RESTRICTED MOST POWERFUL BAYESIAN TESTS

A Dissertation

by

SCOTT DAVID GODDARD

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Valen E. Johnson
Committee Members,	Bani K. Mallick
	Pat R. Goldsmith
	Raymond J. Carroll
Head of Department,	Valen E. Johnson

May 2015

Major Subject: Statistics

Copyright 2015 Scott David Goddard

ABSTRACT

Uniformly most powerful Bayesian tests (UMPBTs) are defined to be Bayesian tests that maximize the probability that the Bayes factor against a fixed null hypothesis exceeds a specified evidence threshold. Unfortunately, UMPBTs exist only in a relatively limited number of testing scenarios, and in particular they cannot be defined for most tests involving linear models. In this dissertation, I generalize the notion of UMPBTs by restricting the class of alternative hypotheses that are considered in the test of a given null hypothesis. I call the resulting class of Bayesian hypothesis tests restricted most powerful Bayesian tests (RMPBTs). I then derive RMPBTs for linear models by restricting the class of possible alternative hypotheses to g-priors.

An important feature of the resulting class of tests is that their rejection regions coincide with the rejection regions of usual frequentist F-tests, provided that the evidence thresholds for the Bayesian tests are appropriately matched to the size of the classical tests. This correspondence leads to the definition of default Bayes factors for many common tests of linear hypotheses. I illustrate the use of RMPBTs in the special cases of ANOVA and one- and two-sample t-tests. I then use RMPBTs to develop a novel Bayesian variable selection method and compare its performance to other Bayesian tests based on g-priors in a sequence of numerical examples.

Finally, a software package for R is detailed which implements the RMPBTs described herein as well as many of the UMPBTs that have been developed.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

This dissertation describes advances in Bayesian hypothesis testing and model selection, with special application to linear models. To begin, we describe some notation and preliminaries. Let $\boldsymbol{\epsilon}$ be a normal random vector of length n with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$ for some $\sigma^2 > 0$. Let $\mathbf{X}$ be a fixed matrix of dimension $n \times p$ for some $p > 0$ and let $\boldsymbol{\beta}$ be a (possibly fixed or random) vector of length p. The general linear model for the random vector $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Often in scientific research it is desirable to determine whether a specific $\mathbf{X}$, call it $\mathbf{X}_1$, is more appropriate than another, which might be called $\mathbf{X}_0$. This could be due to a desire to test some hypothesis H_0 versus an alternative H_1, wherein H_0 and H_1 are stated in terms of $\mathbf{X}_0$ and $\mathbf{X}_1$, or it could form a single step in some general routine to select an $\mathbf{X}$ from among a large number of candidates.

In both cases, Bayesian methods commonly compare the appropriateness of $\mathbf{X}_1$ and $\mathbf{X}_0$ by comparing the probability that each is "right", given $\mathbf{y}$, with some additional assumptions on $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. These conditional probabilities are calculated via Bayes' theorem, which provides a mathematical method for updating prior beliefs about the probabilities of $\mathbf{X}_1$ and $\mathbf{X}_0$ with new information from the data. The simplest form of the theorem can be written

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)},$$

where A and B are events and $\mathbf{P}(\cdot)$ denotes a probability. Bayes theorem can be

easily modified to express the odds of event A given B. In this form, it is called Bayes' rule, and the simplest version is

$$\frac{\mathbf{P}(A|B)}{\mathbf{P}(A^C|B)} = \frac{\mathbf{P}(B|A)}{\mathbf{P}(B|A^C)} \cdot \frac{\mathbf{P}(A)}{\mathbf{P}(A^C)},$$

where A^C denotes the complement of event A. If one regards A^C as the event that some null hypothesis H_0 is true, and A as the event that an alternative hypothesis H_1 is true, and B as the event that a given data set $\mathbf{y}$ is sampled from the population, Bayes' rule can be written as

$$\frac{\mathbf{P}(H_1|\mathbf{y})}{\mathbf{P}(H_0|\mathbf{y})} = \frac{m(\mathbf{y}|H_1)}{m(\mathbf{y}|H_0)} \cdot \frac{\mathbf{P}(H_1)}{\mathbf{P}(H_0)},$$

where $m(\mathbf{y}|H_i)$ is a probability mass function under hypothesis i. With a limiting argument we can show that the relation holds for probability density functions as well, which is the case to which we will restrict attention hereafter. It is easily seen that the posterior odds in favor of the alternative hypothesis equal the prior odds times the ratio of the marginal densities on $\mathbf{y}$. This ratio is known as the Bayes factor, and it is written more explicitly as

$$BF_{10} = \frac{m(\mathbf{y}|H_1)}{m(\mathbf{y}|H_0)} = \frac{\int_{\boldsymbol{\Theta}} f(\mathbf{y}|H_1, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|H_1)\ d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} f(\mathbf{y}|H_0, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|H_0)\ d\boldsymbol{\theta}},$$

where $f(\mathbf{y}|H_i, \boldsymbol{\theta})$ is a density function in $\mathbf{y}$ indexed by the parameter vector $\boldsymbol{\theta}$ given hypothesis i ($i \in \{0, 1\}$), and $\pi(\boldsymbol{\theta}|H_i)$ is a density in $\boldsymbol{\theta}$, given hypothesis i.

