Teacher's Corner

A Default Bayesian Hypothesis Test for ANOVA Designs

Ruud WETZELS, Raoul P. P. P. GRASMAN, and Eric-Jan WAGENMAKERS

This article presents a Bayesian hypothesis test for analysis of variance (ANOVA) designs. The test is an application of standard Bayesian methods for variable selection in regression models. We illustrate the effect of various *g*-priors on the ANOVA hypothesis test. The Bayesian test for ANOVA designs is useful for empirical researchers and for students; both groups will get a more acute appreciation of Bayesian inference when they can apply it to practical statistical problems such as ANOVA. We illustrate the use of the test with two examples, and we provide R code that makes the test easy to use.

KEY WORDS: Bayes factor; Model selection; Teaching Bayesian statistics.

1. INTRODUCTION

Bayesian methods have become increasingly popular in almost all scientific disciplines (e.g., Poirier 2006). One important reason for this gain in popularity is the ease with which Bayesian methods can be applied to relatively complex problems involving, for instance, hierarchical modeling or the comparison between nonnested models. However, Bayesian methods can also be applied in simpler statistical scenarios such as those that feature basic testing procedures. Prominent examples of such procedures include analysis of variance (ANOVA) and the *t*-test; these tests are the cornerstone of data analysis in fields such as biology, economics, sociology, and psychology.

Because Bayesian methods have become more mainstream in recent years, most technically oriented studies now offer at least one course on Bayesian inference in their graduate or undergraduate program. Our own experience in teaching one such course is that students often ask the same questions when Bayesian model selection and hypothesis testing are introduced. First, students are interested to know how they can apply Bayesian methods to testing problems that they face on a regular basis; second, students want to know how prior distributions can be chosen such that a test can be considered default. In this article we address both questions. We apply the Bayesian method to ANOVA designs and explain the rationale and impact of several default prior distributions.

Thus, the first goal of this article is to show how the Bayesian framework of hypothesis testing with the Bayes factor can be carried out in ANOVA designs. ANOVA is one of the most popular statistical methods to assess whether or not two or more population means are equal—in most experimental settings, ANOVA is used to test for the presence of a treatment effect. Because of its importance and simplicity, ANOVA is taught in virtually every applied statistics course. Nevertheless, the Bayesian hypothesis testing literature on ANOVA is scant; the dominant treatment of ANOVA is still classical or frequentist (e.g., Draper and Smith 1998; Faraway 2002) and, although the Bayesian treatment of ANOVA is gaining popularity (e.g., Gelman et al. 2004; Qian and Shen 2007; Ntzoufras 2009; Kaufman and Sain 2010), the latter has dealt almost exclusively with estimation, not testing (for exceptions, see Westfall and Gönen 1996; Sen and Churchill 2001; Ishwaran and Rao 2003; Ball 2005; Gönen et al. 2005; Maruyama 2009). This is all the more surprising because Bayesian hypothesis testing has been well developed for variable selection in regression models (e.g., Liang et al. 2008), of which ANOVA is a special case.

The second goal of this article is to describe the rationale behind a particular family of default priors—*g*-priors—and to use these *g*-priors for default Bayesian tests for ANOVA designs. We hope this work shows students and experimental researchers how Bayesian hypothesis tests can be a valid and practical alternative to classical or frequentist tests.

The outline of this article is as follows. In the first section we briefly cover Bayesian estimation and Bayesian model selection. In the second section we describe the various *g*-priors that have been proposed in the literature on variable selection in regression models. Finally, we present two worked examples that show how the regression framework can be applied to one-way and two-way ANOVA designs.

2. BAYESIAN INFERENCE

2.1 Bayesian Estimation

In Bayesian estimation (e.g., Bernardo and Smith 1994; Lindley 2000; O'Hagan and Forster 2004), uncertainty about parameters is quantified by probability distributions. Suppose,

Ruud Wetzels is Applied Statistician, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam (E-mail: *wetzels.ruud@gmail.com*). Raoul P. P. P. Grasman is Applied Statistician, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam (E-mail: *R.P.P.P.Grasman@uva.nl*). Eric-Jan Wagenmakers is Cognitive Scientist, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam (E-mail: *ej.wagenmakers@gmail.com*). This research was supported by Veni and Vidi grants from the Netherlands Organisation for Scientific Research (NWO).

we have a model $\mathcal{M}$ and we wish to estimate the model parameters $\boldsymbol{\theta}$. Then, we have to define a *prior distribution* over these parameters; $p(\boldsymbol{\theta}|\mathcal{M})$. When data $\boldsymbol{Y}$ come in, this prior distribution $p(\boldsymbol{\theta}|\mathcal{M})$ is updated to yield the *posterior distribution* $p(\boldsymbol{\theta}|\boldsymbol{Y}, \mathcal{M})$ according to Bayes' rule:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{Y}, \mathcal{M}) &= \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{Y}|\mathcal{M})} \\
&= \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{\int_{\Theta} p(\boldsymbol{Y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}} \\
&\propto p(\boldsymbol{Y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}).
\end{aligned}
$$

Hence, the posterior distribution of $\boldsymbol{\theta}$ is proportional to the likelihood times the prior. In Bayesian parameter estimation, the researcher is interested in the posterior distribution of the model parameters $p(\boldsymbol{\theta}|\boldsymbol{Y}, \mathcal{M})$. However, in Bayesian model selection the focus is on $p(\boldsymbol{Y}|\mathcal{M})$, the marginal likelihood of the data under model $\mathcal{M}$.

2.2 Bayesian Model Selection

In Bayesian model selection, competing statistical models or hypotheses are assigned prior probabilities. Consider two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$ with prior probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$.

