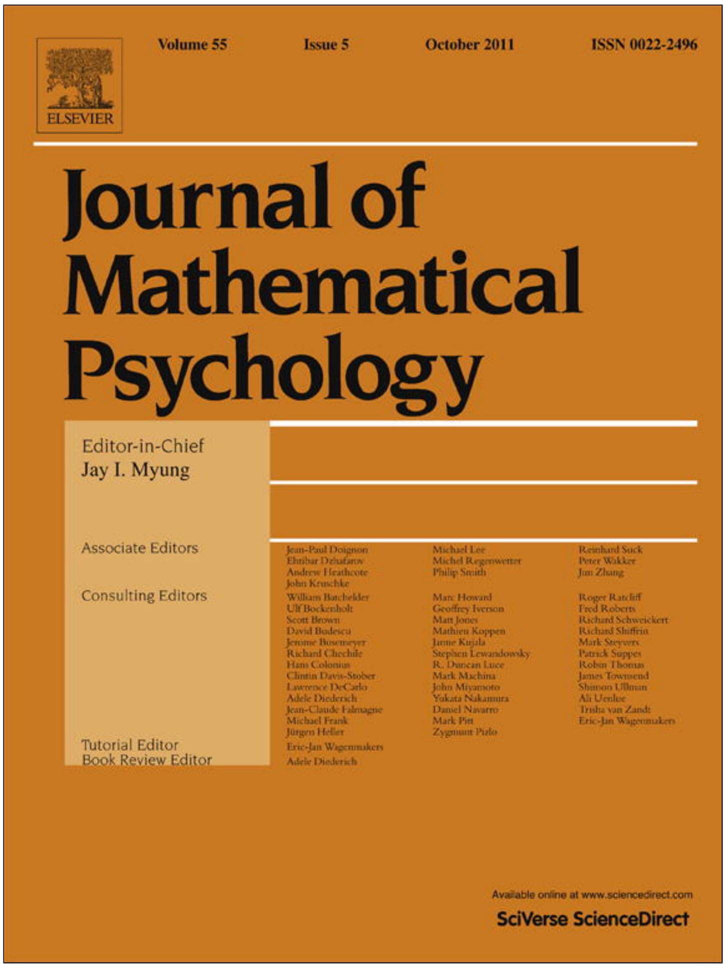
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



Volume 55 Issue 5 October 2011 ISSN 0022-2496

ELSEVIER

Journal of Mathematical Psychology

Editor-in-Chief
Jay I. Myung

Associate Editors
Jean-Paul Doignon
Ehtibar Dzhafarov
Andrew Heathcote
John Kruschke
Michael Lee
Michel Regenwetter
Philip Smith
Reinhard Suck
Peter Wakker
Jun Zhang

Consulting Editors
William Batchelder
Ulf Bockenholt
Scott Brown
David Budescu
Jerome Busemeyer
Richard Chechile
Hans Colonius
Clintin Davis-Stober
Lawrence DeCarlo
Adele Diederich
Jean-Claude Falmagne
Michael Frank
Jürgen Heller
Marc Howard
Geoffrey Iverson
Matt Jones
Mathieu Koppen
Janne Kujala
Stephen Lewandowsky
R. Duncan Luce
Mark Machina
John Miyamoto
Yukata Nakamura
Daniel Navarro
Mark Pitt
Zygmunt Pizlo
Roger Ratcliff
Fred Roberts
Richard Schweickert
Richard Shiffrin
Mark Steyvers
Patrick Suppes
Robin Thomas
James Townsend
Shimon Ullman
Ali Uenlue
Trisha van Zandt
Eric-Jan Wagenmakers

Tutorial Editor
Eric-Jan Wagenmakers

Book Review Editor
Adele Diederich

Available online at www.sciencedirect.com
SciVerse ScienceDirect

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Journal of Mathematical Psychology 55 (2011) 368–378

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

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

Using MCMC chain outputs to efficiently estimate Bayes factors

Richard D. Morey [a,*], Jeffrey N. Rouder [b], Michael S. Pratte [b], Paul L. Speckman [c]

[a] University of Groningen, Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712NS Groningen, The Netherlands
[b] University of Missouri, Department of Psychological Sciences, 210 McAlester Hall, Columbia, MO 65201, United States
[c] University of Missouri, Department of Statistics, 146 Middlebush Hall, Columbia, MO 65201, United States

ARTICLE INFO

Article history:
Received 6 December 2010
Received in revised form
15 June 2011
Available online 3 August 2011

Keywords:
Bayesian statistics
Bayesian inference
Bayes factors
Markov chain Monte Carlo
ANOVA
Encompassing priors

ABSTRACT

One of the most important methodological problems in psychological research is assessing the reasonableness of null models, which typically constrain a parameter to a specific value such as zero. Bayes factor has been recently advocated in the statistical and psychological literature as a principled means of measuring the evidence in data for various models, including those where parameters are set to specific values. Yet, it is rarely adopted in substantive research, perhaps because of the difficulties in computation. Fortunately, for this problem, the Savage–Dickey density ratio (Dickey & Lientz, 1970) provides a conceptually simple approach to computing Bayes factor. Here, we review methods for computing the Savage–Dickey density ratio, and highlight an improved method, originally suggested by Gelfand and Smith (1990) and advocated by Chib (1995), that outperforms those currently discussed in the psychological literature. The improved method is based on conditional quantities, which may be integrated by Markov chain Monte Carlo sampling to estimate Bayes factors. These conditional quantities efficiently utilize all the information in the MCMC chains, leading to accurate estimation of Bayes factors. We demonstrate the method by computing Bayes factors in one-sample and one-way designs, and show how it may be implemented in WinBUGS.

© 2011 Elsevier Inc. All rights reserved.

Frequently, researchers in psychological science must decide which of possibly several theoretical viewpoints is supported by data. For the past century, frequentist statistical methods, such as null hypothesis significance tests and inference by confidence intervals, have been popular in the psychological literature. Although there have been strong arguments for the use of Bayesian methods in psychology for over 50 years (eg, Edwards, Lindman, & Savage, 1963), Bayesian analysis has not been nearly as popular. In fact, there are seemingly more papers in psychology touting the benefits of Bayesian analysis than actually using these analyses to draw conclusions.

One historical reason for this lack of popularity is that obtaining Bayesian quantities often requires significant computational resources. To quantify uncertainty about a statistical parameter, Bayesian methods marginalize over the uncertainty in all other parameters. Marginalizing over all other parameters requires integration over many dimensions, which is often impossible to do analytically. The rise of approximate methods such as Markov Chain Monte Carlo (MCMC; Gelfand & Smith, 1990; Geman & Geman, 1984), and the widespread availability of fast microcomputers has made integration considerably easier, spurring the creation of general tools to perform MCMC analysis (eg, WinBUGS; Lunn, Thomas, Best, & Spiegelhalter, 2000). These tools allow model builders to easily obtain approximate samples from marginal posterior distributions for many useful and relevant models in psychological science (Lee, 2011).

Although the problem of parameter estimation has been largely solved by advances in MCMC methods, model selection in Bayesian contexts remains computationally complicated. We advocate the use of Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995), which we formally define in the next section. The Bayes factor is a Bayesian statistic that quantifies the relative amount of evidence provided by the data for two competing models, and is the ratio of two normalizing constants. MCMC methods used for parameter estimation often make use of the fact that it is possible to sample from distributions without knowledge of normalizing constants. For this reason, MCMC methods designed for parameter estimation, which do not compute normalizing constants, are often not sufficient for model selection.

A number of methods have been proposed to tackle the problem of computing Bayes factors (Meng & Wong, 1996; Raftery, Satagopan, Newton, & Krivitsky, 2007; Verdinelli & Wasserman, 1995), but many of these solutions are difficult to apply, require tailoring to specific problems, or can be unstable in some circumstances. A bright spot among these approaches is computation by the Savage–Dickey density ratio (Dickey,

[*] Corresponding author.
E-mail address: r.d.morey@rug.nl (R.D. Morey).

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.06.004

Author's personal copy

1971; Dickey & Lientz, 1970). Wagenmakers and colleagues (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009) have shown that in some situations the Savage–Dickey ratio does a reasonable job of estimating the Bayes factor using easily-obtained samples from MCMC chains. The fact that the method is implemented with MCMC sampling is highly attractive; one drawback is that the method as presented by Wagenmakers and colleagues is effectively limited to designs with a single effect parameter, such as a t test or regression with a single covariate. It cannot be easily extended, for example, to model selection in factorial designs (ANOVA) in which there are multiple effect parameters. In this article, we discuss an alternative method for computing the Savage–Dickey density ratio, *conditional marginal density estimation* (CMDE; Chen, 1994; Chib, 1995; Gelfand & Smith, 1990). With CMDE estimation of Savage–Dickey density ratios, Bayes factors may be accurately and efficiently estimated from MCMC chains for many designs, including those with multiple effect parameters, such as in ANOVA and multiple regression contexts.