The prior density of the parameter vector $\boldsymbol{\theta}$ under the alternative hypothesis, $\pi(\boldsymbol{\theta}|H_1)$, is not necessarily specified by the research problem at hand. Frequently the Bayes factor is very sensitive to the choice of this density, even for large sample sizes, and the absence of an objective selection opens researchers to criticism of

subjectivity. In 2013, Johnson [2] defined Uniformly Most Powerful Bayesian Tests (UMPBTs), and in so doing, introduced a default prior $\pi(\boldsymbol{\theta}|H_1)$ for one-parameter exponential family models. He concluded:

> Additional research is needed to identify classes of models and testing contexts for which UMPBTs can be defined. The UMPBTs described in this article primarily involve tests of point null hypotheses, or tests that can be reduced to a test of a point null hypothesis after marginalizing over nuisance parameters. Whether UMPBTs can be defined in more general settings remains an open question.

This dissertation describes a set of findings which extend the results in Johnson [2, 1] to tests of a point null hypothesis and model selection methods in the general linear model. Along the way, two other prominent themes from Johnson's work receive additional attention: the equivalence (or non-equivalence) of frequentist and Bayesian tests, and statistical power. These themes underscore the rationale behind the most powerful Bayesian tests described and highlight some interesting features they possess.

The dissertation is organized as follows. Section 2 provides a literature review of Bayesian testing, UMPBTs, Bayesian and frequentist testing equivalence, g priors, Bayesian model selection, and R computing for Bayesian hypothesis testing and model selection. Section 3 describes Restricted Most Powerful Bayesian Tests (RMPBTs) and applies them specifically to linear models using a g prior. Section 4 develops a model selection framework using g prior-RMPBTs and compares it to previously proposed model selection methods in a simulation routine. Section 5 describes an R package, MPBT, which provides easy-to-use functions for implementing the tests in Johnson's work and these chapters. Section 6 provides a summary of

the developments herein and proposes further research. Proofs to the theorems in Section 3 are found in Appendix A.

2. LITERATURE REVIEW

This chapter briefly reviews six topics in the statistical literature which will be treated in the development of this dissertation. Subsection 2.1 gives a general overview of Bayesian hypothesis testing. Subsection 2.2 reviews the short history of Uniformly most powerful Bayesian testing. Subsection 2.3 describes equivalences between Bayesian and frequentist hypothesis tests and Subsection 2.4 reviews Zellner's g prior in the general linear model. Subsection 2.5 summarizes in brief fashion the extensive literature on Bayesian model selection, and Subsection 2.6 touches on the major packages that have been developed for Bayesian hypothesis testing and model selection in R.

2.1 Bayesian Testing

An early development of Bayesian hypothesis testing can be found in Jeffreys [3]. He showed the need to specify a prior density on the parameter of interest under the alternative hypothesis but did not prescribe a specific method to set priors under alternative hypotheses. Instead, he mentioned that multiple such priors may need to be explored, and that prior information should be taken into account. Outside the context of hypothesis testing, Jeffreys [4], many before him (e.g. [5]), and many since (e.g. [6]) have sought non-informative priors that can be assumed in the absence of prior knowledge.

Bayes factors, the critical component in a Bayesian test, were given a general overview in Kass and Raftery [7]. Smith and Spiegelhalter [8] distinguished between global and local Bayes factors, finding in the former a relationship to the Schwarz information criterion, and in the latter a relationship to the Akaike information criterion.

Bayesian tests of specific models are described in Gelman [9], Rouder et al. [10], and Solari, Liseo, and Sun [11], who discussed ANOVA, and Rouder et al. [12], who discussed t-tests. Between these references, there is no consensus regarding prior specification under the alternative hypothesis.

2.2 Uniformly Most Powerful Bayesian Testing

Johnson [2] described the problems associated with subjective Bayesian methods, including the specification of a prior density in hypothesis testing. He commented that "subjective Bayesian testing procedures have not been–and will likely never be–generally accepted by the scientific community." Following this, he defined a Uniformly Most Powerful Bayesian Test for evidence threshold γ [UMPBT(γ)] against a fixed null hypothesis H_0 to be the hypothesis test in favor of an alternative hypothesis H_1 that maximizes the probability that the Bayes factor in favor of H_1 exceeds the evidence threshold γ. That is, the UMPBT(γ) test satisfies

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

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

Next, he showed that UMPBTs exist for one-parameter exponential family models under mild regularity conditions and derived the UMPBTs for one- and two-sample z tests, tests of a binomial success probability, tests of linear regression coefficients when σ^2 is known, and several other models. Finally, he provided approximate UMPBTs for one-sample t tests and tests of linear regression coefficients when σ^2 is unknown.

Johnson [1] revised the approximate UMPBT for the one-sample t-test and provided an approximate UMPBT for the two-sample t-test, while still acknowledging their limited usefulness due to the large sample size required to satisfy the approxi-

mation and their data-dependence.