After observing the data, the relative plausibility of $\mathcal{M}_1$ and $\mathcal{M}_2$ is given by the ratio of posterior model probabilities, that is, the posterior odds:

$$
\frac{p(\mathcal{M}_1|\boldsymbol{Y})}{p(\mathcal{M}_2|\boldsymbol{Y})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}\frac{p(\boldsymbol{Y}|\mathcal{M}_1)}{p(\boldsymbol{Y}|\mathcal{M}_2)}.
$$

Hence, the posterior odds are given by the product of the prior odds and the ratio of marginal likelihoods. The latter component is known as the *Bayes factor* (Jeffreys 1961; Dickey 1971; Berger and Sellke 1987; Kass and Raftery 1995) and quantifies the change from prior to posterior odds; therefore, the Bayes factor does not depend on the prior model probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ and quantifies the evidence that the data provide for $\mathcal{M}_1$ versus $\mathcal{M}_2$.

In linear regression and ANOVA, two models of special interest are the null model, $\mathcal{M}_N$, that does not include any of the predictors (but does include the intercept) and the full model, $\mathcal{M}_F$, that includes all relevant predictors. In this scenario, the main difficulty with the Bayes factor is its sensitivity to the prior distribution for the model parameters under test (Press et al. 2003; Berger 2006; Gelman 2008).

When there is limited knowledge about the phenomenon under study, the prior distribution for the parameters should be relatively uninformative. However, to avoid paradoxical results, the prior distribution cannot be *too* uninformative. In particular, the Jeffreys–Lindley–Bartlett paradox (Bartlett 1957; Jeffreys 1961; Lindley 1980; Shafer 1982; Berger and Delampady 1987; Robert 1993) shows that with vague uninformative priors on the parameters under test, the Bayes factor will strongly support the null model. The reason is that the marginal likelihood $p(\boldsymbol{Y}|\mathcal{M})$ is obtained by averaging the likelihood over the prior; when the prior is very spread out relative to the data, a large part of the prior distribution is associated with very low likelihoods,

decreasing the average. This paradox is illustrated in Figure 3 and will be discussed later in the context of a specific model. The next section details how, in the context of linear regression and ANOVA, one can avoid the Jeffreys–Lindley–Bartlett paradox and nevertheless define prior distributions that are reasonably uninformative.

3. LINEAR REGRESSION, ANOVA, AND THE SPECIFICATION OF *g*-PRIORS

The prior distributions that we will discuss are applicable to model selection in the regression framework. Assume a response vector $\boldsymbol{Y}$ of length n, $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, precision ϕ, and $\boldsymbol{I}_n$ an $n \times n$ identity matrix,

$$
\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_n/\phi).
$$

The mean $\boldsymbol{\mu}$ can be decomposed into an overall common intercept α and the regression coefficients $\boldsymbol{\beta}$. The mean $\boldsymbol{\mu}$ then becomes

$$
\boldsymbol{\mu} = \mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta},
$$

where $\boldsymbol{X}$ represents the $n \times k$ design matrix and $\boldsymbol{\beta}$ is the k-dimensional vector of regression coefficients.

In the ANOVA setting, the independent variables that are controlled in the experiment are called *factors*, which in turn can have different levels of intensity. Then, the regression coefficients are interpreted as level-specific parameters. The design matrix $\boldsymbol{X}$ is constructed using dummy coding (Draper and Smith 1998). Because the matrix $[\mathbf{1}_n, \boldsymbol{X}]$ does not necessarily have full column rank, we need to add a constraint. Here, we adopt the sum-to-zero constraint. By using this constraint, the intercept is the grand mean, and each regression coefficient describes the deviation from this grand mean—consequently, the regression coefficient of the last level equals minus the sum of the other regression coefficients.

In the one-way ANOVA, we examine the effect of a categorical variable $\boldsymbol{X}$ on the continuous response variable $\boldsymbol{Y}$. The null hypothesis is defined as H_0; all levels have the same mean, and the alternative hypothesis is defined as H_1; at least one of the levels has a different mean.

We can translate this frequentist test to a Bayesian model selection situation by comparing the model with all relevant regression coefficients to the model without these coefficients. In the remainder of this article we focus on the one-way and two-way ANOVA and show how these tests can be carried out in a Bayesian fashion.

The sections below list three default prior distributions. We focus on prior distributions for variable selection in regression as this framework provides the basis for the analysis of ANOVA designs (for more information on Bayesian variable selection see Leamer 1978; Zellner 1986, 1987; Mitchell and Beauchamp 1988; Chipman 1996; George and McCulloch 1997). The following subsections detail, in historical order, three versions of the popular *g*-prior.

3.1 Zellner's g-Prior

In the case of linear regression, Zellner's g-prior (Zellner 1986) corresponds to

$$p(\boldsymbol{\beta}|\phi, g, \boldsymbol{X}) \sim N\left(0, \frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) \quad g > 0,$$

with Jeffreys' prior (Jeffreys 1961) on the precision

$$p(\phi) \propto \frac{1}{\phi},$$

and a flat prior on the common intercept α. Note that we assume that the columns of $\boldsymbol{X}$ are centered so that $\mathbf{1}_n^T\boldsymbol{X} = 0$.

This set of prior distributions is of the conjugate Normal-Gamma family, and therefore the marginal likelihood can be calculated analytically. When the design matrix is considered fixed, we are allowed to use it in our prior variance term as $\frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Recall that the variance of the maximum likelihood estimator for $\boldsymbol{\beta}$, var($\hat{\boldsymbol{\beta}}$), equals $\phi^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence, the term g is a scaling factor for the prior: if we choose g to be 1, we give the prior the same weight as the sample; if we choose g to be 2, the prior is half as important as the sample; if we choose g to be n, the prior is $1/n$th as important as the sample.