The outline of this paper is as follows: first, we discuss Bayesian model selection via Bayes factor, and show how Bayes factors may be computed using the Savage–Dickey density ratio. We then introduce methods of estimating the Savage–Dickey density ratio, including the CMDE method. The CMDE method is benchmarked against two alternative methods in a one-sample t test design, in which highly accurate estimates of Bayes factor are known for suitable default priors (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Thereafter, we demonstrate how the CMDE method is applied straightforwardly to models with more than one effect parameter, and benchmark CMDE in a one-way ANOVA design.

1. Bayes factor

In psychology, hypothesis testing is the most widely used method of making inferences from data. The goal of hypothesis testing is to assess the evidence provided by the data for or against a hypothesis. In frequentist null hypothesis significance testing, for instance, hypothesis tests assess the evidence against a null hypothesis. In this paper, we approach hypothesis testing from a model selection perspective, in which the null and alternative hypotheses are treated as separate models. The goal of the analyst is to either decide between the two models given data, or to state the evidence for each provided by the data. Model selection by Bayes factor is consistent with the latter goal: Bayes factor provides a measure of the evidence yielded by the data for one model relative to another.

One way of quantifying how well a model accounts for the data is to calculate the probability density of the data given a model. Let $\boldsymbol{y}$ denote observed data, and let $\boldsymbol{\theta}$ denote a possibly multivariate parameter of interest. Throughout, we use characters in bold to represent vectors or matrices. For the moment, we consider hypotheses formed by restricting $\boldsymbol{\theta}$ to a single point. Let $\mathcal{M}_0$ and $\mathcal{M}_1$ denote the restrictions that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, respectively, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are each points, perhaps in a multidimensional space. We can assess how well the model $\mathcal{M}_0$ accounts for the data by using the marginal probability (or density) of the data given the restriction

$$p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0).$$

This density value by itself does not tell us in an absolute sense whether M_0 can account for the data, because it is affected by factors such as sample size. A more interpretable measure of model fit is the density ratio between the two models:

$$\frac{p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0)}{p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_1)},$$

which is the ratio of the likelihood functions evaluated at $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. If the likelihood ratio is greater than 1 the evidence favors $\mathcal{M}_0$, and *vice versa* if the likelihood ratio is less than 1. Likelihood ratios of around 1 are equivocal.

The above example is limited, because parameter values are restricted to single points. It is more useful to consider models in which parameters are not so restricted. When the parameter $\boldsymbol{\theta}$ may take a range of values under a model $\mathcal{M}_k$, the density of the data is

$$p(\boldsymbol{y}|\mathcal{M}_k) = \int_{\Theta_k} p_k(\boldsymbol{y} \mid \boldsymbol{\theta})\pi_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad (1)$$

where $\boldsymbol{\Theta}_k$ is the parameter space of $\boldsymbol{\theta}$ and π_k is the prior distribution of $\boldsymbol{\theta}$ under model $\mathcal{M}_k$. This marginal likelihood provides a Bayesian measure of the evidence from the data for a model. As mentioned previously, it is more interpretable to compare the relative evidence for two models by creating a ratio. For any two models, denoted $\mathcal{M}_0$ and $\mathcal{M}_1$, this ratio, called the *Bayes factor*, is given by

$$B_{01} = \frac{\int_{\Theta_0} p_0(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0}{\int_{\Theta_1} p(\boldsymbol{y} \mid \boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1} = \frac{p(\boldsymbol{y} \mid \mathcal{M}_0)}{p(\boldsymbol{y} \mid \mathcal{M}_1)}. \quad (2)$$

The Bayes factor quantifies the evidence in the data for $\mathcal{M}_0$ relative to $\mathcal{M}_1$. If the Bayes factor is greater than 1, the evidence favors $\mathcal{M}_0$, and *vice versa* if the Bayes factor is less than 1. Bayes factors of around 1 are equivocal. Another way to understand the Bayes factor is that it expresses how prior odds should be updated in light of data. To see this, an application of Bayes' theorem to the marginal likelihood of $\mathcal{M}_0$ yields

$$p(\boldsymbol{y} \mid \mathcal{M}_0) = \frac{p(\mathcal{M}_0 \mid \boldsymbol{y})p(\boldsymbol{y})}{p(\mathcal{M}_0)}.$$

By rewriting the marginal likelihood of $\mathcal{M}_1$ the same way and forming the ratio, we obtain an alternative expression for the Bayes factor:

$$B_{01} = \frac{p(\mathcal{M}_0 \mid \boldsymbol{y})}{p(\mathcal{M}_1 \mid \boldsymbol{y})} \Bigg/ \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}. \quad (3)$$

In Eq. (3), the first term on the right-hand side corresponds to the posterior odds of $\mathcal{M}_0$ relative to $\mathcal{M}_1$. The second term corresponds to the prior odds of $\mathcal{M}_0$ relative to $\mathcal{M}_1$. The Bayes factor is thus the proportional increase in the odds of the model caused by observing the data.

The Bayes factor is a natural way to quantify the evidence provided by the data for one model over another. Since the Bayes factor is a proportional increase in odds, it is straightforward to interpret: a Bayes factor of 1, for instance, means the data provided no evidence for or against either model. A Bayes factor of 2 means the data caused the odds of $\mathcal{M}_0$ to double relative to what they were before observing the data. A Bayes factor of 1/2 means that the data caused the odds of $\mathcal{M}_0$ to decrease to half of what they were prior to observing the data. If $\mathcal{M}_0$ corresponds to a null hypothesis, then the Bayes factor allows the accumulation of evidence both for and against a null hypothesis, which is not possible when using frequentist p values.

Computing a Bayes factor for comparing any two arbitrary models is often computationally difficult. Eq. (1) shows why this is so. In most interesting models, the integration over all parameters will be multidimensional. MCMC methods, which allow for samples from the marginal posterior distribution, do not immediately lend themselves to computing the marginal likelihood; algorithms based on MCMC samples, such as bridge sampling (Meng & Wong, 1996), require custom tuning in application to avoid numerical instabilities.

The situation is considerably simpler, however, if we restrict our attention to nested models. Let $\boldsymbol{\xi} = (\boldsymbol{\theta}', \boldsymbol{\phi}')'$ denote the parameters

Author's personal copy

of a general model, denoted $\mathcal{M}_1$. The null model, denoted $\mathcal{M}_0$, is constructed by setting parameters $\boldsymbol{\theta}$ to a point $\boldsymbol{\theta_0}$ while leaving the nuisance parameters $\boldsymbol{\phi}$ unconstrained. This nested-model setting accounts for the vast majority of the hypothesis tests in psychology, including traditional t tests, ANOVA, and regression. When the two models are nested, the Savage–Dickey density ratio, which is computationally much simpler than other sampling methods, may be used to compute the Bayes factor.

2. The Savage–Dickey method

For the nested-model setup above, the Savage–Dickey method provides a convenient way to compute the Bayes factor, provided certain conditions are met. The marginal probability of the data under the null may be expressed as a restriction of the model $\mathcal{M}_1$: $p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0}, \mathcal{M}_1)$. Consequently, the Bayes factor for the null model relative to the general one is:

$$B_{01} = \frac{p(y|\mathcal{M}_0)}{p(y|\mathcal{M}_1)} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0}, \mathcal{M}_1)}{p(\boldsymbol{y} \mid \mathcal{M}_1)}.$$

Because all quantities are conditioned on $\mathcal{M}_1$, this dependence may be dropped from the notation without ambiguity:

$$B_{01} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0})}{p(\boldsymbol{y})}.$$

An application of Bayes theorem yields

$$p(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0}) = \frac{p(\boldsymbol{\theta} = \boldsymbol{\theta_0} \mid \boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{\theta} = \boldsymbol{\theta_0})}.$$

Substituting and simplifying yields:

$$B_{01} = \frac{p(\boldsymbol{\theta} = \boldsymbol{\theta_0} \mid \boldsymbol{y})}{p(\boldsymbol{\theta} = \boldsymbol{\theta_0})}. \tag{4}$$

Eq. (4) shows that the Bayes factor for the nested restriction is the proportional change from the prior to the posterior in the marginal density of $\boldsymbol{\theta}$ at the point $\boldsymbol{\theta}_0$ under the general model. The construction of the Savage–Dickey ratio is depicted in Fig. 1, which shows hypothetical marginal prior (dotted line) and marginal posterior (solid line) distributions for a parameter θ. Suppose we are interested in the Bayes factor for the restriction $\theta = 100$ against the general model where θ is unrestricted. The prior density at $\theta = 100$ is 0.027. After observing data, we examine the posterior distribution; for the purposes of computing the Bayes factor, we are interested in the posterior density at $\theta = 100$. In this case, the density is 0.242. The increase in density from the prior to the posterior shows that the restriction $\theta_0 = 100$ is highly plausible. The ratio of posterior density to the prior density at $\theta = 100$ is $0.242/0.027 = 9.1$. Consequently, the Bayes factor is $B_{01} = 9.1$.

