality. The Bayes factor for one comparison, say the main effect of orientation, does not depend on the true values of the other factors and interactions. This orthogonality mirrors that in conventional ANOVA analysis, and a necessary condition for it is separate g parameters across main effects and interactions.

8.3. A note on fixed, random, and mixed interactions

There is a trend in Bayesian analysis to treat effects as random in ANOVA designs. For one-way ANOVA, the Bayes factor for balanced designs is the same whether the effects are modeled as fixed or random lending credence to the notion that constraint from priors is in some abstract way comparable to explicitly imposing a sum-to-zero constraint. Unfortunately, this general comparability does not hold for interactions. Consider the 2×2 factorial case in which in the random-effects model there are 4 interaction effects, and the constraint comes from the prior in which they are treated as exchangeable. Contrast this to the fixed-effect model where three sum-to-zero constraints are imposed and there is subsequently one interaction parameter. We explore how imposing the sum-to-zero constraints affects the Bayes factor through evaluation of an example.

The table in Fig. 5 shows hypothetical data from Model $\mathcal{M}_\alpha$ in which there are only orientation effects. Classically, the F-value for orientation effect in the fixed-effects model is obtained by dividing MS_A by MS_E, and it evaluates to $F(1, 36) = 17.0$, which, because the degrees-of-freedom in the error term is high, results in a small p-value of 0.0002. For the random-effects model, the F-value is obtained by dividing MS_A by MS_I, the interaction term, and it evaluates to $F(1, 1) = 28.3$. Although this F-value is high, the corresponding p-value is 0.12 because there is a single degree-of-freedom in the error term. In classical statistics, evidence for an orientation effect in this example is more easily detected when the effects are modeled as fixed rather than random. This makes sense: it should be easier to conclude that two levels differ than it is to conclude that all possible levels differ when there are only two in a design.

Our default Bayes factors follow these classical patterns. Fig. 5 shows the resulting Bayes factors for the three contrasts and for four different types of effects models. In the first model, darkest bars, the orientation and frequency are both considered fixed. In the second model, orientation is fixed and frequency is random, and their interaction is mixed with 2 parameters (dark gray bars). Included too is the complementary model (light gray bars) with random orientation and fixed frequency, and the random effects model (white bars), which has 4 interaction parameters. For all four models, there is evidence for a null frequency effect and for a null interaction. These results are appropriate as the data were generated without these effects. There is a discrepancy across the models in the assessment of the orientation main effect. If frequency is considered fixed, the resulting Bayes factors yield strong evidence for an orientation effect; conversely, if frequency is considered random, the evidence is equivocal. Whereas the data are generated with a strong orientation effect, these random frequency models are hiding the underlying structure. The reason they do so is that the random interactions are heavily parameterized. It this case, this heavy parameterization leads to interactions so flexible that they may account for main effect patterns.

This example highlights the usefulness of fixed-effects modeling. In many cases, random-effect models are inappropriate because they are too flexible for the experimental design and the

questions of interest. Because of this increased flexibility, random and mixed interactions should be used with great care. Overall, we think the trend on Bayesian analysis to use random effects as a default is unhelpful and analysts will be served better by careful consideration of context in deciding between fixed and random effects. We think the prevailing rule-of-thumb that sum-to-zero constraints should be imposed for manipulated variables and not imposed for sampled levels is a good one.

9. Within-subject and mixed designs

The above development is appropriate to what are commonly referred to as *between-subject* designs, in which participants are nested within factors. Each participant performs under a single, specific combination of factors, and systematic variability across participants enters into the residual error terms. In within-subject designs, in contrast, participants are crossed with the levels of the factors, and each participant performs in all combinations of the factors. It is reasonable to expect that participants vary substantially, and this variation induces a correlation in performance across conditions. A common approach is to include a separate factor for participant effects. Consider, for example, an experiment in which each participant identifies Gabor gratings at varying orientations. In the psychological literature, this design is commonly referred to as a *one-way within-subject design*, where the one-way refers to the stimulus variable, orientation, and the within-subject refers to the fact that the levels are crossed with participants. Even though this design is called one-way, it is in fact a two-factor design with factors for participants and orientation. Likewise, what is commonly termed a *two-way within-subject design* has three factors: one participant factor and two stimulus factors.

The one-way within-subject design may be modeled with a two-way ANOVA model. The following is appropriate when the stimulus variable is modeled as a fixed effect:

$$\mathcal{M}_f : \boldsymbol{y} = \mu\mathbf{1} + \sigma\left(\boldsymbol{X}_\alpha\boldsymbol{\alpha} + \boldsymbol{X}^*_\beta\boldsymbol{\beta}^* + \boldsymbol{X}^{\cdot *}_\gamma\boldsymbol{\gamma}^{\cdot *}\right) + \boldsymbol{\epsilon}, \tag{18}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}^*$ are parameter vectors that describe the effect of participants and the levels of the stimulus factor, respectively. Included for full generality is the mixed interaction term $\boldsymbol{\gamma}^{\cdot *}$. This term may be estimated if the design is replicated, that is, each participant yields several observations in each condition. In repeated measures designs, in which participants yield a single observation in each condition, it is not possible to distinguish the participants-by-treatment interaction term from the residual. In this case, the appropriate full model is $\mathcal{M}_{\alpha+\beta^*}$.

Mixed designs occur when some factors are manipulated in a within participant manner and others are manipulated in a between participants manner. These designs may be treated analogously to within-subject designs. In mixed designs, the design matrix on participant parameters codes which factors are manipulated in a within-subject manner and which are manipulated in a between-subjects manner.

10. Theoretical properties of Bayes factors with multiple g-parameter priors

In Section 4.3, we listed three desirable properties of the one-sample Bayes factor with a g-prior. These were *scale invariance*, *consistency* and *consistency in information*. Some of these properties are known to apply to the Bayes factor in (9). Scale invariance, for example, is assured because there is a scale-invariant prior on (μ, σ^2), and the model is parameterized in terms of standardized effects rather than unstandardized effects.

Consistency is a more complicated concept in a factorial setting because there are multiple large-sample limits to be considered. Take the case of the two-factor design in which the sample size, N, is the product of three quantities: the number of levels of the first and second factors (a, b), and the number of replicates in a cell r, $N = abr$. The sample size may be increased to the limit by increasing any of these three quantities. Perhaps the simplest case is when r, the number of replicates in a cell, is increased to the limit while a and b are held constant. In this case, the model dimensionality is held constant as sample size increases. A more difficult case is when r is held constant and the number of levels of a factor is increase; i.e., when say a is increased. In this case, increases in sample size correspond to an increase in model dimensionality. This second case is quite important for within-subject designs. In these designs, researchers increase sample size by adding additional subjects rather than by increasing the replicates per subject. Adding additional subjects entails adding more levels, that is, increasing model dimensionality. Hence, it is important to show consistency in the large-model-dimension limit too.