An obvious problem with this prior distribution is how to set parameter g. If g is set low, then the prior distribution for $\boldsymbol{\beta}$ is relatively peaked and informative. If g is set high then this prior is relatively spread out and uninformative. However, as described in the previous section, a prior that is too vague can result in the Jeffreys–Lindley–Bartlett paradox.

Various settings for g have been studied and proposed. A popular setting is $g = n$, corresponding to the so-called "unit information prior." The intuition is that this prior contains as much information as present in a single observation (Kass and Wasserman 1995); the argument is that the precision of the sample estimate of $\boldsymbol{\beta}$ contains the information of n observations. Then the amount of information in an imaginary single observation is this quantity divided by n, hence $g = n$. Another well-known choice of g is to set it equal to the square of the number of predictors of the regression model: $g = k^2$ (i.e., the Risk Inflation Criterion, Foster and George 1994). Furthermore, Fernandez, Ley, and Steel (2001) suggested to take $g = \max\{n, k^2\}$ as a "benchmark prior."

A quantity of interest is the so-called shrinkage factor $g/(g+1)$. It can be used to estimate the posterior mean of $\boldsymbol{\beta}$, which is the least squares estimate of $\boldsymbol{\beta}$ multiplied by the shrinkage factor

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}, g] = \frac{g}{g+1}\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$. A low value of g pulls the posterior mean of $\boldsymbol{\beta}$ to zero, whereas a high value of g yields results similar to the least squares estimate. Note that, somewhat confusingly, a low shrinkage factor means more shrinkage and vice versa.

To compute the Bayes factor in the one-way ANOVA design, we compare the full model, $\mathcal{M}_F$ to the null model, $\mathcal{M}_N$. Then, the Bayes factor is given by

$$\mathrm{BF}[\mathcal{M}_F : \mathcal{M}_N] = (1+g)^{(n-k-1)/2}[1+g(1-R^2)]^{-(n-1)/2},$$

(1)

where k equals the number of predictors of $\mathcal{M}_F$, n is the sample size, and R^2 the coefficient of determination of $\mathcal{M}_F$ (note that R^2 for $\mathcal{M}_N$ equals zero as it contains no predictors).

Equation (1) shows that, in its general formulation, Zellner's g-prior is potentially vulnerable to the Jeffreys–Lindley–Bartlett paradox: when $g \to \infty$ with n and k fixed, the Bayes factor, $\mathrm{BF}[\mathcal{M}_F : \mathcal{M}_N]$, will go to 0, favoring the null model regardless of the observed data (see Figure 3 for an example).

Another problem with the Zellner g-prior is that, when the evidence in favor of the full model goes to infinity (i.e., R^2 goes to 1), the Bayes factor converges to the upper bound $(1 + g)^{(n-k-1)/2}$. Liang et al. (2008) termed this undesirable property the "information paradox."

3.2 Jeffreys–Zellner–Siow (ZJS) Prior

To test whether a parameter μ is zero or nonzero (with μ the mean of a normal distribution), Jeffreys (1961, pp. 268–270) suggested to apply a Cauchy prior. The Cauchy prior was the simplest distribution to satisfy consistency requirements that Jeffreys considered important for hypothesis testing. One such requirement is that a researcher does not want to favor one model over another on the basis of a single datum.

Extending Jeffreys' suggestion to variable selection in the regression model, Zellner and Siow (1980) proposed a multivariate Cauchy prior on the regression coefficients and a flat prior on the common intercept. However, as the marginal likelihood is not analytically tractable, this approach did not gain much popularity.

Recently, however, Liang et al. (2008) represented the JZS prior as a mixture of g-priors, that is, an Inverse-Gamma(1/2, n/2) prior on g and Jeffreys' prior on the precision ϕ:

$$p(\phi) \propto \frac{1}{\phi}$$

$$p(\boldsymbol{\beta}|\phi, g, \boldsymbol{X}) \propto \int N\left(0, \frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) p(g)dg$$

$$p(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2}e^{-n/(2g)}.$$

This formulation combines the computational advantages of the g-prior with the statistical advantages of the Cauchy prior. Note that again we assume that the columns of $\boldsymbol{X}$ are centered.

By assigning a prior to g, we avoid having to assign g a specific value; moreover, the prior on g allows us to estimate g from the data and obtain data-dependent shrinkage. Equation (2) gives the expected value of the shrinkage factor $g/(g + 1)$ with the JZS approach:

$$\mathbb{E}\left[\frac{g}{g+1}|\boldsymbol{Y}, \mathcal{M}\right]$$
$$= \left\{\int_0^\infty (1+g)^{(n-k-3)/2}[1+g(1-R^2)]^{-(n-1)/2}\right.$$
$$\left.\times g^{(-1/2)}e^{-n/(2g)}dg\right\} \Big/ \left\{\int_0^\infty (1+g)^{(n-k-1)/2}\right.$$
$$\left.\times [1+g(1-R^2)]^{-(n-1)/2}g^{-3/2}e^{-n/(2g)}dg\right\}.$$

(2)

It can be seen from Equation (2), and later from Equation (4), that the expected value of $g/(g + 1)$ increases with R^2 (Zeugner and Feldkircher 2009). Hence, there is less shrinkage when more variance is explained by the model.

In the JZS approach, the Bayes factor comparing the full model to the null model is:

$$
\mathrm{BF}[\mathcal{M}_F : \mathcal{M}_N] = \frac{(n/2)^{1/2}}{\Gamma(1/2)}
$$
$$
\times \int_0^\infty (1+g)^{(n-k-1)/2}[1+g(1-R^2)]^{-(n-1)/2}g^{-3/2}e^{-n/(2g)}dg.
$$
(3)

As pointed out by Liang et al. (2008), the integral is one-dimensional and easily approximated using standard software packages such as R (R Development Core Team 2004).