Alternatively, suppose we are interested in the restriction for $\theta = 104$. The prior density at $\theta = 104$ is 0.026, which is nearly the same as the prior density at 100, the previous null. However, in this case the posterior density (0.004) is much smaller than the prior density. The Bayes factor is $0.004/0.026 = 0.17$, indicating that the unrestricted model is better supported by the data.

Typically, the exact value of the posterior density at any given restriction is unknown. In order to compute the Savage–Dickey density ratio, the posterior density must be estimated at the desired restriction using approximate samples from the posterior distribution of the unrestricted model. Because the Savage–Dickey ratio only requires these easy-to-obtain samples, it is possible to forget that the Bayes factor resulting from the Savage–Dickey ratio is still a ratio of marginal likelihoods. The Bayes factor is not a test that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; it is, instead, a comparison of a model in which $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the unrestricted model from which the samples were drawn. However, there are many possible models in which $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, all differing in the possible models placed on the nuisance parameters. One important consideration is whether the null model implied by the Savage–Dickey density ratio is the one we actually wish to test against.

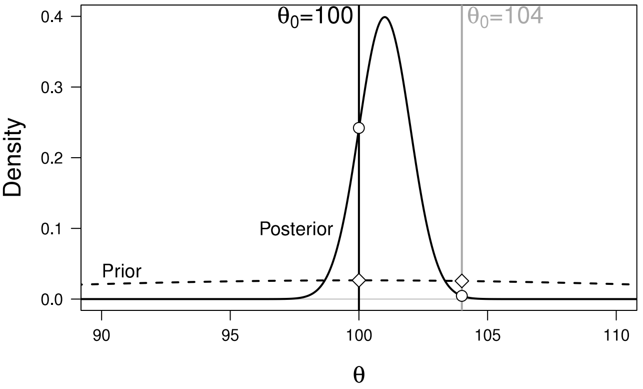


Fig. 1. A demonstration of the Savage–Dickey density ratio. See text for details.

To ensure that the Bayes factor we compute using the Savage–Dickey ratio is the the ratio of marginal densities that we intend, the following condition must hold for the priors under the null and alternative:

$$\pi_1(\boldsymbol{\phi} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0}) = \pi_0(\boldsymbol{\phi}).$$

This requirement is easily met by models which specify priors in which the nuisance parameters are independent of the parameters of interest. In case the priors are not independent, Verdinelli and Wasserman (1995) suggest a correction factor that can be applied to the Savage–Dickey ratio to yield the desired Bayes factor.

The Savage–Dickey ratio is almost always more tractable than analytical integration over all parameters. In order to compute the Savage–Dickey ratio, two quantities are needed: the evaluation of the marginal prior density and of the marginal posterior density of $\boldsymbol{\theta}$ at the specific point $\boldsymbol{\theta_0}$ under the unrestricted model. We denote the marginal prior density at $\boldsymbol{\theta}_0$ as $p_0(\boldsymbol{\theta})$, and the marginal posterior density at $\boldsymbol{\theta}_0$ as $p_0(\boldsymbol{\theta} \mid y)$. In most applications, evaluation of the marginal prior density is straightforward using analytic methods.[1] The evaluation of the marginal posterior density, however, is more complicated, as discussed below.

Wetzels et al. (2009) note that an estimate of the marginal posterior density at a point may be obtained from the output of an MCMC analysis. MCMC chains provide approximate samples from the marginal posterior distribution of $\boldsymbol{\theta}$; the density of the posterior at $\boldsymbol{\theta_0}$, $p_0(\boldsymbol{\theta} \mid \boldsymbol{y})$, is the desired quantity. The challenge is finding good density estimates at a point given MCMC posterior samples. Wagenmakers et al. (2010) suggest *logspline density estimation* (Stone, Hansen, Kooperberg, & Truong, 1997). In this approach, samples from the marginal posterior density of $\boldsymbol{\theta}$ can be used to build an approximation to the logarithm of the density function. The approximation is done via a spline, and when exponentiated provides an estimate of the density at $\boldsymbol{\theta}_0$. Logspline density estimation may be performed conveniently in the *R* statistical language (R Development Core Team, 2009), and the method generally outperforms kernel-based density estimation.[2] A second

[1] In case sampling methods are needed to compute the prior density, the methods described in this paper can be used to estimate them accurately.

[2] Wetzels et al. (2009) propose that a kernel density estimator be used to estimate the density function from the sample, and that smoothing splines be fit to these outputs to estimate the density at the restriction point. One drawback of this method is that there is no constraint that the density estimate at the restriction is positive, and we found in our simulations that if the density was low (as it would be when the general model was supported by the data), the estimate was negative for a sizable fraction of the simulations. Negative density estimates cannot be used to estimate the Bayes factor.

Author's personal copy

approach comes from Wetzels et al. (2009), who use a normal approximation to the posterior distribution. Although marginal posteriors will not be normal in general, the Bayesian central limit theorem (Bernardo & Smith, 2000) implies that marginal posteriors will be asymptotically normal, and so in many cases the assumption will yield estimates that are not too far away from the true value. The mean and variance of the posterior of $\boldsymbol{\theta}$ are computed from the MCMC results; using these estimates of the normal density parameters, the posterior density at $\boldsymbol{\theta}_0$ can be estimated.

There are two separate potential problems with the above two methods. The normal approximation method of Wetzels will clearly suffer to the degree that the posterior density of the parameter of interest is not normal. Deviation from normality, especially in the region of the restriction of interest, will lead to *low quality* (higher mean squared error) estimates of the Savage–Dickey density ratio. It is plausible that estimates using the normal approximation method will be low quality, especially in low sample sizes.

The potential problem with logspline density estimation is qualitatively different. Wagenmakers et al. have benchmarked the logspline approach on models where θ, the parameter of interest, describes the location of a univariate normal distribution. Although this is an important case, and underlies the one-sample t test, most contexts demand more than restrictions of a single location parameter. For example, one-way ANOVA posits separate location parameters for each group, and the appropriate null model is that all of these effects are restricted to be zero. If Savage–Dickey ratio computation of the Bayes factor is to be broadly applicable, it is important that methods of estimating marginal posterior densities work well for multivariate restrictions.

Logspline density estimation does not scale well as the dimensionality of the parameter space increases. The reason for this is the "curse of dimensionality" (Scott, 1992). Suppose, for instance, that we are interested in the bivariate null hypothesis that $(\theta_1, \theta_2) = (0, 0)$. In order to estimate the density at the null point, we need an amount of data in some neighborhood of (0, 0); say, between some $-\epsilon$ and ϵ in both dimensions. Suppose the posterior probability that $-\epsilon < \theta_1 < \epsilon$ is 0.01, and likewise for θ_2, and that they are nearly independent. In this case, only 1 of every 10,000 samples from the posterior distribution will be in the neighborhood of the null hypothesis, compared with 1 out of every 100 for each univariate hypothesis. If we add another similar dimension, the proportion within the neighborhood of (0, 0, 0) becomes 1 in 1,000,000. In general, obtaining quality estimates of the density around a point when there are so few samples around the point is practically impossible. Hence, it seems doubtful that logspline density estimation will be broadly applicable in everyday designs.

3. Improved Savage–Dickey estimates

Logspline density estimates and normal approximations use marginal posterior samples of $\boldsymbol{\theta}$ as input but do not rely on samples of $\boldsymbol{\phi}$, the parameters in common across the full and restricted models. At first glance, using samples from $\boldsymbol{\theta}$ may appear reasonable; after all, the marginal posterior density of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ is exactly the quantity of interest. Yet, the sample of $\boldsymbol{\phi}$, in conjunction with the data, provided all the information used to sample $\boldsymbol{\theta}$ in the MCMC chain. Gelfand and Smith (1990) noted that they also contain all the information available in the chain to compute the density of $\boldsymbol{\theta}$ at a point $\boldsymbol{\theta}_0$.

Gelfand and Smith propose the following approach for estimating the marginal posterior density at a restriction. The key insight is that the marginal posterior samples of $\boldsymbol{\phi}$ may be combined with known information about the conditional distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \boldsymbol{y})$:

$$\begin{aligned} p_0(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \int_{\boldsymbol{\Phi}} p_0(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \boldsymbol{y}) p(\boldsymbol{\phi} \mid \boldsymbol{y})\, d\boldsymbol{\phi} \\ &= \mathrm{E}_{\boldsymbol{\phi}|\boldsymbol{y}} \left[p_0(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \boldsymbol{y}) \right]. \end{aligned} \tag{5}$$

In practice, we do not know the form of the marginal posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{y})$. But we do have MCMC samples approximating this distribution, which enables us to approximate the expected value in Eq. (5) by

$$\mathrm{E}_{\boldsymbol{\phi}|\boldsymbol{y}} \left[p_0(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \boldsymbol{y}) \right] \approx \frac{1}{T} \sum_{t=1}^{T} p_0(\boldsymbol{\theta} \mid \boldsymbol{\phi}^{(t)}, \boldsymbol{y}) \tag{6}$$

where $\boldsymbol{\phi}^{(t)}$ is the tth MCMC sample from the marginal posterior of $\boldsymbol{\phi}$, out of T total MCMC iterations. Because the method uses the conditional information to compute the marginal density, Chen (1994) called the resulting estimator the conditional marginal density estimator, or CMDE.