Min (2011) studied the consistency properties of a more general class of priors in various large sample limits. He proved two facts of relevant here. First, if r is increased and the model dimensionality (a, b) is held constant, then Bayes factor (9) is consistent; that is, it approaches zero when the null holds and ∞ when the specified model holds. Second, the Bayes factor is consistent in the large a or large b limit when r is held constant. Therefore, researchers may use multiple g-priors in between-subject, within-subject, and mixed designs with assurance of correct limiting behavior.

To our knowledge, consistency in information, which refers to the correct limit as the R^2 approaches zero or 1, has not been studied in multiple g-parameter priors. It is known to hold for single-g parameter priors (Liang et al., 2008). Consistency in information is not as critical to us as consistency in cell replicates or in model dimensionality, and the lack of theoretical work on this particular type of consistency should not dissuade adoption.

11. Inference with multiple random effects: memory and language

Our development of default Bayes factors for ANOVA is exceedingly general. In this section, we illustrate the generality with an application to memory and language studies. Inference is more complicated in memory and language because in typical designs, researchers sample items from a corpus as well as people from a participant pool. The goal is to generalize the results back to these corpra and populations. Consider a researcher who wishes to know if nouns are read at a different speed than verbs. Suppose the researcher samples 25 nouns and 25 verbs, and asks 50 participants to read each of these 50 words. In this case, there are three factors. The one of substantive interest is the part-of-speech factor (noun vs. verb), which may be modeled as a fixed effect. A second factor is an item factor. Individual nouns and verbs are assumed to have their own systematic effects above and beyond their part-of-speech mean. The final factor is the effect of participants, and each participant is assumed to have his or her own systematic effect.

In many language studies, and in almost all memory studies, researchers average the results across items to construct participant-level scores. These participant-level scores are then submitted to a conventional ANOVA analysis. In the current example, a mean noun and verb reading time can be tabulated for each participant, and these scores may be submitted to a paired t-test to assess the part-of-speech effect. This averaging approach, however, is known to be flawed because the Type I error rate will be inflated over nominal values. Clark (1973) noted that averaging treats items as fixed rather than as random effects, and the correlation in performance

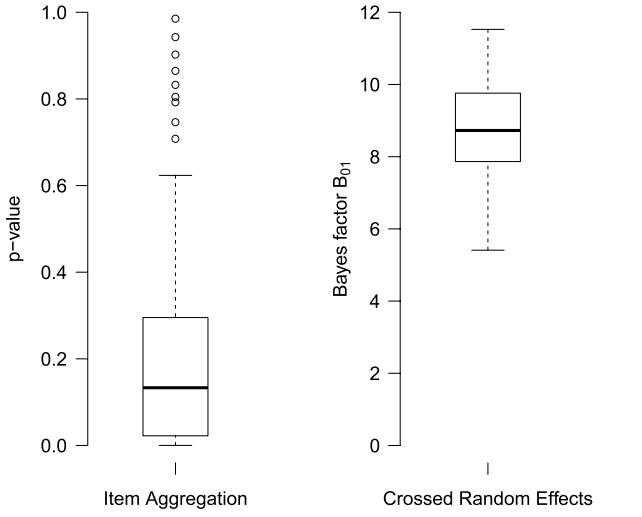

Fig. 6. Simulation of a word-naming experiment with systematic variation across participants and items. Data were generated from a null model in which there was no part-of-speech effect. The p-values are obtained from a t-test on participant-specific noun and verb means. The distribution of these p-values deviates substantially from a uniform, with an over-representation of small values. The Bayes factor are from the same data, but the model includes crossed random effects of people and items. The Bayes factor favors the no part-of-speech effect null model.

across items leads to downward bias in the estimate of residual variability.

To demonstrate this downward bias, we performed a small simulation in which there is no true part-of-speech effect. Participants and items varied, and their individual effects are normally distributed with a standard deviation of 100 ms. The residual error distribution has a standard deviation of 150 ms. We performed 100 replicates in the simulation to explore the distribution of p-values, which is shown in the left box plot in Fig. 6. If there were no distortions due to averaging, then these p-values should be uniformly distributed. The p-values deviate from a uniform distribution, and there is a dramatic over-representation of small values. For a nominal 0.05 level, the observed Type I error rate is 0.34.

Fortunately, researchers in linguistics are well aware of the problem of inflated Type I error rates when items are aggregated. One recommended solution is to specify mixed linear models that treat people and items as crossed random effects (Baayen, Tweedie, & Schreuder, 2002). Mixed models may be analyzed in many popular packages including `Proc Mixed` in SAS, SPSS, and `NMLE` in R. These more advanced models provide suitable Type I error control, that is, if there truly is no part-of-speech effect, the resulting p-values are uniformly distributed. Surprisingly, memory researchers have not adopted crossed random-effects modeling as readily as their linguistics colleagues (cf., Pratte, Rouder, & Morey, 2010).

We show here Bayes factors for crossed-random effects may be conveniently calculated. We implemented the following models to assess the part-of-speech effect for the above example in which 50 participants read 25 nouns and 25 verbs. In this case, there are a total of $N = 50 \times 50 = 2500$ observations. Let $\boldsymbol{X}_\alpha^*$ be a 2500×1 design matrix that indicates whether the item is a noun or verb, and let $\boldsymbol{X}_\lambda$ and $\boldsymbol{X}_\tau$ be 2500×50 design matrices that map people and items into observations respectively. The full model is

$$\mathcal{M}_1 : \boldsymbol{y} = \mu\mathbf{1} + \sigma\left(\boldsymbol{X}_\alpha^*\alpha^* + \boldsymbol{X}_\lambda\boldsymbol{\lambda} + \boldsymbol{X}_\tau\boldsymbol{\tau}\right) + \boldsymbol{\epsilon}, \tag{19}$$

where α^* is a part-of-speech effect, and $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ are person and item random effects, respectively. The null model to assess part-of-speech effects is

$$\mathcal{M}_0 : \boldsymbol{y} = \mu\mathbf{1} + \sigma\left(\boldsymbol{X}_\lambda\boldsymbol{\lambda} + \boldsymbol{X}_\tau\boldsymbol{\tau}\right) + \boldsymbol{\epsilon}. \tag{20}$$

The Bayes factor for the two models is straightforwardly computed via (9), and the results are shown in the right box plot in Fig. 6. The Bayes factor for all 100 replicates of the experiment favor the null model between a factor of 6 and 12. This result is desirable as the data were simulated with no part-of-speech effect. Note here how researchers can state positive evidence for a lack of an effect.