A drawback of the JZS prior is that the Bayes factor is not analytically available. However, the JZS prior is not vulnerable to the Jeffreys–Lindley–Bartlett paradox nor to the information paradox (Liang et al. 2008).

3.3 Hyper-g Priors

As an alternative to the JZS prior, Liang et al. (2008) proposed a family of prior distributions on g and termed this the hyper-g approach:

$$
p(g) = \frac{a-2}{2}(1+g)^{-a/2} \quad g > 0,
$$

which is a proper distribution if $a > 2$ (Strawderman 1971; Cui and George 2008). Because this distribution leads to indeterminate Bayes factors when $a \le 2$, Liang et al. (2008) studied the behavior of this prior for $2 < a \le 4$. Interestingly, this family of priors on g corresponds to the following prior on the shrinkage factor $g/(1 + g)$:

$$
\frac{g}{1+g} \sim \mathrm{Beta}\left(1, \frac{a}{2} - 1\right).
$$

By choosing a, one can tune the prior on the shrinkage factor. When $a = 4$, the prior is uniform between 0 and 1, whereas when a is very close to 2, the prior distribution for the shrinkage factor will have most mass near 1. Figure 1 shows the effect of various a on the prior distribution for the shrinkage factor $g/(g + 1)$. Furthermore, Dellaportas, Forster, and Ntzoufras (in press) showed that the posterior densities of the parameters are, in terms of posterior shrinkage, insensitive to the choice of a within the recommended range. Only for very high values of a (in their simple linear regression example, $a \approx 20$) was posterior shrinkage considerable.

The expected value of the shrinkage factor $g/(g + 1)$ with the hyper-g approach is

$$
\mathbb{E}\left[\frac{g}{g+1}\middle| \boldsymbol{Y}, \mathcal{M}\right]
$$
$$
= \frac{2}{k+a}\frac{{}_2F_1[(n-1)/2, 2; (k+a)/2+1; R^2]}{{}_2F_1[(n-1)/2, 1; (k+a)/2; R^2]}, \quad (4)
$$

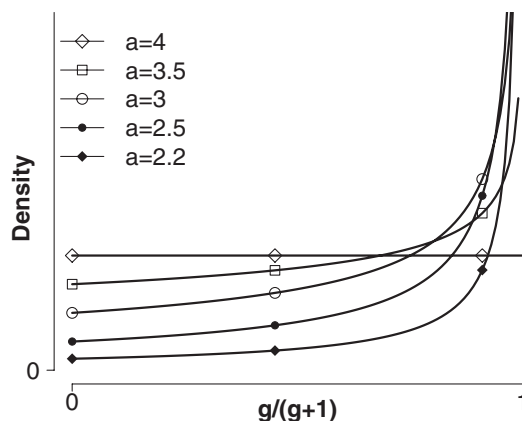where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function (Abramowitz and Stegun 1972)

Figure 1. Effect of parameter a on the shrinkage factor $g/(g + 1)$. When $a = 4$, the prior is uniform between 0 and 1, whereas when a is very close to 2, the prior distribution for the shrinkage factor has most mass near 1. Higher values for $g/(g + 1)$ result in less shrinkage.

$$
{}_2F_1(a, b; c; z)
$$
$$
= \frac{\Gamma(c)}{\Gamma(c-b)\Gamma(b)}\int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}dt \quad c > b > 0.
$$

Just as with the JZS prior, the hyper-g approach estimates g and allows for data-dependent shrinkage.

In order to compare the two models that are important in the one-way ANOVA design, we need to calculate the Bayes factor (note that, this Bayes factor is also available in closed form using the Gaussian hypergeometric function):

$$
\mathrm{BF}[\mathcal{M}_F : \mathcal{M}_N] = \frac{a-2}{2}\int_0^\infty (1+g)^{(n-k-1-a)/2}
$$
$$
\times [1+g(1-R^2)]^{-(n-1)/2}dg. \quad (5)
$$

Just as with the JZS prior, the hyper-g approach is not vulnerable to the Jeffreys–Lindley–Bartlett paradox, nor to the information paradox (when $a \le n - k + 1$, Liang et al. 2008).

4. A BAYESIAN ONE-WAY ANOVA

To illustrate the differences between the various priors and the effects they have on the Bayes factor for ANOVA designs, we first discuss the one-way ANOVA. We follow Box and Tiao (1973) and use example data from an experiment that was set up to investigate to what extent yield of dyestuff differs between batches. The experiment featured six batches with five observations each. Figure 2 shows the box and whisker plot of yield of dyestuff for the different batches. The left plot shows the original data from Box and Tiao (1973). To illustrate the behavior of the Bayes factor when the null hypothesis is true, the right plot shows the same data but with equal means (i.e., the difference between the batch mean and the overall mean was subtracted from the batch data).

First, we carried out a classical one-way ANOVA to compute the F statistic and the corresponding p-value for both datasets. For the original dataset, we compute $F(5, 24) = 4.60$, $p = 0.004$, suggesting that at least one of the batches has a different yield. In the modified dataset with equal means, we compute $F(5, 24) = 0$, $p = 1$, suggesting that the yield of dyestuff is
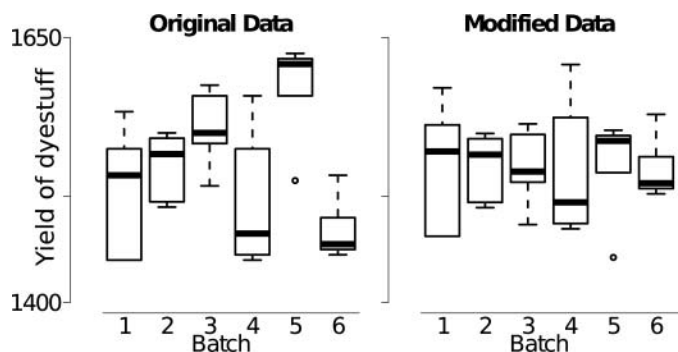
Figure 2. Boxplots of yield of dyestuff per batch. The left plot (original data) shows the original data from Box and Tiao (1973). The right plot (modified data) shows the same data but with the difference between the batch mean and the overall mean subtracted from the batch data.
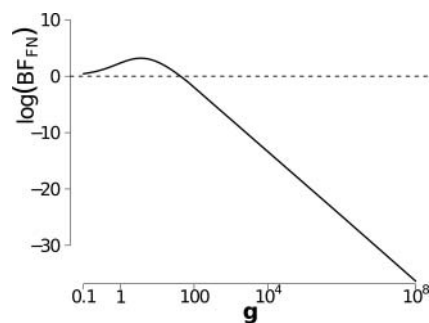