Gelfand and Smith (1990) use the Rao–Blackwell theorem (Blackwell, 1947) to show that any kernel density estimator that does not make use of $\boldsymbol{\phi}$ will have mean squared error as large or larger than one that is conditioned on $\boldsymbol{\phi}$. The Rao–Blackwell theorem shows that conditioning on a sufficient statistic will tend to improve an unbiased estimator: the conditioned estimator will have MSE less than or equal to the unconditioned estimator. This is intuitive, because sufficient statistics include all the information necessary in the data to estimate a parameter. In our case, we seek to estimate the expected value in Eq. (5). Since the complete sample is always sufficient, conditioning any unbiased estimator on $\boldsymbol{\phi}$ will tend to improve it, and will never make it worse. Although the Rao–Blackwell theorem applies to unbiased estimators, one would expect that the general strategy of using all the information available to estimate a quantity would yield better estimates than any strategy that does not. Chen and Shao (1997), for instance, show that the CMDE consistently outperforms a kernel density estimate in a simple constrained-parameter linear regression example.

Computing the CMDE requires that the full conditional posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$ be known completely; the normalizing constant must be known so that the conditional posterior distribution integrates to 1. Although this may appear to be a difficult restriction to meet, it is often possible to build models in such a way that the normalizing constant will be known. For example, if the prior for $\boldsymbol{\theta}$ is a known form and conjugate or semi-conjugate (Gelman, Carlin, Stern, & Rubin, 2004), then the normalizing constant will be known. When building the model it is sufficient to specify a single conjugate prior – the prior on the parameter of interest – to make using the CMDE possible. In cases where the model cannot be formulated in this manner, a generalization of CMDE called is possible; we address this generalization briefly in the discussion.

In this report, we assess the feasibility of CMDE across two applications. We will show that CMDE has tremendous advantages and outperforms competitor methods. We specifically highlight the following critical advantages:

High quality. CMDE provides higher quality estimates of the Savage–Dickey density ratio than other methods; these estimates have both lower variability and lower bias.

Extension to multiple effects. CMDE is easily applied in multivariate contexts where multiple parameters are restricted simultaneously. Consequently, it may be used broadly.

Practical ease of computation. CMDE may be easily computed in flexible MCMC programs such as WinBUGS (Lunn et al., 2000) or JAGS (Plummer, 2003). Computation does not rely on external routines, such as spline approximation, and there is no

Author's personal copy

need to switch between multiple software applications. Example WinBUGS code may be found in the Appendix.

Availability of error estimates. CMDE is a sampling-based approach to integration. Because Monte Carlo integration is approximate, there will be some error in the estimate of the posterior marginal density. Fortunately, under mild conditions, this error is guaranteed to decrease to 0 as the number of samples is increased, but estimates based on finite numbers of samples will always contain some error. The error in MCMC chains is difficult to estimate, due to the fact that MCMC chains are autocorrelated. However, there are a number of methods available for estimating MCMC error even with autocorrelation (Heidelberger & Welch, 1983; Roberts, 1996), and these methods are applicable to CMDE estimates as well. For example, WinBUGS offers an estimate of MCMC error based on Roberts' batch means method. If WinBUGS is used to compute the CMDE and the Bayes factor, an estimate of the MCMC error in the Bayes factor estimate is provided automatically.

The remainder of this paper comprises two applications of CMDE to common hypotheses. The first of these is a CMDE computation of the Bayes factor for the one-sample t test. The computation of this Bayes factor is well understood, and Rouder et al. (2009) provide a one-dimensional integral expression that may evaluated to high precision by Gaussian quadrature. Hence, Rouder et al. 's numeric solution serves as a highly accurate estimate against which we may compare the success of the CMDE, logspline and normal approximation approaches. We show that CMDE outperforms the other two approaches. The second application is one-way ANOVA, and this application affords assessment of CMDE for multivariate restrictions. To check the quality of the estimates from each method, we derive an expression for one-way ANOVA Bayes factor that involves only a one-dimensional integral. Consequently, these integrals may be evaluated with Gaussian quadrature, enabling a comparison of the CMDE to a highly-accurate estimate.

4. One-sample *t* test

4.1. Model and priors

We first outline the one-sample t test model, and then describe how estimates of the Bayes factor may be obtained. As is conventional to assume, the likelihood of the data is normal. It is convenient to parameterize the model in terms of standardized effect size $\delta = \mu/\sigma$:

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

where $i = 1, \ldots, N$ indexes participant. We place a conventional noninformative Jeffreys prior on σ^2:

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

We must also place a prior on δ, the parameter of interest in the general model. In Bayesian parameter estimation, it is typical to place a flat prior on unbounded parameters such as δ, giving equal prior weight on all real values. This non-informative prior minimizes the effect of the prior on the posterior. However, for model selection, a flat prior would be inappropriate: in fact, the Bayes factor would not exist. The Savage–Dickey density ratio provides insight into the reason. Instead of a flat prior density, consider a normal prior density with very large variance. As the prior variance increases to approximate a flat prior, the density at any single point becomes arbitrarily small. The posterior density at the same point, however, does not. To compute the Bayes factor by means of the Savage–Dickey density ratio, the posterior density at $\delta = 0$ is divided by the prior density at $\delta = 0$. Since the prior density can be made arbitrarily close to 0, the Bayes factor can be made arbitrarily large. With a flat prior (corresponding to infinite prior variance) the Bayes factor thus becomes infinitely large.

To avoid this problem, Rouder et al. followed Jeffreys (1961) and in placing an informative t distributed prior with a single degree of freedom, also known as the Cauchy distribution, on δ:

$$\delta \mid r \sim \text{Cauchy}(r),$$

where r serves as a scale term on effect size that is set *a priori* by the analyst. The scaled Cauchy has density function

$$p(\delta \mid r) = \frac{1}{\pi r\left[1 + \left(\frac{\delta}{r}\right)^2\right]}, \quad \delta \in \mathbb{R}.$$

Rouder et al. choose a value of $r = 1$ as a default value, which corresponds to a standard Cauchy prior on effect. Liang, Paulo, Molina, Clyde, and Berger (2008) note that the Cauchy distribution is the result of a mixture of normal random variables, if the normal random variables have variances (denoted here by g) drawn from an inverse gamma distribution:

$$\delta \mid g \sim \text{Normal}(0, g)$$

$$g \sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r^2}{2}\right).$$

This inverse gamma distribution has density function

$$p(g \mid r) = r\pi^{-\frac{1}{2}} g^{-\frac{3}{2}} \exp\left\{-\frac{r^2}{g}\right\}, \quad g > 0.$$

Integrating out g yields a marginal Cauchy distribution on δ with scale r.

4.2. Benchmark Gaussian quadrature computation

The Bayes factor for the t test model may be found by computing the ratio of the marginal likelihoods under $\mathcal{M}_0$ and $\mathcal{M}_1$:

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

$$\mathcal{M}_1 : \delta \sim \text{Cauchy}(r).$$

Rouder et al. show that the Bayes factor may be expressed as a one-dimensional integral, and consequently can be evaluated using Gaussian quadrature. An easy-to-use web applet is provided at http://pcl.missouri.edu/bayesfactor. To use this applet, researchers input the t statistic and sample size; the applet outputs the corresponding Bayes factor. The results from numeric integration by Gaussian quadrature are highly accurate and serve as a benchmark to evaluate the Savage–Dickey density ratio estimation methods.

4.3. Savage–Dickey density ratio estimation

The Savage–Dickey density ratio comprises two elements: an evaluation of the marginal prior and of the marginal posterior at the restriction. The evaluation of the marginal prior at $\delta = 0$ is straightforward. Because the marginal prior distribution of δ is a scaled Cauchy distribution, the prior density at 0 is

$$p_0(\delta) = \frac{1}{r\pi}.$$

To estimate the marginal posterior density, we sampled from the marginal posterior distributions of all parameters using Gibbs sampling (Geman & Geman, 1984; see Rouder & Lu, 2005) for a tutorial for psychologists). We implement the Gibbs sampler in an R package available from the first author's website at http://drsmorey.org/research/rdmorey. Although we chose to

Author's personal copy

implement the Gibbs sampler in R, any method of sampling from the marginal posterior distributions will enable the computation of the CMDE, including WinBUGS and JAGS. An example showing how to estimate the t test Bayes factor using the CMDE estimate using WinBUGS is given in the Appendix.

Logspline density estimates were obtained by submitting the MCMC samples of δ to the `dlogspline()` routine in the logspline package in R (Stone et al., 1997) with default settings. The normal approximation method was implemented as in Wetzels et al. (2009), by matching the mean and variance of the MCMC samples δ and finding the density at 0 of the resulting normal distribution.