12. Alternative g-priors

In our development, we use separate g parameters for each factor. There are obvious alternatives. One is to use a single g parameter for all effects regardless of factor; a second is to use a separate g-prior for each effect. In this section, we compare our choice to these alternatives.

12.1. A single g-prior

In the single-g prior, there is one g parameter for all main effects and interactions, i.e., $\boldsymbol{G} = g\boldsymbol{I}$. Wetzels et al. (2012), for example, discuss this approach. Clearly, a single-g prior is more computationally efficient as the integral in (9) is single-dimensional for all models. Nonetheless, we think the single-g prior is inferior to the multiple-g prior for general use. When there is one g, the pattern of effects on one factor calibrates the prior on the others through the single g. For instance, take the case of two researchers who wish to test the effect of part-of-speech (noun vs. verb) on word reading times. The first researcher uses one fixed effect, part-of-speech, and presents each word for 300 ms. The second researcher crosses part-of-speech with a second variable, presentation time, which is manipulated across two levels: 299 and 301 ms. If these researchers use a single-g prior, the value of g will be lower for the second researcher to reflect the assuredly null effect of presentation time. Hence, the Bayes factor for tests of part-of-speech will differ, and, in particular, the second researcher will be more likely to interpret small observed part-of-speech effects as evidence for a true effect. The multiple-g prior allows the inference about one factor to be independent of the patterns of effects in the other factors.

12.2. A separate g-parameter for each effect

Each element in the diagonal of $\boldsymbol{G}$ may be specified as a unique parameter, and the marginal joint prior on effects is consequently the independent Cauchy in (4). This prior, which is also a multiple g-parameter prior, has potentially many more g parameters than the previous multiple g-parameter prior as there may be several levels for each factor. To differentiate this prior from the previous one, we call this prior the *independent Cauchy* prior and reserve the term *multiple g-parameter prior* for the recommended one in which each factor rather than each effect is modeled with a separate g parameter.

We argue that the independent Cauchy prior is not ideal for ANOVA designs. Researchers use ANOVA specifically when effects can be decomposed into factors. Factors, by their very nature, have a group structure that imply a certain degree of coherence within a factor. For instance, consider the orientation and frequency factor in the previous example. The different levels of orientation have a coherency in that they all describe a unified property; different levels of frequency also have a coherency. This coherence is captured by the exchangeability of level (Gelman, 2005) as implemented by the correlations in the multivariate Cauchy. With this prior, effects of levels of a factor cannot be arbitrarily different from one another.

There are some designs/models where the independent Cauchy prior is more appropriate than the recommended multiple g-parameter prior, and these designs do not have a factor structure.

For example, consider the question of whether various diverse chemical compounds are agonists for a specific neural receptor. Without some knowledge of the structure of the compounds, there may be little coherency among them with regard to the ensuing receptor activity. In this case, the analyst is not interested in the mean effect of the compounds, or the variation around this mean. Instead, the analyst assesses whether any specific compound serves as an agonist, and there is no hypothetical correlation or structure among the levels. The appropriate model is a cell-means model in which there is a separate standardized effect parameter for each cell, and an appropriate prior is the independent Cauchy prior. The multiple g-parameter prior, in contrast, embeds possible structure among factors and is more appropriate for ANOVA designs in which the analyst is concerned about main effects and interactions.

13. A comparison to default regression priors

In modern statistics it is common to think of ANOVA and regression in a unified linear model framework. Yet, we think researchers should be mindful of some differences when considering categorical and continuous covariates. In the previous development, we advocated priors that led to Bayes factors that were invariant to the location and scale of measurement of the dependent variable. With continuous covariates, it is desirable to consider an additional theoretical property: the Bayes factor should be invariant to the location and scale of the independent variable. For example, consider a researcher wishes to study intelligence as a function of height, the Bayes factor should not depend on whether height is measured in inches or centimeters (or, for that matter, light years or ångströms), The following Bayes factor, from Zellner and Siow (1980), obeys this property.

Let the linear model in (6) hold with the condition that each column of $\boldsymbol{X}$ sums to zero. This condition is not substantive and provides no constraint; it simply guarantees that μ may be interpreted as a grand mean. Zellner and Siow placed the noninformative prior $\pi(\mu, \sigma^2) = 1/\sigma^2$ and the following prior on standardized slopes $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \mid g \sim \text{Normal}(\mathbf{0}, g(\boldsymbol{X}'\boldsymbol{X})^{-1}), \qquad g \sim \text{Scaled Inverse-}\chi^2(N),$$

where the scaled inverse-χ^2 distribution has density

$$f(x; h) = r^{-2}(2\pi)^{-1/2}(x/h)^{-3/2}e^{-h/(2x)}, \tag{21}$$

where h is a scale parameter. It is helpful to rewrite this prior so that the scale factor of N is in the variance of $\boldsymbol{\theta}$ rather than in g:

$$\boldsymbol{\theta} \mid g \sim \text{Normal}(\mathbf{0}, g(\boldsymbol{X}'\boldsymbol{X}/N)^{-1}), \qquad g \sim \text{Inverse-}\chi^2. \tag{22}$$

The difference between the Zellner and Siow prior and the ones we develop for ANOVA (Eq. (7)) is the introduction of a new scaling matrix $\boldsymbol{X}'\boldsymbol{X}/N$ as well as the use of a single g-parameter. Because $\boldsymbol{X}$ is set to be zero-centered, this scaling term can be thought of as the variance or noise power of the covariates. The scaling term is a matrix and includes the covariances between covariates, making it appropriate for nonorthogonal covariates. A helpful interpretation is that there is a g-prior on *double standardized effects*, where effects is standardized to both the variability in the dependent variable and covariates, and are, consequently, without units. This scaling by $\boldsymbol{X}'\boldsymbol{X}/N$ is necessary for the resulting Bayes factor to be invariant to the scale of the independent variable. With this scaling in the prior, Liang et al. (2008) derive the following expression for the resulting Bayes factor against the null model:

$$B_{f0} = \int_0^\infty (1+g)^{(N-p-1)/2}\left[1+g\left(1-R^2\right)\right]^{-(N-1)/2} \times \frac{\sqrt{N/2}}{\Gamma(1/2)} g^{-3/2}e^{-n/(2g)}dg. \tag{23}$$

The introduction of the scaling term $\boldsymbol{X}'\boldsymbol{X}/N$ strikes us as very reasonable for regression applications with continuous covariates, but less so for ANOVA applications with categorical covariates. Consider the basic random effects model given in (10) with a design matrix given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

To meet the requirement that each column sums to zero, we subtract a constant 1/2 from each entry:

$$\boldsymbol{D} = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & -1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