Figure 3. An illustration of the Jeffreys–Lindley–Bartlett paradox when the Zellner g-prior is applied to the dyestuff data from Box and Tiao (1973). When g increases from 1 to 4, the Bayes factor in favor of the full model increases as well. By increasing g much further, the Bayes factor can be made arbitrarily close to 0, signifying infinite support for the null model.

equal for all batches, although such an inference in favor of the null hypothesis is not warranted in the Fisherian framework of p-value significance testing.

Next, we designed a Bayesian hypothesis test to contrast two models. The full model, $\mathcal{M}_F$, contains a grand mean α and the predictors for batches 1–5. The predictor for batch 6 is omitted because of the sum-to-zero constraint. The null model, $\mathcal{M}_N$, contains no predictors. Therefore, our test concerns the following two models:

$$\mathcal{M}_F : \boldsymbol{\mu} = \mathbf{1}_n\alpha + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{X}_3\beta_3 + \boldsymbol{X}_4\beta_4 + \boldsymbol{X}_5\beta_5$$
$$\mathcal{M}_N : \boldsymbol{\mu} = \mathbf{1}_n\alpha.$$

The results from the Bayesian hypothesis test for the data with unequal group means, reported in Table 1, show that the two Zellner g-priors and the JZS prior yield only modest Bayes factor support in favor of $\mathcal{M}_F$; the two hyper-g priors yield more convincing support in favor of $\mathcal{M}_F$: overall, the results suggest that the data may be too sparse to allow an unambiguous conclusion. Importantly, a Bayes factor of 3 arguably does not inspire as much confidence as one would glean from a p-value as low as 0.004 (Berger and Sellke 1987). This result highlights the general conflict between Bayes factors and p-values in terms of their evidential impact (e.g., Edwards, Lindman, and Savage 1963; Sellke, Bayarri, and Berger 2001).

When the models are compared using the modified data, Table 1 shows that the two Zellner g-priors and the JZS prior yield considerable Bayes factor support in favor of the null

Table 1. Bayes factors and shrinkage factors for the one-way ANOVA example on the dyestuff data, see Figure 2. The Bayes factor compares the full model to the null model, testing for a main effect of batch.

| | Unequal means | | Equal means | |
Prior	$\mathrm{BF}_{F:N}$	$\mathbb{E}[g/(g+1)|\boldsymbol{Y}]$	$\mathrm{BF}_{F:N}$	$\mathbb{E}[g/(g+1)|\boldsymbol{Y}]$
Zellner $g = n$	2.0	0.97	1.87×10^{-4}	0.97
Zellner $g = k^2$	2.9	0.96	2.90×10^{-4}	0.96
JZS	3.1	0.90	8.51×10^{-4}	0.86
Hyper-g $a = 3$	9.9	0.71	0.17	0.25
Hyper-g $a = 4$	10.1	0.65	0.29	0.22

model $\mathcal{M}_N$; the two hyper-g priors also provide evidence in favor of $\mathcal{M}_N$, albeit less extreme. Moreover, the relation between R^2 and the shrinkage factor now becomes clear: for each prior where g is estimated (i.e., JZS, hyper-g with $a = 3$, and hyper-g with $a = 4$), the shrinkage factor is lower when the null model is preferred, as is the case for the modified data.

Finally, we use the original dyestuff data with unequal means to illustrate the Jeffreys–Lindley–Bartlett paradox for the one-way ANOVA model. Under Zellner's g-prior with $g = n$ or $g = k^2$, the Bayes factor was in favor of the full model. However, Figure 3 shows that by increasing g the Bayes factor can be made arbitrarily close to 0, indicating impressive evidence in favor of the null model.

5. A BAYESIAN TWO-WAY ANOVA

To illustrate the Bayesian two-way ANOVA, we use a slightly more complex example from Faraway (2002). As part of an investigation of toxic agents, a total of 48 rats were allocated to three poisons (I, II, III) and four treatments (A, B, C, D). The dependent variable is the reciprocal of the survival time in tens of hours, which can be interpreted as the rate of dying. Figure 4 shows the box-and-whisker plot of the survival times in the different experimental conditions.
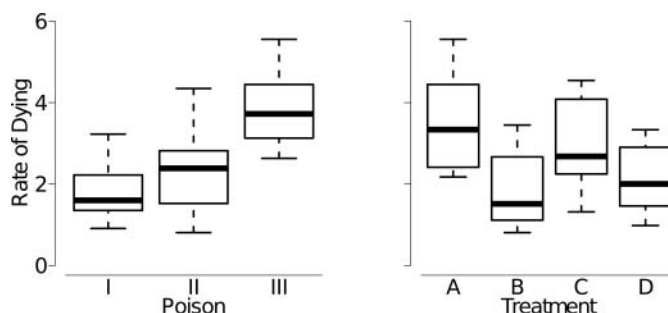

Figure 4. Rate of dying per poison and per treatment. Poison group I (the reference level for poison) has a mean of 1.80; the means of groups II and III are 0.47 and 2.00 higher, respectively. Treatment group A (the reference level for treatment) has a mean of 3.52; the means of groups B, C, and D are 1.66, 0.57, and 1.36 lower, respectively.