The CMDE estimate of the marginal posterior density at $\delta = 0$ was computed by MCMC integration of the full conditional posterior density. The conditional posterior density of δ is

$$\delta \mid \sigma^2, g, \boldsymbol{y} \sim \text{Normal}\left(\frac{\bar{y}}{\sigma}\left(\frac{N}{N+\frac{1}{g}}\right), \frac{1}{N+\frac{1}{g}}\right). \tag{7}$$

The proof of this statement is straightforward and omitted for brevity. The value of this density at $\delta = 0$ is

$$p_0(\delta \mid \sigma^2, g, \boldsymbol{y}) = \sqrt{\frac{N+\frac{1}{g}}{2\pi}} \exp\left\{-\frac{N^2\bar{y}^2}{2\left(N+\frac{1}{g}\right)\sigma^2}\right\}. \tag{8}$$

On every iteration of the MCMC chain, we compute Eq. (8) given the values of σ^2 and g for that iteration. This yields a chain of values, the mean of which is the CMDE, by Eq. (6). We denote the CMDE marginal posterior estimate as $\hat{p}(\delta \mid \delta = 0, \boldsymbol{y})$:

$$\hat{p}_0(\delta \mid \boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} p_0(\delta \mid (\sigma^2)^{(t)}, g^{(t)}, \boldsymbol{y}).$$

4.4. Results

In order to test the three methods of estimating the Savage–Dickey ratio, we generated data with a range of observed effect sizes and sample sizes (16 and 256), and estimated posterior quantities for chains of varying numbers of iterations (100 and 10,000). Because the t test Bayes factor only depends on t, our data were 60 t values in equal increments from 0 to 6. For each simulated data set, we computed the three Savage–Dickey Bayes factor estimates, as well as the benchmark Gaussian quadrature value. The error in each of the three Savage–Dickey methods was the ratio of the sampling-based estimate to the benchmark quadrature value. For instance, if a logspline estimate of the Bayes factor was 10 times the value obtained from Gaussian quadrature, the corresponding Bayes factor error was 10.

Fig. 2A shows the error in the Bayes factor for the logspline method as a function of the Gaussian quadrature estimate for $M = 100$ MCMC iterations. The open circles show the error when sample size $N = 256$; the gray triangles show the error when $N = 16$. For both sample sizes, the logspline estimates show a marked pattern: as the Bayes factor decreases (that is, as the evidence favors the alternative over the null), estimates of the Bayes factor become biased toward the null model. In extreme cases, this bias toward the null model can be more than a factor of 10,000.

One possible way of improving the logspline density estimates is to increase the number of MCMC iterations. A sample of $M = 100$ MCMC iterations is quite small, and would not necessarily be expected to yield good estimates of the Bayes factor. Fig. 2B shows the error from the logspline estimates for $M = 10{,}000$ MCMC iterations, an increase of two orders of magnitude. Although we have drastically increased the number of MCMC iterations, the logspline estimated Bayes factors show the same distortions as with $M = 100$. Although the bias when $M = 10{,}000$ is less than when $M = 100$, the errors can still reach factors of 1000. Increasing the number of iterations does increase the accuracy of the estimates, but the logspline does not offer an estimate of error. In practice, researchers have no way of assessing the accuracy of their logspline-derived Bayes factor estimates.

Figs. 2C and D shows the error for the normal approximation method for $M = 100$ and $M = 10{,}000$ MCMC iterations, respectively. Estimates of the Bayes factors are greatly improved over the logspline method. Even so, there is a systematic bias for low sample sizes ($N = 16$; gray triangles) in which estimated Bayes factors are biased toward the null model. This bias is a direct consequence of the normal distribution being a poor approximation to the posterior for low sample sizes; the bias exists regardless of the length of the MCMC chain, but is most noticeable for $M = 10{,}000$ because the error due to MCMC sampling is small.

Errors in estimating the Bayes factor using the CMDE method with $M = 100$ and $M = 10{,}000$ MCMC iterations are shown in Figs. 2E and F. The CMDE method is superior to the other two methods over the entire range of Bayes factors considered. Low sample sizes, which led to bias in the normal approximation, lead only to a slight increase in the variability of CMDE estimates; however, the CMDE appears unbiased across all Bayes factors. The lack of bias from the CMDE method is not surprising. Because the CMDE is the sample mean of random variables that all have the same expected value, and this expected value is exactly the quantity of interest, it is guaranteed to be unbiased after the initial burn-in of the MCMC chain. This is a guarantee not provided by the other methods. However, the CMDE also outperforms the other methods in efficiency: the CMDE estimates at just 100 MCMC iterations are nearly as high-quality as the estimates of the normal approximation at 10,000 samples, and are much higher quality than the logspline estimates at 10,000 samples.

One surprising result from the foregoing simulations is the relatively poor performance of logspline density estimation in this context. Previously, our concern with the logspline method was generalizing the technique for multivariate null hypotheses, such as in ANOVA. Here, we see logspline does poorly even in a univariate case. Using spline density estimates in multivariate contexts is computationally intensive; given the poor univariate results and the difficulty of extending splines, we eliminate the splines from consideration in our next example. In the next section, we evaluate the performance of the remaining two Savage–Dickey methods, normal approximation and CMDE, in a multivariate context.

5. One-way, between-subjects ANOVA

Our first example showed how the CMDE method can be applied to Rouder et al. 's t test Bayes factor. Although the t test is one of the first statistical tests that students learn in introductory statistics classes, it is not as commonly used in practice as other statistical tests, such as ANOVA. The main feature of ANOVA is a multivariate null hypothesis in which all group effects are zero. We assess the performance of both normal approximation and CMDE by comparing each to the following Gaussian quadrature solution.

5.1. Model and priors

The one-way ANOVA model we present here is a generalization of Rouder et al. 's t test. Following ANOVA conventions, we assume that the ith ($i = 1, \ldots, N_j$) observation in the jth ($j = 1, \ldots, J$) group is normally distributed:

$$y_{ij} \sim \text{Normal}(\mu + \delta_j\sigma, \sigma^2)$$

Author's personal copy

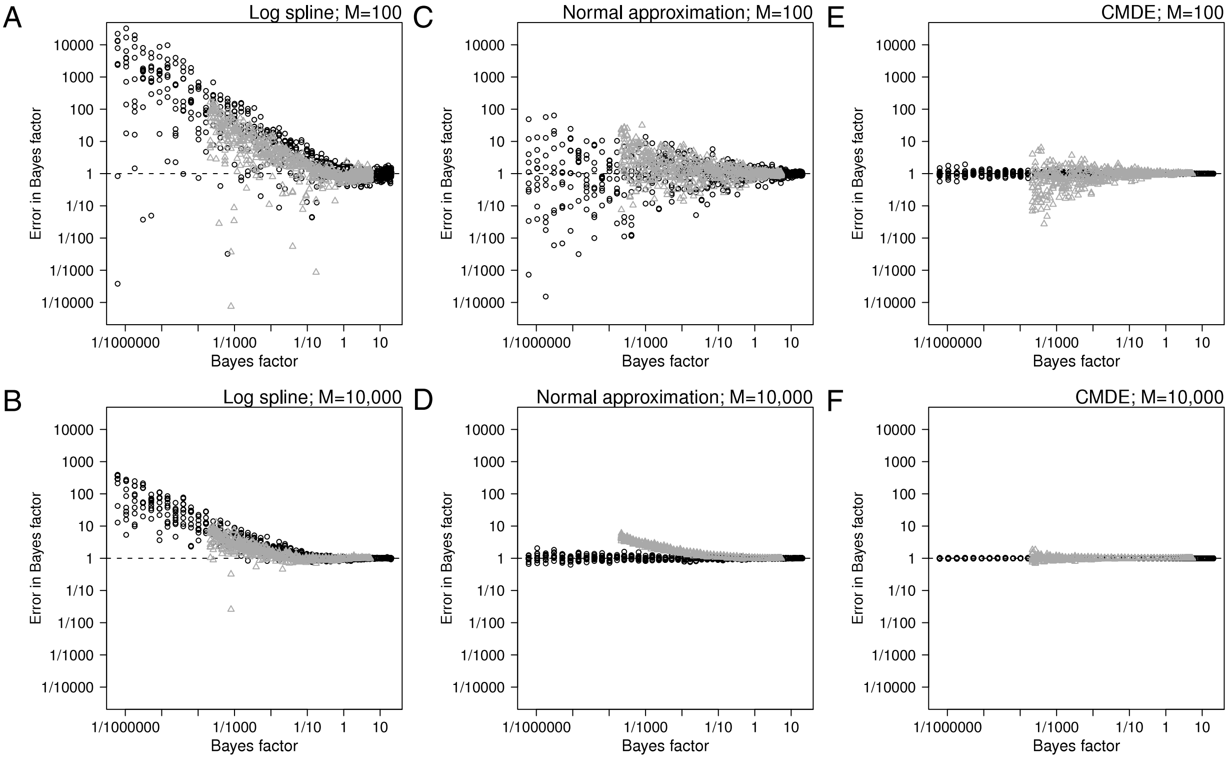


Fig. 2. Multiplicative error in Bayes factor estimation in the t test model for three methods: Logspline (left column), normal approximation to the posterior (center column), and CMDE (right column). The error is the ratio of the estimated Bayes factor to the true Bayes factor. For each method, the top and bottom plots show errors for 100 and 10,000 MCMC iterations, respectively. Sample size $N = 16$ is represented by the gray triangles; $N = 256$ is represented by the black circles.

where μ is the grand mean, σ^2 is the error variance, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)'$ is the vector of standardized effects for each group. Under the null model $\mathcal{M}_0$, $\boldsymbol{\delta} = \mathbf{0}$.