The resulting scale term is

$$\boldsymbol{D}'\boldsymbol{D}/N = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix},$$

which is singular and cannot be inverted in the usual sense. One alternative is to take a generalized inverse, which is proposed by Bayarri and Garcia-Donato (2007). Even with the generalized inverse, we are unsure that the scale term is well calibrated for the ANOVA case. Consider a balanced one-way ANOVA with a levels and r replicates and design matrix $\boldsymbol{X} = \boldsymbol{I}_a \otimes \mathbf{1}_r$. Note that $(\boldsymbol{X}'\boldsymbol{X}/N)^{-1} = a\boldsymbol{I}_a$. Here, the prior variance on the effects is proportional to the number of levels, which is indeed an unsatisfying specification.[6] Moreover, it has undesirable implications for consistency. The Zellner–Siow priors lead to consistent Bayes factors if model dimensionality is held constant, but they do not lead to consistent Bayes factors if model dimensionality increases with sample size (Berger, Ghosh, & Mukhopadhyay, 2003; Liang et al., 2008). In the current example, consistency would fail as a increases and r is kept constant. Because a scales the prior variance, in the limit, this variance would grow without bound. Consequently, the Bayes factor would favor the null regardless of the data. This behavior contrasts unfavorably with the ANOVA priors in (7), which lead to consistent Bayes factors even as model dimensionality is increased with sample size.

We recommend that researchers choose priors based on whether the covariate is categorical or continuous. It is appropriate to use (7) for categorical covariates and (22) for continuous ones. Models with both types of covariates may necessitate prior (7) on the categorical effects and prior (22) on the continuous effects.

14. Computational issues

The critical step in computing Bayes factor is integrating the likelihood function with respect to the prior distribution on parameters. All of the models discussed have parameters μ, and σ^2, and the non-null models have additional parameters $\boldsymbol{\theta}$ and $\boldsymbol{g}$. Computation of Bayes factors, therefore, requires evaluation of high-dimensional integrals. For the proposed models, it is possible to derive closed-form expressions for the integrals over μ, σ^2, and $\boldsymbol{\theta}$, and impossible to do so for $\boldsymbol{g}$. Although the closed-form integration greatly reduces the dimensionality of integration relative to the number of parameters, the resulting integral in (9) is still potentially multidimensional. The evaluation of this integral is the topic of this section.

[6] This dependence of prior variance on the number of levels holds even if one centers the design matrix so that the columns are orthogonal to $\mathbf{1}$ and then takes the generalized inverse.

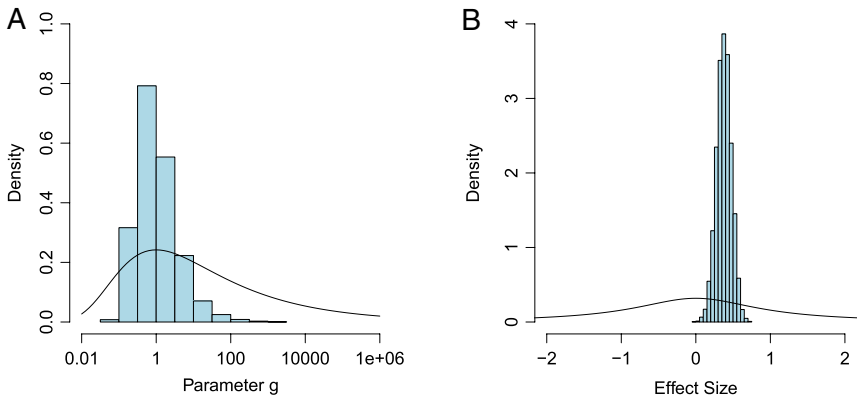

Fig. 7. Concentration of posterior relative to the prior. A. The posterior for g is not too concentrated relative to its prior, and Monte Carlo integration is relatively efficient. B. The posterior for effect-size (α) is more concentrated relative to prior, indicating that the closed-form integration over effects and variance is very helpful. The figure shows the case for a two-group design with 100 observations in each group. The model is a one-way, fixed effects model with four parameters: μ, σ^2, α and g.

We evaluate the multidimensional integral in (9) with a straightforward form of Monte-Carlo sampling. Note that the Bayes factor in (9) may be expressed as an expected value:

$$B_{10} = \mathrm{E}_{\boldsymbol{g}}[S(\boldsymbol{g})],$$

where the expectation is with respect to the prior distributions on $g_1, \ldots, g_r$. This expectation value may be approximated by

$$\mathrm{E}_{\boldsymbol{g}}[S(\boldsymbol{g})] = \frac{1}{L}\sum_{\ell=1}^{L} S(\boldsymbol{g}_\ell),$$

where $\boldsymbol{g}_\ell$ is an r-dimensional random vector sampled from the joint prior on $\boldsymbol{g}$.

The efficiency of this method is a function of the concentration of $S(\boldsymbol{g})$ relative to the prior on $\boldsymbol{g}$. If $S(\boldsymbol{g})$ has large values on just a small range of $\boldsymbol{g}$ relative to the prior, then it will take a great many samples of $\boldsymbol{g}$ to accurately estimate the integral. Conversely, if $S(\boldsymbol{g})$ is spread across a wide range, then Monte Carlo integration may converge quickly. Fortunately, for commonly used designs in experimental psychology, $S(\boldsymbol{g})$ is relatively diffuse, and Bayes factors in (9) may be evaluated quickly and accurately with Monte Carlo integration.

Fig. 7(A) illustrates why simple Monte Carlo integration of g parameters may be effective. Shown are posterior and prior distributions of g for a fixed-effects ANOVA model with two levels on a single factor. There are four parameters for this model: μ, σ^2, α (standardized effect), and g. The critical quantity is $S(g)$, and it is proportional to the ratio of posterior and prior densities. As can be seen, the posterior is not much more concentrated than the prior, implying that $S(g)$ is also fairly diffuse. Although the figure demonstrates the diffuseness of the posterior for a single-g parameter prior, this diffuseness is general and applies to multiple g-parameter priors as well. Fig. 7(B) illustrates that the closed-form integration of the other parameters is necessary. Shown are the posterior and prior distributions of α, on which a standard Cauchy prior is placed. The posterior is quite concentrated, especially relative to the prior, and this concentration will slow convergence of most numerical methods. In fact, the analytic integration of $\mu, \sigma^2, \boldsymbol{\theta}$ in the Appendix is critical for the convenient evaluation of (9).

In several cases, such as one-way ANOVA, or certain regression models, there may be a single g parameter. In this case, the integration in (9) is over a single dimension. We have found that Gaussian quadrature (Press et al., 1992) provides for quick and accurate estimation of Bayes factors, and recommend it for these cases. With large samples, researchers should be cognizant of numerical precision issues, and may have to fine-tune default quadrature algorithms. We have yet to explore Gaussian quadrature integration across multiple dimensions, but given the properties of $S(\boldsymbol{g})$, it may serve as a reasonable alternative to Monte Carlo integration in this context.