First, we carried out a classical two-way ANOVA to compute the F statistics and the corresponding p-values. First, we investigate whether the interaction terms should be incorporated in the model. We compute $F(6, 36) = 1.1$, $p \approx 0.39$, suggesting that poison and treatment do not interact, although, again, such inference in favor of the null hypothesis is not warranted in the Fisherian framework of p-value significance testing.

Because the interaction effects were not significant, we remove them from the model. Then, for the main effect of treatment, we compute $F(3, 42) = 27.9$, $p < 0.001$, suggesting that at least one of the treatments has an effect on rate of dying. For the main effect of poison we compute $F(2, 42) = 71.7$, $p < 0.001$, suggesting that at least one of the poisons has an effect on rate of dying.

Again, we compare the classical results to the Bayesian alternatives. We define the necessary models that are needed to test for each of the main effects and for the interaction effects. To test for the effect of the interaction terms we define two models: the full model containing the main and interaction effects $\mathcal{M}_{PT}$, and the same model without the interaction effects $\mathcal{M}_{P+T}$. To test for the main effects we define the no-interaction model with the effects of treatment $\mathcal{M}_T$; the no-interaction model with the effects of poison $\mathcal{M}_P$; and the null model $\mathcal{M}_N$.

$$
\begin{aligned}
\mathcal{M}_{PT} : \boldsymbol{\mu} = {} & \mathbf{1}_n \alpha + X_I \beta_I + X_{II} \beta_{II} + X_A \beta_A + X_B \beta_B + X_C \beta_C \\
& + X_{I \times A} \beta_{I \times A} + X_{I \times B} \beta_{I \times B} + X_{I \times C} \beta_{I \times C} \\
& + X_{II \times A} \beta_{II \times A} + X_{II \times B} \beta_{II \times B} + X_{II \times C} \beta_{II \times C}, \\
\mathcal{M}_{P+T} : \boldsymbol{\mu} = {} & \mathbf{1}_n \alpha + X_I \beta_I + X_{II} \beta_{II} + X_A \beta_A + X_B \beta_B + X_C \beta_C, \\
\mathcal{M}_T : \boldsymbol{\mu} = {} & \mathbf{1}_n \alpha + X_A \beta_A + X_B \beta_B + X_C \beta_C, \\
\mathcal{M}_P : \boldsymbol{\mu} = {} & \mathbf{1}_n \alpha + X_I \beta_I + X_{II} \beta_{II}, \\
\mathcal{M}_N : \boldsymbol{\mu} = {} & \mathbf{1}_n \alpha.
\end{aligned}
$$

We compare the reduced models to the larger model to test for the effect of the predictors that were left out. If the larger model is preferred over the reduced model then the tested effects matter. However, these models cannot be compared directly using the methods outlined above, as these methods always feature the null model. Instead, we first calculate the Bayes factor comparing the larger model $\mathcal{M}_L$ to the null model, $\mathrm{BF}[\mathcal{M}_L : \mathcal{M}_N]$, and the reduced model $\mathcal{M}_R$ to the null model, $\mathrm{BF}[\mathcal{M}_R : \mathcal{M}_N]$. The desired Bayes factor, $\mathrm{BF}[\mathcal{M}_L : \mathcal{M}_R]$, is then obtained by taking the ratio of Bayes factors

$$
\mathrm{BF}[\mathcal{M}_L : \mathcal{M}_R] = \frac{\mathrm{BF}[\mathcal{M}_L : \mathcal{M}_N]}{\mathrm{BF}[\mathcal{M}_R : \mathcal{M}_N]}.
$$

We do not present the shrinkage factors because the model comparison is not between the null model and the full model but between two models with many predictors each.

A test for the interaction involves the comparison between $\mathcal{M}_{PT}$ and $\mathcal{M}_{P+T}$. Table 2 shows the results for the Bayes factors that test for the presence of the interaction terms. The different priors do not change the overall conclusion: all priors support the model without the interaction terms. Hence, we drop the interaction terms from the ANOVA model and proceed with the main effects only.

By comparing $\mathcal{M}_{P+T}$ to $\mathcal{M}_P$, we can test for a main effect of treatment. Table 2 shows that all Bayesian hypothesis tests

Table 2. Bayes factors for the two-way ANOVA for the rats dataset from Faraway (2002) plotted in Figure 4. The Bayes factor compares the relevant models to each other to test for main effects of poison and treatment, and their interaction.

Prior	$\mathrm{BF}_{PT:P+T}$	$\mathrm{BF}_{P+T:P}$	$\mathrm{BF}_{P+T:T}$
Zellner $g = n$	2.61×10^{-04}	6.87×10^{07}	3.09×10^{12}
Zellner $g = k^2$	1.45×10^{-05}	3.41×10^{08}	4.36×10^{11}
JZS	5.37×10^{-04}	4.52×10^{07}	1.24×10^{12}
Hyper-g $a = 3$	9.41×10^{-04}	2.95×10^{07}	1.81×10^{11}
Hyper-g $a = 4$	1.34×10^{-03}	2.07×10^{07}	6.72×10^{10}

favor the model that includes the treatment effect, regardless of the specific choice of prior distribution.

By comparing $\mathcal{M}_{P+T}$ to $\mathcal{M}_T$ we can test for a main effect of poison. The Bayesian hypothesis tests show that all methods favor the full model over the null model, regardless of the specific choice of prior distribution (see Table 2). The support for the model with a main effect of poison is considerably higher then the support for the main effect for treatment.