We first place priors on μ and σ^2. Because these parameters are common to both models, we place noninformative Jeffreys priors on both:

$$[\mu] \propto 1$$

$$\left[\sigma^2\right] \propto \frac{1}{\sigma^2}.$$

Of critical importance is the prior on $\boldsymbol{\delta}$ under the unrestricted model. Here, we choose a multivariate generalization of the Cauchy distribution, the multivariate Student's $t(1)$ distribution (Kotz & Nadarajah, 2004):

$$\boldsymbol{\delta} \sim t\left(\nu = 1, \boldsymbol{\Sigma}_0 = r^2 \boldsymbol{I}_J\right),$$

where ν denotes the degrees of freedom of the multivariate t distribution, $\boldsymbol{\Sigma}_0$ denotes a covariance matrix, and $\boldsymbol{I}_J$ denotes the $J \times J$ identity matrix. This multivariate t distribution has density function

$$p(t \mid r) = \frac{\Gamma\left(\frac{J+1}{2}\right)\pi^{-\frac{J+1}{2}} r^{-J}}{\left(1 + \frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{r^2}\right)^{-\frac{J+1}{2}}}.$$

In the t test model, we made use of the fact that the Cauchy distribution arises from a mixture of univariate normal distributions with variances drawn from an inverse gamma distribution. A generalization of this fact applies in the multivariate case. If

$$\boldsymbol{\delta} \mid g \sim \text{Multivariate Normal}\left(\mathbf{0}, g\boldsymbol{I}_J\right), \quad \text{and}$$

$$g \sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r^2}{2}\right),$$

then the marginal prior distribution of $\boldsymbol{\delta}$ is a multivariate Student $t(1, r^2\boldsymbol{I}_J)$.

5.2. Benchmark Gaussian quadrature computation

The Bayes factor for balanced designs ($N_j = N$ for all j) is given by

$$B_{01} = \frac{1}{\frac{r}{\sqrt{2\pi}}\int_0^\infty g^{-\frac{3}{2}} \exp\left\{-\frac{r^2}{2g}\right\}(Ng+1)^{-\frac{NJ-J}{2}} K^{-\frac{NJ-1}{2}}\,dg}, \qquad (9)$$

where

$$K = 1 + \frac{Ng}{\left(\frac{J-1}{NJ-J}\right)F + 1},$$

and F is the F statistic available from conventional ANOVA analysis (Liang et al., 2008).[3] Because the integral is one-dimensional, it may be evaluated to high precision with Gaussian quadrature and will serve as a benchmark for evaluating the Savage–Dickey methods.

5.3. Savage–Dickey density ratio estimation

The Savage–Dickey density ratio comprises two elements: first, an evaluation of the marginal prior, and second, the marginal posterior at the restriction. Once again, the evaluation of the marginal prior at $\boldsymbol{\delta} = \mathbf{0}$ is straightforward. Because the marginal prior distribution of $\boldsymbol{\delta}$ is a multivariate t distribution, the prior density at $\mathbf{0}$ is

$$p_0(\boldsymbol{\delta}) = \Gamma\left(\frac{J+1}{2}\right)\pi^{-\frac{J+1}{2}} r^{-J}.$$

[3] The one-way ANOVA Bayes factor may be derived from Eq. (13) in Liang et al. (2008), with the prior in their Eq. (15), substituting for n the within-group sample size N. See also Zellner and Siow (1980).

Author's personal copy

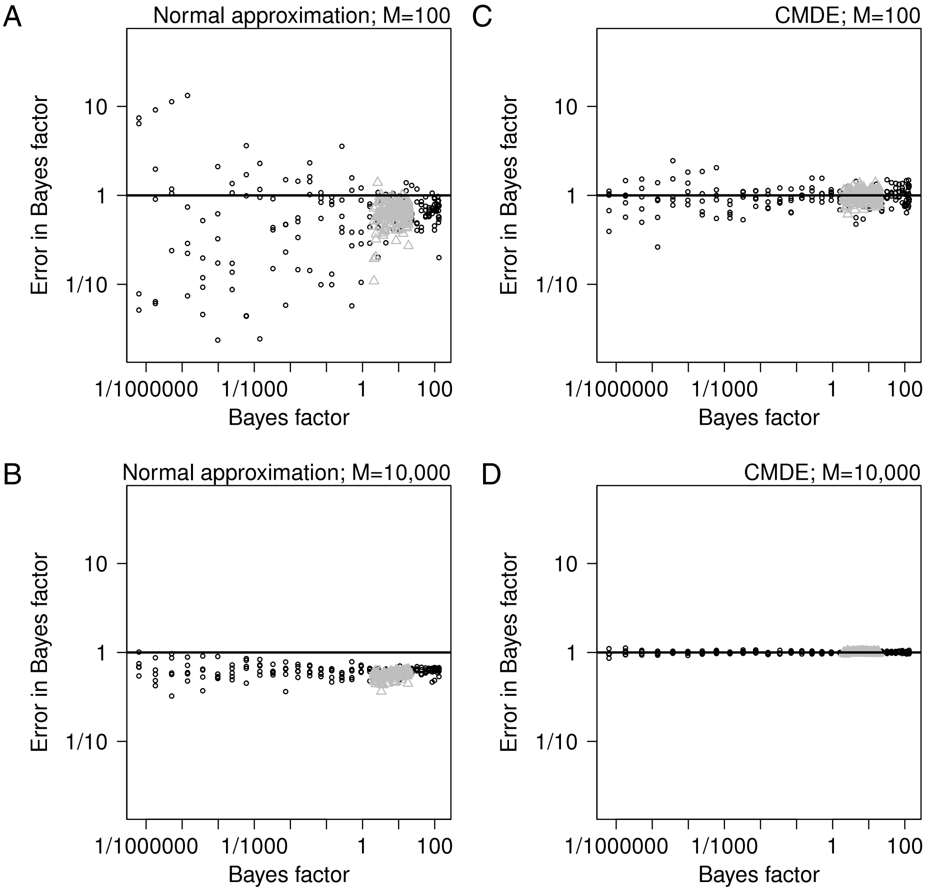


Fig. 3. Errors in Bayes factor estimation in the one-way ANOVA model for two methods: normal approximation to the posterior (left column), and CMDE (right column). For each method, the top and bottom plots show errors for 100 and 10,000 MCMC iterations, respectively. Sample size $N = 16$ is represented by the gray triangles; $N = 128$ is represented by the black circles.

To estimate the marginal posterior density at $\boldsymbol{\delta} = \mathbf{0}$, we again derived full conditional distributions for all parameters and used Gibbs sampling to sample from the marginal posterior distributions (code provided at the first author's website). We computed the normal approximation in the same way here as in the previous t test example; the only difference is that in this example we fit a trivariate normal (three means and six covariance parameters) rather than a univariate normal. Computing the CMDE estimate is similar to the t test case. The full conditional posterior distribution of $\boldsymbol{\delta}$ is

$$\boldsymbol{\delta} \mid \mu, \sigma^2, g, \boldsymbol{y} \sim \text{Multivariate Normal}\,(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta) \tag{10}$$

where

$$\boldsymbol{\Sigma}_\delta = \left(\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{g}\boldsymbol{I}_J\right)^{-1}, \tag{11}$$

$$\boldsymbol{\mu}_\delta = \frac{1}{\sigma}\boldsymbol{\Sigma}_\delta \boldsymbol{Z}'(\boldsymbol{y} - \mu\mathbf{1}), \tag{12}$$

and $\boldsymbol{Z}$ is the $\sum_{j=1}^{J} N_j$ by J design matrix mapping the vector of observations $\boldsymbol{y}$ to their respective groups. Proof of this fact is straightforward and is omitted for brevity. The conditional density at $\boldsymbol{\delta} = \mathbf{0}$ is

$$p_0(\boldsymbol{\delta} \mid \sigma^2, g, \boldsymbol{y}) = (2\pi)^{-\frac{J}{2}} \left|\boldsymbol{\Sigma}_\delta\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\mu}\right\}. \tag{13}$$

On every iteration of the chain, we compute the quantity in Eq. (13) given the values of σ^2, g, and μ on that iteration of the chain. The marginal posterior estimate at the restriction is then simply the mean of those values across all iterations, as in Eq. (6).