There are alternative approaches to evaluating (9) than Gaussian quadrature or simple Monte Carlo integration. In cases where model dimensionality is large, it may prove necessary to use other sampling approaches, such as importance sampling (Ross, 2002) or bridge sampling (Meng & Wong, 1996). Another alternative is to use MCMC-based approaches, such as evaluating the Savage–Dickey density ratio or its generalizations (Chib, 1995; Dickey & Lientz, 1970; Morey, Rouder, Pratte, & Speckman, 2011; Verdinelli & Wasserman, 1995) or transdimensional MCMC (Carlin & Chib, 1995; see Lodewyckx et al., 2011 for a review). One approach that we have tried that does not seem to work as well as hoped is the Laplace approximation (Gelman et al., 2004); the accuracy of the approximation is poor in some cases because the tails on the posterior of g diminish very slowly. The current approach of Monte Carlo sampling directly from the priors seems more convenient than MCMC approaches because the analyst need not worry about mixing.

We have implemented Gaussian quadrature (for one g-parameter priors) and Monte Carlo sampling (for multiple g-parameter priors) in the `bayesfactorPCL` package for the `R` statistical software package. The package is currently in beta development and can be found at https://r-forge.r-project.org/projects/bayesfactorpcl/.

15. Bayes factor for a nonlinear application: skill acquisition

The current development is for linear ANOVA and regression models. Yet, many mathematical psychologists are interested in the analysis of nonlinear models. The current development must be modified, in some cases significantly, to accommodate nonlinearity. There are two general approaches that may prove useful: Laplace approximation and Savage–Dickey density ratio evaluation. The Laplace approximation is based on assuming that the posterior can be well-approximated with a multivariate normal, which can be integrated analytically. Sarbanés Bové and Held (2011) use the Laplace approximation to develop Bayes factors with g priors for the class of generalized linear models (McCullagh & Nelder, 1989). Some psychological process models are members of the generalized linear model class, including Bradley–Terry–Luce scaling models (Bradley & Terry, 1952; Luce, 1959) and a wide class of signal-detection models (DeCarlo, 1998). Several other psychological processing models fall outside the GLM class; examples include response time models with shift parameters that denote the lower bound of support, such as the T_{ER} shift parameter in Ratcliff's (1978) diffusion model.

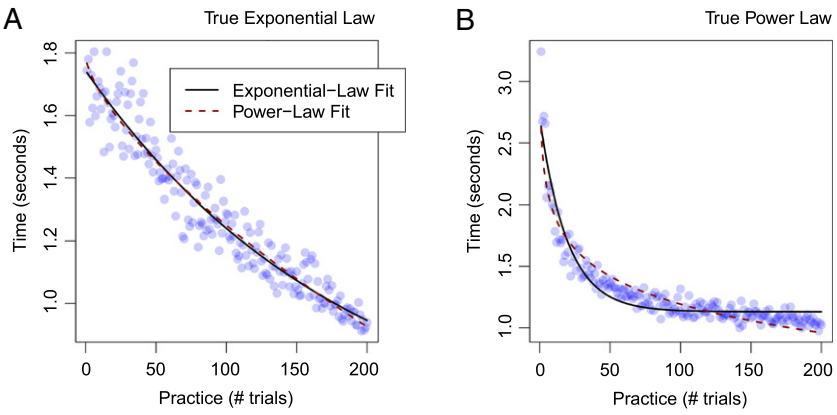

Fig. 8. Mean task completion time as a function of practice for simulated data. Means are averages over 30 participants. Solid and dashed lines show best lest-squares fit for three parameter exponential and power laws, respectively. A. Each individual's data follow an exponential law and were generated with $\mathcal{M}_e$ in (26). B. Each individual's data followed a power law and were generated with $\mathcal{M}_p$ in (27).

In this section we provide an example of a nonlinear application. We develop a few nonlinear models, and compute the multiple-g prior Bayes factor using the Savage–Dickey density estimation approach (Dickey & Lientz, 1970; Morey et al., 2011). Our example is skill acquisition, and we address the particularly hard problem of assessing whether the speeding of the time to complete a task falls as a power function or exponential function of the amount of practice. Prior to the work of Heathcote, Brown, and Mewhort (2000), it was generally accepted that learning followed a power law with practice (e.g., Newell & Rosenbloom, 1981), and this power law speed up was explained by a straightforward race-among-exemplars process (Logan, 1988, 1992). Estimating learning curves has been a particularly vexing problem; although averaging across participants seems attractive to reduce noise, the functional form of the averaged data may not accurately reflect the functional form of individuals (Estes, 1956). Heathcote et al. (2000) provide a particularly lucid description of the problem for assessing whether learning is a power law or an exponential law. The power law describes a more shallow decrease in learning than the exponential. Averaging data is known to artifactually shallow the form of learning. Heathcote et al. showed that if all individuals followed an exponential decrease (a steep form of learning), then the averaged data would approximate a power law even though it was not characteristic of any one individual.

Fig. 8 provides some perspective on the difficulties of adjudicating between power and exponential laws of learning. The panel shows mean task completion time as a function of practice. The left panel shows data where all individuals follow an exponential law, but there is variation in scale and shift across individuals that shallow the aggregate curve. Here, best fitting power law and exponential laws are quite similar, and any analysis of mean data would be inconclusive. The right panel shows data where all individuals follow a power law. Neither power-law nor exponential-law fits to the data are perfect, and each misses in slight but systematic ways. Here we see that it is quite difficult to adjudicate between the functional forms form the analysis of mean data aggregated across individuals.

We develop Bayesian hierarchical nonlinear power-law and exponential-law models with multiple g-priors and use Savage–Dickey density ratio estimation to compute the Bayes factor between the two models. In the experimental setup, a set of a individuals each perform a task J times, and the dependent variable is the time taken to complete the task. Let t_{ij} denote this time for the ith participant on the jth trial, $i = 1, \ldots, a,\ j = 1, \ldots, J$. As people repeat the task, they learn how to do it faster, and their times decrease. We model skill acquisition as a three-parameter, shifted lognormal:

$$\log(t_{ij} - \psi_i) = \mu + \alpha_i + \beta x_j + \epsilon_{ij}, \tag{24}$$

where ψ_i serves an individual-specific shift parameter, μ is a grand mean on the log scale, α_i is a subject effect, x_j is a covariate related to the level of practice, and β is a slope. Noise terms are

$$\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2).$$

The lognormal is a unimodal distribution with an elongated right tail and a relatively soft rise, and all three of these properties are characteristic of response time distributions. Not surprisingly, it has been recommended as both a descriptive model and process model of RT (e.g., Ulrich & Miller, 1993).