6. CONCLUSION

ANOVA is one of the most often-used statistical methods in the empirical sciences. However, Bayesian hypothesis tests are rarely conducted in ANOVA designs; instead, most theoretical development has concerned the more general problem of selecting variables in regression models (e.g., Mitchell and Beauchamp 1988; George and McCulloch 1997; Kuo and Mallick 1998; Casella and Moreno 2006; O'Hara and Sillanpää 2009). Here we showed how the regression framework can be seamlessly carried over to ANOVA designs, at the same time illustrating various default prior distributions, such as Zellner's g-prior, the JZS approach, and the hyper-g approach (for a similar approach see Bayarri and García-Donato 2007).

Of course, other Bayesian model specifications for ANOVA are possible; ours has the advantage that it follows directly from the regression approach that has been studied in detail. A further didactical advantage is that many students are already familiar with linear regression and the extension to ANOVA is conceptually straightforward. In addition, software programs implemented in R make it easy for students and teachers to apply Bayesian regression and ANOVA to inference problems of practical interest; in addition, this software allows users to compare the substantive Bayesian conclusions to those drawn from the classical p-value approach. In general, the software implementation of the theoretical framework provides students with the opportunity of considerable hands-on experience with Bayesian hypothesis testing, something that is likely to increase not only their understanding, but also their motivation to learn.

We feel it is important for students to realize that there is likely no single correct prior distribution; in fact, it can be informative to use different priors in a sensitivity analysis. If different plausible prior distributions lead to different substantive conclusions it is best to acknowledge that the data are ambiguous.

Although not the focus of this article, post hoc comparisons can easily be accommodated within the present framework. For instance, one might be interested in testing which group

mean is different from the reference category mean. Then it is straightforward to calculate a Bayes factor to compare those means, using a procedure resembling a Bayesian *t*-test (Gönen et al. 2005). Another possibility is to apply model averaging and calculate an inclusion probability for each predictor over all possible models (Clyde 1999; Hoeting et al. 1999).

Note that, although the Bayes factor already has a dimension penalty built in—sometimes called the Bayesian Ockham's razor (Berger and Jefferys 1992)—this is not a penalty against multiple comparisons. To correct for multiple comparisons, the prior on the model itself must be chosen appropriately (see Stephens and Balding 2009; Scott and Berger 2010, and references therein).

In sum, we have outlined a default Bayesian hypothesis test for ANOVA designs by a direct and simple extension of the framework for variable selection in regression models. In the course of doing so we have discussed three of the most popular default priors. We hope that empirical researchers and students can better appreciate and understand Bayesian hypothesis testing when they see how it can be applied to practical research problems for which ANOVA is often the method of choice.

APPENDIX: R FUNCTIONS TO COMPUTE THE ANOVA BAYES FACTOR

```
###
# For all functions:
# y is the response vector
# x is the design matrix
# R-scripts with the ANOVA examples can be
  found at
# www.ruudwetzels.com
###

## (1) Function to compute the Bayes Factor
## with Zellner's g-prior prior

zellner.g = function(y, x, g){
output = matrix(,1,2)
colnames(output) = c('BF_10','g/(g+1)')
n = length(y)
r2 = summary( lm(y ~ x) )$r.squared
k = dim(x)[2] - 1
output[1] = ( 1+g )^( (n-k-1)/2 )
  *( 1+g*(1-r2) )^(-(n-1)/2)
output[2] = g / (g+1)
return(output)}

## (2) Function to compute the Bayes Factor
## with Jeffreys-Zellner-Siow prior

zellnersiow = function(y, x){
output = matrix(,1,2)
colnames(output) = c('BF_10','g/(g+1)')
n = length(y)
r2 = summary( lm(y ~ x) )$r.squared
k = dim(x)[2] - 1

BF.integral = function(g, n = n, k = k,
  r2 = r2){
```

```
(1+g)^((n-k-1)/2)*(1+g*(1-r2))^(-(n-1)/2)
  *g^(-3/2)*exp(-n/(2*g))}
output[1] = ((n/2)^(1/2)/gamma(1/2))
  *integrate(BF.integral,0,Inf,n=n,k=k,r2=r2)
  $value

shrinkage.integral=function(g,n=n,k=k,
  r2=r2){
(1+g)^((n-k-1-2)/2)*(1+g*(1-r2))^(-(n-1)/2)
  *g^(1-3/2)*exp(-n/(2*g))}
g.=integrate(shrinkage.integral,0,Inf,n=n,
  k=k,r2=r2)$value

output[2] = g. / integrate(BF.integral,0,
Inf,n=n,k=k,r2=r2)$value
return(output)}

## (3) Function to compute the Bayes Factor
## with Liang et al. hyper-g prior

hyper.g = function(y, x, a){
output = matrix(,1,2)
colnames(output) = c('BF_10','g/(g+1)')
n = length(y)
r2 = summary( lm(y ~ x) )$r.squared
k = dim(x)[2]-1

BF.integral=function(g, n=n, k=k, a=a,
  r2=r2){
(1+g)^((n-k-1-a)/2)*(1+g*(1-r2))^(-(n-1)/2)}

output[1]=((a-2)/2)*integrate(BF.integral,0,
  Inf,n=n,a=a,k=k,r2=r2)$value
output[2]=(2/(k+a))*(f21hyper((n-1)/2,2,
  (k+a)/2+1,r2)/f21hyper((n-1)/2,1,(k+a)/2,r2))
return(output)}
```

[Received December 2010. Revised April 2012.]