5.4. Results

To evaluate the quality of estimates from the normal approximation and CMDE methods, we generated data with a variety of sample sizes N, for $J = 3$. Bayes factor error is once again defined as the multiplicative error relative to the quadrature result. Fig. 3A shows the error in the normal approximation for $M = 100$ MCMC iterations. Gray triangles and black circles show the Bayes factor with $N = 16$ and $N = 128$ participants per group, respectively. For all Bayes factors from the normal approximation, the estimates appear to be biased toward the alternative model. The bias does not decrease appreciably when the number of MCMC iterations is high ($M = 10{,}000$ iterations, Fig. 3B). We suspect that this bias in estimation is due to bias in the estimation of the variance and covariance parameters. Because samples from MCMC chains are not independent of one another, estimates of variances and covariances will be biased. Biased estimates of the variance and covariance terms will lead to biased Savage–Dickey estimates from the normal approximation.

Figs. 3C and D show the error in the CMDE Bayes factor estimates for $M = 100$ and $M = 10{,}000$, respectively. The estimates are highly accurate, even with only 100 MCMC iterations. As expected, they are not biased, and have significantly lower variance than the estimates of the Bayes factors obtained by the normal approximation. For the one-way ANOVA Bayes factor, the

Author's personal copy

CMDE yields much higher quality estimates than the multivariate normal approximation.

5.5. Bayes factor vs. p-values

These computations provide an opportunity to compare inference by Bayes factor with inference by p-values from F statistics in one-way designs. Fig. 4 shows the one-way ANOVA Bayes factor as a function of the sample size.[4] Each line represents a different observed effect size η^2, the proportion of variance accounted for by the ANOVA:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}.$$

When $\eta^2 = 0$ the data are highly consistent with the null model, and increasing η^2 gives increasing evidence for the unrestricted model. For these small effect sizes, at small sample sizes the Bayes factor reveals evidence for the null hypothesis, as expected. For $\eta^2 = 0$, the evidence for the null hypothesis is at about 10 to 1 at a sample size of 8 and increases to 100 to 1 at a sample size of 128. For $\eta^2 > 0$, the Bayes factor begins to accumulate evidence for the alternative hypothesis as the sample size increases.

Fig. 4 reveals the miscalibration of p values: p values tend to overstate the evidence against the null model. The inability of p values to consider the distribution of the test statistic under the alternative hypotheses biases them against the null model (Berger & Sellke, 1987; Sellke, Bayarri, & Berger, 2001). The square points in Fig. 4 show combinations of η^2 and N that yield significant p values. Several of these significant p values correspond to equivocal Bayes factors, or even Bayes factors favoring the null model. Although a frequentist would take the significant p value to mean that the null hypothesis can be rejected, the Bayes factor shows that when the reasonable alternative models are considered, the evidence against the null is sometimes marginal at best. The miscalibration of p values will grow with the sample size, due to the Jeffreys–Lindley paradox (Lindley, 1957).

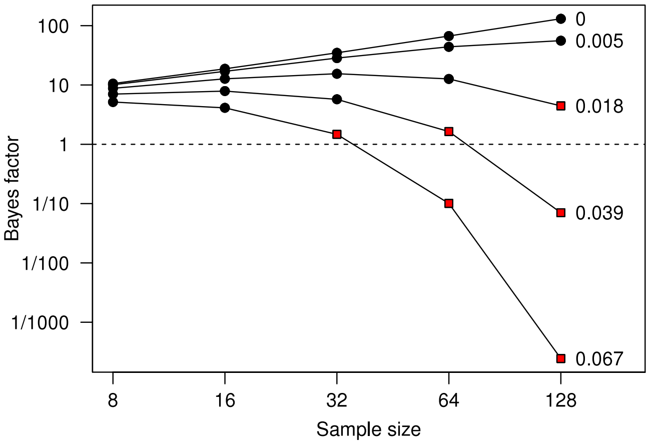


Fig. 4. The one-way ANOVA Bayes factor for $J = 3$ groups as a function of sample size in a group. Each line represents a different observed effect size η^2. Circles show sample sizes and η^2 values that are nonsignificant by an F test at $\alpha < 0.05$; squares show significant tests.

6. Discussion

In the preceding development, we have described the CMDE method for obtaining efficient estimates of posterior densities. The CMDE method is useful in computing the Bayes factor via the Savage–Dickey method in the case where the normalizing constant on the parameter of interest is known. In general, it will be expected to outperform other methods that do not make use of all the information in the MCMC chain.

We especially expect the CMDE to outperform the kernel density estimates, logsplines, and the normal approximation, because the latter methods ignore information. For kernel density estimates and logsplines, information is ignored when most of the posterior density of $\boldsymbol{\theta}$ is far away from the null value $\boldsymbol{\theta}_0$. If samples of $\boldsymbol{\theta}$ are far away from $\boldsymbol{\theta}_0$, we obtain only poor information about the density around $\boldsymbol{\theta}_0$. Every iteration of the chain contains information about the density at $\boldsymbol{\theta}_0$ through the full conditional distribution of $\boldsymbol{\theta}$, which is ignored. Likewise, the normal approximation ignores the fact that the marginal distribution of $\boldsymbol{\theta}$ can be thought of as a mixture of full conditional distributions (Gelfand & Smith, 1990). The CMDE uses the form of this mixture, while the normal approximation makes the unnecessary assumption that the resulting mixture is normal.

There are, however, some situations where CMDE is not preferred. One example is when sampling methods themselves are unnecessary. Consider, for instance, the logic underlying the normal approximation method: the joint posterior distribution will be asymptotically distributed as a multivariate normal, with a mode at the maximum likelihood estimate and covariance matrix that is the inverse of the Fisher information (Bernardo & Smith, 2000). If one is willing to assume normality, a reasonable estimate of the marginal density can be had with no sampling at all, and the CMDE will be unnecessary. Avoiding sampling can speed up inference dramatically, for an often modest penalty in accuracy. A second example is when the marginal likelihood may be analytically integrated across some parameters to reduce the dimensionality of the problem. We have used this approach to construct our benchmarks, and clearly they are superior to any sampling-based approach. In fact, it is often possible to use priors that greatly simplify the direct computation of marginal likelihood; an instructive example is the use of a multivariate Cauchy prior on effects by Liang et al. (2008) for linear models.

There are important extensions of the basic idea of using conditional marginal estimation, discussed next. In the first extension, we use conditional quantities to estimate Bayes factors where the null region is an area rather than a point. In the second extension, we discuss a generalization for the case in which the normalizing constant on the parameter of interest is unknown.

6.1. Extension 1: area null hypotheses

Although the CMDE is employed here to estimate densities at a point, the general strategy of using conditional quantities to estimate marginal quantities extends beyond the Savage–Dickey ratio. Wetzels, Grasman and Wagenmakers (2010) note that the Savage–Dickey density ratio may be thought of as a special case of the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005). Encompassing priors are useful when two hypotheses for comparison are nested within a more general model.

Consider the one-sample t test model defined previously. Morey and Rouder (in press) constructed tests in which under the null, the parameter of interest was restricted to a small equivalence range $A = \{\delta : \delta_l < \delta < \delta_u\}$. The two models to be compared are:

$$\mathcal{M}_n : \delta \in A$$
$$\mathcal{M}_s : \delta \notin A$$
$$\delta \sim \text{Cauchy}(r).$$

[4] Group "data" for Fig. 4 were $(N + 1)$-tiles of the standard normal distribution, with means manipulated to yield a given observed effect size.

Author's personal copy

Because typically A would include 0, $\mathcal{M}_n$ and $\mathcal{M}_s$ are the models where δ is "negligible" and "substantial", respectively. Morey and Rouder used approximations that required no sampling to compute the corresponding Bayes factor.

An alternative computational approach, the encompassing prior approach, uses the fact that the two models are nested within the "encompassing" model $\mathcal{M}_1 : \delta \sim \text{Cauchy}(r)$. Klugkist et al., for instance, suggest using MCMC methods to sample from the posterior distribution of δ in the encompassing model, and computing:

$$\frac{\text{Count of MCMC samples for which } \delta \in A}{\text{Count of MCMC samples for which } \delta \notin A} \approx \frac{p(\delta \in A \mid \boldsymbol{y})}{p(\delta \notin A \mid \boldsymbol{y})}.$$

The ratio of these two counts is approximately the posterior odds of $\delta \in A$ versus $\delta \notin A$. The Bayes factor can be obtained by dividing the posterior odds by the prior odds, as in Eq. (3).