The lognormal is a shift-scale-shape model; parameter ψ describes the shift, parameters μ, α, and β describe the scale; parameter σ^2 describes the shape. If $x_j = j$, then the expected value of response time follows an exponential law:

$$E(t_{ij}) = \psi_i + K_i \exp(\beta \times j),$$

where $K_i = \exp(\mu + \alpha_i + \sigma^2/2)$ serves as a constant. Alternatively, if $x_j = \log(j)$, then response time follows a power law:

$$E(t_{ij}) = \psi_i + K_i \times j^{\beta}.$$

The fact that the power and exponential laws correspond to different covariates is very useful. Before stating the models, we standardized the linear covariates ($x_j = j$) and logarithm covariates ($x_j = \log(j)$) so they have a mean of 0 and a variance of 1. For the linear covariates, let $\boldsymbol{w} = 1, \ldots, J$ denote the vector of covariates, and let $\bar{w}$ and s_w be the mean and (population) standard deviation of $\boldsymbol{w}$, respectively. The standardized linear covariate, denoted c_j, is $c_j = (w_j - \bar{w})/s_w$. Likewise for logarithm covariates, let $\boldsymbol{v} = (\log(1), \ldots, \log(J))$, $\bar{v}$, and s_v be the vector of covariates, its mean, and its (population) standard deviation, respectively. The standardized log covariate, denoted d_j, $d_j = (v_j - \bar{v})/s_v$. The following three submodels of (24) serve respectively as a null model (no skill acquisition), an exponential-law model, and a power-law model:

$$\mathcal{M}_0 \ \log(t_{ij} - \psi_i) = \mu + \sigma\alpha_i + \epsilon_{ij}, \tag{25}$$

$$\mathcal{M}_e \ \log(t_{ij} - \psi_i) = \mu + \sigma(\alpha_i + \beta c_j) + \epsilon_{ij}, \tag{26}$$

$$\mathcal{M}_p \ \log(t_{ij} - \psi_i) = \mu + \sigma(\alpha_i + \beta d_j) + \epsilon_{ij}, \tag{27}$$

where effects are standardized with respect to σ^2 as in the previous development. Priors are needed for all parameters. As before, noninformative priors are placed on common parameters:

$$\pi(\boldsymbol{\psi}, \mu, \sigma^2) = 1/\sigma^2,$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_a)$. Separate g-parameter priors are placed on participant and covariate slope:

$$\boldsymbol{\alpha} \mid g_1 \sim \text{Normal}(\mathbf{0}, g_1 \boldsymbol{I}_a),$$
$$\beta \mid g_2 \sim \text{Normal}(0, g_2),$$
$$g_k \overset{\text{i.i.d.}}{\sim} \text{Inverse-}\chi^2(1), \quad k = 1, 2.$$

The key objective is to compute a Bayes factor between the power and exponential law: B_{pe}. This Bayes factor is given by $B_{pe} = B_{p0}/B_{e0}$, where B_{p0} is the Bayes factor between $\mathcal{M}_p$ and $\mathcal{M}_0$, and B_{e0} is the Bayes factor between $\mathcal{M}_e$ and $\mathcal{M}_0$. If the shift ψ_i is zero for all people, then these Bayes factors could be calculated with (9) by placing linear models on $\log y_{ij}$. Unfortunately, empirical distributions of response time robustly exhibit substantial shifts (Rouder, 2005). Therefore, we implement an alternative Savage–Dickey approach to calculate B_{p0} and B_{e0} as follows.

Models $\mathcal{M}_p$ and $\mathcal{M}_0$ are nested and differ by only the inclusion of slope β. Dickey and Lientz (1970) noted that in some cases the Bayes factor can be expressed in this case as a ratio of posterior and prior densities of the parameter of interest, which in this case is β. Because Dickey and Lientz attributed the idea to Savage, this ratio is often called the Savage–Dickey density ratio. The ratio, denoted here as D_{0p}, is

$$D_{0p} = \frac{p(\beta = 0 \mid \boldsymbol{y})}{p(\beta = 0)}, \tag{28}$$

where the numerator and denominator are the posterior and prior probabilities, respectively, that $\beta = 0$ under Model $\mathcal{M}_p$. The Savage–Dickey ratio is equal to the Bayes factor B_{0p} under the following independence conditions. Let $\boldsymbol{\eta}$ denote all parameters that are in $\mathcal{M}_0$, and let π_0 denote the prior density on these parameters. Let π_p be the priors of the same parameters under $\mathcal{M}_p$. Then, the independence condition is given by

$$\pi_p(\boldsymbol{\eta} \mid \beta = 0) = \pi_0(\boldsymbol{\eta}).$$

Fortunately, this condition is satisfied by the standardized effect models in (26) and (27), and, consequently, $D_{0p} = B_{0p}$. The evaluation of the denominator, $p(\beta = 0)$ is straightforward. The marginal prior on β is a Cauchy, and its density evaluated at $\beta = 0$ is $1/\pi$ (where π is the mathematical constant). The evaluation of $p(\beta = 0 \mid \boldsymbol{y})$, the posterior evaluated at $\beta = 0$, is more complicated. Wagenmakers, Lodewyckx, Kuriyal, and Grasman (2010) and Wetzels, Grasman, and Wagenmakers (2010), who first recommended computation of the Savage–Dickey ratio in psychology, used the posterior samples of β to estimate the density at $\beta = 0$ through either splines or through a normal approximation. Chen (1994) and Gelfand and Smith (1990) recommend an alternative called *conditional marginal density estimation* (CMDE) in which the probability that $\beta = 0$ is computed on each iteration of the chain and averaged. Morey et al. (2011) discuss these methods at length and performed a set of simulations to characterize their properties. In all cases, CMDE outperformed the other density estimation methods and is implemented here.

For our case, it is relatively straightforward to set up an MCMC chain to estimate parameters and the posterior density $p(\beta = 0 \mid \boldsymbol{y})$. Derivation of conditional posteriors is straightforward (see Rouder & Lu, 2005, for a tutorial in such derivations). Parameters $\boldsymbol{\psi}$ may be efficiently sampled with Metropolis–Hastings steps; parameters μ, $\boldsymbol{\alpha}$, and β may be treated as one multivariate block and have a joint conjugate prior; variance parameters σ^2, g_1 and g_2 have conjugate priors. These conjugate priors are convenient as the corresponding parameters may be sampled with Gibbs steps. The key step is evaluating the conditional posterior density of β at the point $\beta = 0$, and since this conditional is a normal density, evaluation is computationally convenient.