REFERENCES

Abramowitz, M., and Stegun, I. (1972), *Handbook of Mathematical Functions*, New York: Dover. [107]

Ball, R. (2005), "Experimental Designs for Reliable Detection of Linkage Disequilibrium in Unstructured Random Population Association Studies," *Genetics*, 170, 859–873. [104]

Bartlett, M. S. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534. [105]

Bayarri, M. J., and García-Donato, G. (2007), "Extending Conventional Priors for Testing General Hypotheses in Linear Models," *Biometrika*, 94, 135–152. [109]

Berger, J. (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385–402. [105]

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–352. [105]

Berger, J. O., and Jefferys, W. H. (1992), "The Application of Robust Bayesian Analysis to Hypothesis Testing and Occam's Razor," *Statistical Methods and Applications*, 1, 17–32. [110]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *p* Values and Evidence," *Journal of the American Statistical Association*, 82, 112–139. [105,108]

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley. [104]

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading: Addison-Wesley. [107,108]

Casella, G., and Moreno, E. (2006), "Objective Bayesian Variable Selection," *Journal of the American Statistical Association*, 101, 157–167. [109]

Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors," *Canadian Journal of Statistics*, 24, 17–36. [105]

Clyde, M. (1999), "Bayesian Model Averaging and Model Search Strategies" (with discussion), *Bayesian Statistics*, 6, 157–185. [109]

Cui, W., and George, E. (2008), "Empirical Bayes vs. Fully Bayes Variable Selection," *Journal of Statistical Planning and Inference*, 138, 888–900. [107]

Dellaportas, P., Forster, J., and Ntzoufras, I. (in press), "Joint Specification of Model Space and Parameter Space Prior Distributions," *Statistical Science*, in press. [107]

Dickey, J. M. (1971), "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters," *The Annals of Mathematical Statistics*, 42, 204–223. [105]

Draper, N., and Smith, H. (1998), *Applied Regression Analysis*, New York: Wiley-Interscience. [104,105]

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242. [108]

Faraway, J. (2002), "Practical Regression and ANOVA using R," available at *http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf*. [104,108,109]

Fernandez, C., Ley, E., and Steel, M. (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427. [106]

Foster, D., and George, E. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975. [106]

Gelman, A. (2008), "Objections to Bayesian Statistics," *Bayesian Analysis*, 3, 445–450. [105]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [104]

George, E., and McCulloch, R. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [105,109]

Gönen, M., Johnson, W. O., Lu, Y., and Westfall, P. H. (2005), "The Bayesian Two-Sample *t* test," *The American Statistician*, 59, 252–257. [104,109]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [109]

Ishwaran, H., and Rao, J. (2003), "Detecting Differentially Expressed Genes in Microarrays using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455. [104]

Jeffreys, H. (1961), *Theory of Probability*, Oxford, UK: Oxford University Press. [105,106]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 377–395. [105]

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934. [106]

Kaufman, C. G., and Sain, S. R. (2010), "Bayesian Functional ANOVA Modeling Using Gaussian Process Prior Distributions," *Bayesian Analysis*, 5, 123–150. [104]

Kuo, L., and Mallick, B. (1998), "Variable Selection for Regression Models," *Sankhya: The Indian Journal of Statistics,* Series B, 60, 65–81. [109]

Leamer, E. (1978), "Regression Selection Strategies and Revealed Priors," *Journal of the American Statistical Association*, 73, 580–587. [105]

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [104,106,107]

Lindley, D. (1980), "L. J. Savage—His Work in Probability and Statistics," *The Annals of Statistics*, 8, 1–24. [105]

Lindley, D. V. (2000), "The Philosophy of Statistics," *The Statistician*, 49, 293–337. [104]

Maruyama, Y. (2009), "A Bayes Factor with Reasonable Model Selection Consistency for ANOVA Model," *Arxiv preprint arXiv:0906.4329*. [104]

Mitchell, T., and Beauchamp, J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [105,109]

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Hoboken, NJ: Wiley. [104]

O'Hagan, A., and Forster, J. (2004), *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference* (2nd ed.), London: Arnold. [104]

O'Hara, R., and Sillanpää, M. (2009), "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–118. [109]

Poirier, D. J. (2006), "The Growth of Bayesian Methods in Statistics and Economics Since 1970," *Bayesian Analysis*, 1, 969–980. [104]

Press, S., Chib, S., Clyde, M., Woodworth, G., and Zaslavsky, A. (2003), *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, New York: Wiley-Interscience. [105]

Qian, S., and Shen, Z. (2007), "Ecological Applications of Multilevel Analysis of Variance," *Ecology*, 88, 2489–2495. [104]

R Development Core Team. (2004), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, ISBN 3–900051–00–3. [107]

Robert, C. (1993), "A Note on Jeffreys–Lindley Paradox," *Statistica Sinica*, 3, 601–608. [105]

Scott, J., and Berger, J. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [110]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [108]

Sen, S., and Churchill, G. (2001), "A Statistical Framework for Quantitative Trait Mapping," *Genetics*, 159, 371–287. [104]

Shafer, G. (1982), "Lindley's Paradox," *Journal of the American Statistical Association*, 77, 325–351. [105]

Stephens, M., and Balding, D. J. (2009), "Bayesian Statistical Methods for Genetic Association Studies," *Nature Reviews Genetics*, 10, 681–690. [110]

Strawderman, W. (1971), "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," *The Annals of Mathematical Statistics*, 42, 385–388. [107]

Westfall, P., and Gönen, M. (1996), "Asymptotic Properties of ANOVA Bayes Factors," *Communications in Statistics-Theory and Methods*, 25, 3101–3123. [104]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with *g*-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, pp. 233–243. [105]

Zellner, A. (1987), *An Introduction to Bayesian Inference in Econometrics*, Malabar, FL: RE Krieger Pub. Co. [105]

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: University Press, pp. 585–603. [106]

Zeugner, S., and Feldkircher, M. (2009), "Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging," IMF Working Papers, 9, 202. [107]