The method of simply counting the regions samples fall in, however, ignores the rich information in the MCMC chain. There is information in every iteration of the MCMC chain about the conditional posterior probability that $\delta \in A$, and this information may be used to estimate the marginal posterior odds. Because the form of the conditional distribution for δ is known (we used it to compute the CMDE), we can compute the conditional posterior probability of $\mathcal{M}_n$:

$$\Pr(\mathcal{M}_n \mid \sigma^2, g, \boldsymbol{y}) = \Phi\left[\frac{\delta_u - \mu_\delta}{\sqrt{\tau_\delta}}\right] - \Phi\left[\frac{\delta_l - \mu_\delta}{\sqrt{\tau_\delta}}\right]$$

where Φ represents the CDF of the standard normal distribution and μ_δ and τ_δ represent the mean and variance of the full conditional distribution of δ from Eq. (7).

The posterior probability of model $\mathcal{M}_n$ is easy to estimate, given samples of σ^2 and g obtained using MCMC techniques. The following estimate is analogous to the CMDE, for the area null:

$$\Pr(\mathcal{M}_n \mid \boldsymbol{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\left(\Phi\left[\frac{\delta_u - \mu_\delta^{(t)}}{\sqrt{\tau_\delta^{(t)}}}\right] - \Phi\left[\frac{\delta_l - \mu_\delta^{(t)}}{\sqrt{\tau_\delta^{(t)}}}\right]\right).$$

From this estimate, the Bayes factor test $\mathcal{M}_n$ versus $\mathcal{M}_s$ is trivial to compute given the prior odds. As a demonstration, we provide WinBUGS code to estimate the Bayes factor in the Appendix. In general, when computing Bayes factors using the encompassing approach, the full conditionals should be used, if available. Similar logic can be used to estimate credible regions, posterior means, posterior variances, and other posterior quantities.

6.2. Extension 2: unknown normalizing constants

Computation of CMDE is contingent on knowing the normalizing constant of the full conditional distribution of the parameter of interest. In many Bayesian analyses, however, these full-conditional posterior normalizing constants are not all known. Hence, it may seem that CMDE is too specialized to be of general interest. However, we note that the CMDE does not require that *all* normalizing constants be known; rather, the CMDE simply requires that a single normalizing constant be known: the normalizing constant for the full conditional on the parameter of interest. The knowledge of the normalizing constant can be assured by making the prior on the parameter of interest conjugate to the likelihood. Conjugate priors are known for many common models (Tanner, 1998), and if a model lacks a conjugate prior on the parameter of interest, then the model may be slightly modified to accommodate the requirements of CMDE.

In cases where this condition cannot be met, Chen (1994) suggests a generalization of the CMDE method that uses an importance sampling method to estimate the marginal density, called importance-weighted marginal density estimation (IWMDE). Chen shows that the CMDE has the lowest asymptotic variance of all IWMDEs, but argues that in some cases the IWMDE would still be preferable, such as if the CMDE is difficult or costly to compute. Further, Chen and Shao (1997) show that the IWMDE outperforms a kernel density estimate in a constrained-parameter linear regression example. Thus, even in cases where the CMDE cannot be used, the basic logic of the CMDE may still be used to obtain efficient estimates of Bayes factors or other desirable posterior quantities.

7. Conclusion

In foregoing examples, we have applied conditional marginal density estimation of Savage–Dickey ratios to compute Bayes factors. We show that this approach is tractable for a one-sample t test and one-way, between-subjects ANOVA. Bayes factors obtained via CMDE are computationally convenient, and are far more accurate than previously recommended kernel density, logspline, or normal approximations methods. In cases where the necessary normalizing constant for computing the CMDE is unknown, generalizations of the CMDE are applicable (Chen, 1994).

CMDE is an application of a more general strategy of using conditional quantities to estimate marginal ones with MCMC. This theme extends broadly to other situations, such as testing null hypotheses that are areas instead of points. In general, whenever a marginal quantity is desired from an MCMC chain, the corresponding conditional quantity should be used to compute it if possible. Otherwise, information in the MCMC chain is ignored, leading to inefficient estimates of the desired quantity.

An R package to perform sampling and compute Bayes factors for the applications in this paper is available from the first author's website at http://drsmorey.org/research/rdmorey.

Appendix. Computing Bayes factors in WinBUGS

One of the benefits of the conditional-marginal approach is that Bayes factors may be naturally computed in software like WinBUGS. In this Appendix, we demonstrate how this can be done using the t test implemented in WinBUGS.

To compute a Bayes factor based on a kernel density or a spline requires extra computation, typically in a separate software package; for example, Wetzels et al. (2009) used R to compute Bayes factors from Gibbs sampler chains obtained from WinBUGS. Using the conditional-marginal approach, the marginal quantities are computed from the conditional quantities. WinBUGS uses Gibbs sampling to compute chains of conditional quantities, making it possible to add extra code to the WinBUGS model to obtain CMDE estimates.

We demonstrate this in the WinBUGS code below. In addition to the standard model declaration, we define variables which enable us to compute the conditional quantities of interest.

```
model
{
  #######
  # For model
  #######

  for(i in 1:N)
  {
    y[i] ~ dnorm(mu, prec);
  }
  mu ~ dnorm(0,invg*prec);
  prec ~ dgamma(precShape, precRate);
  # rsqr = r^2
  invg ~ dgamma(.5, rsqr/2);

  #######
  # For Bayes factors
  #######
```

Author's personal copy

```
  # Conditional precision and mean of delta
  precDelta <- invg + N
  meanDelta <- ybar*sqrt(prec)*N/precDelta

  # log of conditional density of delta at 0
  logPostDensDelta <- -0.5*log(2*pi/precDelta)-
                          0.5*precDelta*pow(meanDelta,2)

  # Divide conditional density by prior density
  # The mean of this chain will estimate the JZS BF
  BFpoint <- exp(logPostDensDelta - logPriorDensDelta)

  # Conditional area between bounds
  postAreaDelta <- phi((deltaUpper - meanDelta)*sqrt(precDelta))-
                   phi((deltaLower-meanDelta)*sqrt(precDelta))

  # Divide the conditional area by prior area
  # The mean of this chain will estimate the encompassing BF
  BFarea <- postAreaDelta/(1-postAreaDelta)/
              (priorAreaDelta/(1-priorAreaDelta))
}
```

The first part of the WinBUGS model declaration above defines the one-sample t test model, with the substitution of precisions for variances, standard for WinBUGS, and gamma prior on $1/\sigma^2$ instead of the Jeffreys prior. Sufficiently small values for `precShape` and `precRate` can be chosen such that the Jeffreys prior is approximated.

The second part of the WinBUGS model declaration contains variables used to compute the Bayes factors corresponding to two hypothesis tests: the point null and the area null. To test the point null hypothesis $\delta = 0$, we may use the result in Eq. (8), computing the CMDE in the WinBUGS code. The first two lines are the conditional posterior precision and mean for δ, which we use to compute the conditional density at 0, `logPostDensDelta`. The logarithm of the prior density at 0 is then subtracted, and the result is exponentiated to yield a density ratio in `BFpoint`. We can also test the area null hypotheses described in the discussion of this paper. The full conditional probability that δ is "negligible" is computed in `postAreaDelta`, which is then used to compute the Bayes factor of the "negligible" model versus the "substantial" model.

Once the model is analyzed in WinBUGS the JZS point null Bayes factor and the area null Bayes factor will be well-estimated by the mean of the `BFpoint` and `BFarea` chains. Moreover, WinBUGS gives an estimate of the Monte Carlo standard error by default, which provides researchers with a measure of the error in the Bayes factor estimate.

References

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. URL: http://www.jstor.org/stable/2289131.

Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian theory*. Chichester, England: John Wiley & Sons.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, *18*, 105–110. doi:10.1214/aoms/1177730497.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association*, *89*(427), 818–824. URL: http://www.jstor.org/stable/2290907.

Chen, M.-H., & Shao, Q.-M. (1997). Performance study of marginal posterior density estimation via Kullback–Leibler divergence. Test 6, 321–350.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321. URL: http://www.jstor.org/stable/2291521.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. URL: http://www.jstor.org/stable/2958475.

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*(1), 214–226. URL: http://www.jstor.org/stable/2239734.

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

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

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144. URL: http://www.jstor.org/stable/170841.

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

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. URL: http://www.jstor.org/stable/2291091.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

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

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

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. URL: http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337.

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

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

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

Morey, R.D., & Rouder, J.N. (in press). Bayes factor approaches for testing interval null hypotheses. Psychological Methods.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*.

R Development Core Team, (2009). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL: http://www.R-project.org.

Raftery, A. E., Satagopan, J. M., Newton, M. A., & Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian statistics 8* (pp. 1–45). Alicante, Spain: Oxford University Press.

Roberts, G. (1996). Markov chain concepts related to sampling algorithms. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

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., 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.

Scott, D. W. (1992). *Multivariate density estimation*. New York: John Wiley & Sons.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *American Statistician*, *55*, 62–71. URL: http://www.jstor.org/stable/2685531.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, *25*(4), 1371–1425. URL: http://www.jstor.org/stable/2959054.

Tanner, M. A. (1998). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. New York: Springer.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618. URL: http://www.jstor.org/stable/2291073.

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

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian ttest. *Psychonomic Bulletin & Review*, *16*, 752–760.

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

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.