Bayes factors were computed by Savage–Dickey density ratios for the two sample data sets in Fig. 8. The data for Fig. 8(A) was generated from $\mathcal{M}_e$ with sizable uncorrelated individual variation in ψ and α; the data in Fig. 8(B) was generated from $\mathcal{M}_p$, again with sizable uncorrelated individual variation. The resulting Bayes factor for the data in Fig. 8(A) was $B_{pe} = 7.4 \times 10^{-31}$, indicating strong evidence for the exponential law. This behavior is desirable as the true data were generated with the exponential-law model. Likewise, the resulting Bayes factor for the data in Fig. 8(B) was $B_{pe} = 1.4 \times 10^6$, which is also desirable as the data were generated by the power-law model. As can be seen, g-prior Bayes factor assessment of the functional form of the learning curve is possible and convenient, and it is more principled and more powerful than assessment from averaged data. In general, we suspect that both Savage–Dickey density ratios and Laplace approximations will be useful for developing g-prior Bayes factors in other nonlinear settings.

16. Conclusions

One goal in the analysis of experimental data is the assessment of theories that specify constraints among observables. In service of this goal, there is a general need to be able to state evidence for invariances as well as for effects. The Bayes factor provides an approach to measuring evidence from data for competing theoretical positions, including those that specify invariances. In this regard, it provides a principled approach to accumulating evidence for null hypotheses, an advantage not shared with null hypothesis significance testing.

One of the necessary conditions for increased adoption of Bayes factors is development of default priors with associated algorithms for Bayes factor computation. Our approach herein is an *objective* approach in which priors are chosen based on desirable theoretical properties of the resulting Bayes factor. The defaults we advocate, combinations of multivariate and independent Cauchy priors on effects, lead to well-behaved Bayes factors that are invariant to changes in measurement scale. The resulting computation is convenient, especially when the number of g-parameters is relatively small. We therefore propose the default Bayes factor as a replacement for null hypothesis significance testing in regression, ANOVA, and other linear models.

We conclude with a few caveats. The current class of default priors does not free the researcher from making important decisions in analysis. Some of these decisions, such as specifying whether model factors are to be treated as fixed or random, or specifying which models are to be compared, are not unique to Bayesian statistics. Other decisions, such as those involving the choice of priors, are more specialized. The most consequential of these prior specifications is the prior on effect parameters, and we have recommended a combination of multivariate and independent Cauchy priors as a default position. This recommendation is not a hard and fast rule, and it may be adapted as needed to reflect context about parameter variation, or, perhaps, the goals of inference. For example, researchers who *a priori* believe that small effects in a domain are substantially important and highly probable may wish to use a less diffuse prior than the Cauchy, or, perhaps, a Cauchy of reduced scale.[7]

Finally, we cannot escape the observations that (a) testing is not the most appropriate analysis in many situations, and that (b) testing is performed far too frequently in psychology. In fact,

[7] The development is generalized to a scaled Cauchy prior on standardized effects with scale $\sqrt{h}$ by placing a scaled inverse-χ^2 prior on g with scale h (see Eq. (21)). The marginal likelihood in (8) and Bayes factor in (9) correspond to the case that $h = 1$, but are easily generalized for a scaled inverse-χ^2 prior on g.

we characterize the field as having a testing fetishism in which common-sense exploratory analyses are sometimes ignored in favor of ill-suited hypothesis tests. For example, researchers are quick to resort to ANOVA as the main tool of analysis, even in cases where null hypotheses are implausible *a priori* (Cohen, 1994; Morey & Rouder, 2011). They routinely test all main effects and interactions, often complete with post hoc tests of which levels are significantly different. We note that structure in data is sometimes not elucidated adequately by assessment of effects in linear models, and that exploratory and graphical methods often offer a more suitable alternative (Gelman, 2005; Wilkinson & the Task Force on Statistical Inference, 1999). Researchers using testing, including Bayes factors, should keep in mind that they are trying to divine structure from comparisons among a set of models that are highly simplified representations of nature. The implicit justification for testing is that a comparison among simplified models still yields useful insights into the structure of the data and the relative applicability of various theoretical positions. This justification will not always be present, and when it is not, other methods of analysis are better suited. Although we hope that researchers adopt Bayes factors for their testing, we would find it problematic if they substituted a fetishism with *p*-values for one with Bayes factors. In summary, the following compact tag line, adopted from a current beer commercial,[8] seems highly appropriate: "I don't always test, but when I do, I prefer Bayes factors".

Acknowledgments

The authors thank Brandon Turner and Eric-Jan Wagenmakers for detailed and constructive comments. This research is supported by NSF Grant SES 1024080.

Appendix

Proof of (8). We have the model specification and priors on $\mu, \boldsymbol{\theta}$ and σ^2:

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}, \mu, \sigma^2) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{\theta})\right),$$

$\pi(\mu, \sigma^2) = 1/\sigma^2$, and

$$\pi(\boldsymbol{\theta} \mid \sigma^2, \boldsymbol{g}) = \frac{1}{(2\pi)^{P/2}(\sigma^2)^{P/2}|\boldsymbol{G}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{G}^{-1}\boldsymbol{\theta}\right),$$

where $\mathbf{1}' = (1, \ldots, 1)$ is a vector of length p.

The required marginal likelihood, m, is

$$m = \int_{g_1} \cdots \int_{g_r} f(\boldsymbol{y} \mid \boldsymbol{g})\pi(\boldsymbol{g})\, dg_1 \cdots dg_r,$$

where

$$f(\boldsymbol{y} \mid \boldsymbol{g}) = \int_{\sigma^2}\int_{\mu}\int_{\boldsymbol{\theta}} f(\boldsymbol{y}, \boldsymbol{\theta}, \mu, \sigma^2 \mid \boldsymbol{g})\, d\boldsymbol{\theta}\, d\mu\, d\sigma^2, \tag{29}$$

with

$$f(\boldsymbol{y}, \boldsymbol{\theta}, \mu, \sigma^2 \mid \boldsymbol{g}) = f(\boldsymbol{y} \mid \boldsymbol{\theta}, \mu, \sigma^2)\pi(\boldsymbol{\theta} \mid \sigma^2, \boldsymbol{g})\pi(\mu, \sigma^2). \tag{30}$$

The proof follows by showing that $f(\boldsymbol{y} \mid \boldsymbol{g}) = T_m(\boldsymbol{g})$, where T_m is defined in (8).

Combining terms,

$$f(\boldsymbol{y}, \boldsymbol{\theta}, \mu, \sigma^2 \mid \boldsymbol{g}) = \frac{1}{(2\pi)^{(N+P)/2}(\sigma^2)^{(N+P)/2+1}|\boldsymbol{G}|^{1/2}} \times \exp\left(-\frac{Q}{2\sigma^2}\right),$$

where

$$Q = (\boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{\theta}) + \boldsymbol{\theta}'\boldsymbol{G}^{-1}\boldsymbol{\theta}.$$

Now let $\boldsymbol{P}_0 = \frac{1}{N}\mathbf{1}\mathbf{1}'$ and complete the square in μ to obtain

$$\begin{aligned} Q &= (\boldsymbol{y} - \mathbf{1}\mu)'(\boldsymbol{y} - \mathbf{1}\mu) + N\left(\mu - \frac{1}{N}\mathbf{1}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right)^2 \\ &\quad - \frac{1}{N}(\mathbf{1}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}))^2 + \boldsymbol{\theta}'\boldsymbol{G}^{-1}\boldsymbol{\theta} \\ &= (\boldsymbol{y} - \mathbf{1}\mu)'(\boldsymbol{I} - \boldsymbol{P}_0)(\boldsymbol{y} - \mathbf{1}\mu) + N\left(\mu - \frac{1}{N}\mathbf{1}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right)^2 \\ &\quad + \boldsymbol{\theta}'\boldsymbol{G}^{-1}\boldsymbol{\theta}. \end{aligned}$$

Thus

$$\int_{-\infty}^{\infty} f(\boldsymbol{y}, \boldsymbol{\theta}, \mu, \sigma^2 \mid \boldsymbol{g})\, d\mu = \frac{1}{(2\pi)^{(N+P-1)/2}(\sigma^2)^{(N+P-1)/2+1}|\boldsymbol{G}|^{1/2}\sqrt{N}} \exp\left(-\frac{Q_2}{2\sigma^2}\right),$$

where

$$Q_2 = (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\theta})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\theta}) + \boldsymbol{\theta}'\boldsymbol{G}^{-1}\boldsymbol{\theta},$$

with $\tilde{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{y}$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{X}$. (We have used the fact that $(\boldsymbol{I} - \boldsymbol{P}_0)$ is a perpendicular projection.)

Next, let $\boldsymbol{V}_g = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{G}^{-1}$ so that

$$Q_2 = \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} + (\boldsymbol{\theta} - \boldsymbol{V}_g^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}})'\boldsymbol{V}_g(\boldsymbol{\theta} - \boldsymbol{V}_g^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}) - \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}'\boldsymbol{V}_g^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}.$$

We then have

$$\begin{aligned} &\int_{-\infty}^{\infty}\int f(\boldsymbol{y}, \boldsymbol{\theta}, \mu, \sigma^2 \mid \boldsymbol{g})\, d\mu\, d\boldsymbol{\theta} \\ &= \frac{1}{(2\pi)^{(N-1)/2}(\sigma^2)^{(N-1)/2+1}|\boldsymbol{G}|^{1/2}\sqrt{N}|\boldsymbol{V}_g|^{1/2}} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}'\boldsymbol{V}_g^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}})\right). \end{aligned}$$

Finally, integrating out σ^2 yields (8). □

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological), 53*, 111–142.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: subject variability and morphological family effects in the mental lexicon. *Brain and Language, 81*, 55–65.

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika, 94*, 135–152.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159–165.

Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations to the Bayes factor in model selection problems and consistency issues. *Journal of Statistical Planning and Inference, 112*, 241–258.

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

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika, 39*, 324–355.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton-Mifflin.

[8] The commercial for Dos Equis brand beer ends with the tag line, "I don't always drink beer, but when I do, I prefer Dos Equis". See www.youtube.com/watch?v=8Bc0WjTT0Ps.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 473–484.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association*, *89*, 818–824.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997–1003.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, *4*, 907–949.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

DeCarlo, L. M. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205.

De Finetti, B. (1992). *Probability, induction and statistics: the art of guessing*. Wiley.

Dehaene, S., Naccache, L., Le Clech, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., et al. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

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

Estes, W. K. (1956). The problem of inference from curves based on grouped data. *Psychological Bulletin*, *53*, 134–140.

Eves, H. (1980). *Elementary matrix theory*. Boston, MA: Allyn & Bacon.

Fechner, G. T. (1966). *Elements of psychophysics*. New York: Holt, Rinehart and Winston.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gelfand, A., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Gelman, A., & Shalizi, C.R. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* (in press).

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: on the misperception of random sequences. *Cognitive Psychology*, *17*, 295–314.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in world war II. *Biometrika*, *66*, 393–396.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge, England: Cambridge University Press.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185–207.

Jaynes, E. (1986). Bayesian methods: general background. In J. Justice (Ed.), *Maximum-entropy and Bayesian methods in applied statistics*. Cambridge: Cambridge University Press.

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

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions. volume 2* (2nd ed.). New York: Wiley.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, *42*, 551–560.

Kotz, S., & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge: Cambridge University Press.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Lee, M., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

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

Liu, C. C., & Aitkin, M. (2008). Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *56*, 362–375.

Lodewyckx, T., Kim, W., Tuerlinckx, F., Kuppens, P., Lee, M. D., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

Logan, G. D. (1988). Towards an instance theory of automization. *Psychological Review*, *95*, 492–527.

Logan, G. D. (1992). Shapes of reaction time distributions and shapes of learning curves: a test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 883–914.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Masin, S. C., Zudini, V., & Antonelli, M. (2009). Early alternative derivations of Fechner's law. *Journal of the History of the Behavioral Sciences*, *45*, 56–65.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Meng, X., & Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Min, X. 2011. Objective Bayesian inference for stress–strength models and Bayesian ANOVA. University of Missouri, Department of Statistics.

Morey, R.D., Romeign, W.-J., & Rouder, J.N. The humble Bayesian: model checking from a fully Bayesian perspective (in press).

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

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *54*.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723–739.

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Neter, J., Kutner, M. H., Wasserman, W., & Nachtschiem, C. J. (1996). *Applied linear regression models*. Chicago: McGraw-Hill, Irwin.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*, 224–232.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, F. P. (1992). *Numerical recipes in C: the art of scientific computing* (2nd ed.). Cambridge, England: Cambridge University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ross, S. M. (2002). *Simulation* (3rd ed.). London: Academic Press.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*, 377–381.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin and Review*, *18*, 682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. New York: CRC Press.

Sarbanés Bové, D., & Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, *6*, 1–24.

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

Shibley Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592.

Sternberg, S. (1969). The discovery of processing stages: extensions of Donder's method. In W. G. Kosner (Ed.), *Attention and Performance II* (pp. 276–315). Amsterdam: North-Holland.

Ulrich, R., & Miller, J. O. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*, 513–525.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: an apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618.

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

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics and Data Analysis*, *54*, 2094–2102.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for anova designs. *American Statistician*,.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: a formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1415–1434.

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