2.3 Bayesian and Frequentist Testing Equivalence

An important property of the UMPBTs described in Johnson [2] is that the rejection regions for these tests (i.e., the values of $\mathbf{y}$ for which $BF_{10} > \gamma$) can be made to coincide with the rejection regions for classical uniformly most powerful tests (UMPTs) by setting γ as a particular function of the size of the classical test α. In this way, a p-value and a Bayes factor which result in the same conclusions can be computed and compared. In one example of a phase II clinical trial, Johnson found that a p-value of 0.05 corresponded to a posterior probability in favor of the null of between 0.13 and 0.30 if equal prior probabilities were assigned to each hypothesis. In [1], Johnson found that for a range of common tests, a p-value of 0.05 corresponded to a posterior probability of the null hypothesis (still assuming equal prior probabilities on the hypotheses) of between 0.17 to 0.25, while a p-value of 0.01 corresponded to a posterior probability of between 0.05 and 0.08. In light of this, he went on to estimate that between 17-25% of marginally significant scientific findings are false.

These findings track closely to earlier attempts to quantify a p-values by finding a corresponding Bayes factor. Edwards, Lindman, and Savage [13] found some special cases where the two approaches could be compared and noted that "Often evidence which, for a Bayesian statistician, strikingly supports the null hypothesis leads to rejection of that hypothesis by standard classical procedures". Bernardo [14] developed a prior distribution under the alternative and a prior probability of the null using information theoretic arguments, enabling him to calculate the asymptotic posterior probability in favor of the null hypothesis for a given p-value and found that a result significant at the 0.05 level corresponded to a posterior probability in favor of the null hypothesis of about 0.2. Dickey [15] estimated that the 0.05 p-value corresponded

to a posterior probability of between 0.25 and 0.58, depending on the sample size. Berger and Sellke [16] and Berger and Delampady [17] examined normal and binomial models and found lower bounds on Bayes factors and posterior probabilities over a wide class of priors, concluding that the discrepancy between p-values and Bayes factors puts them in dramatic conflict. Berger, Boukai, and Wang [18] showed that Bayesian tests are virtually equivalent to frequentist tests if a conditional frequentist method is utilized.

Although the various methods employed for calculating a Bayes factor based on a p-value produce similar and striking results, they are open to criticism for resorting to approximations or relying on subjective choices of priors. By matching rejection regions of the Bayesian and frequentist tests, the UMPBTs demonstrate more robustly and objectively that p-values can exaggerate evidence against the null hypothesis in certain cases where UMPBTs exist.

2.4 g Priors

The g prior was first suggested by Zellner [19]. He assumed that a length-n random vector $\mathbf{y}$ could be modeled by the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e}$ had a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. With a nuisance parameter prior density $\pi(\sigma^2) \propto 1/\sigma^2$, he defined the g prior density on the parameter vector $\boldsymbol{\beta}$ to be

$$\pi(\boldsymbol{\beta}|\sigma^2) = (2\pi)^{-p/2} \cdot |g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|^{-1/2} \cdot \exp\{-1/(2g\sigma^2)\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\},$$

or a normal density with mean $\mathbf{0}$ and covariance matrix $g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ for some $g > 0$. Here, p is defined to be the dimension of $\boldsymbol{\beta}$.

The g prior's conjugacy and interpretability have facilitated its widespread adoption, but numerous methods for setting g have been proposed and there is no consensus of opinion regarding which is best. In a hypothesis testing or model selection scenario, these proposals include $g = n$ (where n is the sample size) according to the unit information prior (UIP) [7]; $g = p^2$ (where p is the number of covariates in the model) according to the risk information criterion (RIC) [20]; $g = \max(n, p^2)$ [21]; $g = \log(n)^3$ according to the Hannan-Quinn information criterion [22]; $g = \max(\hat{F} - 1, 0)$ where $\hat{F}$ is the usual F statistic, which is a local Empirical Bayes prior (EBL) [23]; and a global Empirical Bayes prior (EBG) [23]. Liang et al. [23] reviewed some proposed methods and discussed a second approach, which is to place a prior distribution on g. They argued that the multivariate Cauchy priors advocated in Zellner and Siow [24] are equivalent to an Inverse Gamma(1/2, $n/2$) prior on g. They also introduced a hyper-g prior, which takes the form:

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

Guo and Speckman [25] put the Jeffreys prior on g and found that the resulting Bayes factor is consistent.

2.5 Bayesian Model Selection

For the purposes of this dissertation, the terms "model selection" and "variable selection" will be used interchangeably, which is a reflection of the fact that, in many linear modeling scenarios, the data analyst will fully specify the model up to the inclusion of certain candidate predictors *a priori*. Selecting between models, in such a case, consists merely in selecting predictors.

Bayesian model selection is a vast area which encompasses many of the other topics reviewed here. Review articles are provided by Wasserman [26] and O'Hara and Sillanpää [27]. The latter authors classified the major methods in the literature into four types: indicator model selection, stochastic search variable selection, adaptive shrinkage, and model space approaches. The first three classes place priors directly on the coefficients of the candidate variables and calculate their posterior distributions and posterior inclusion probabilities. The fourth class adds a prior on the model size to calculate posterior distributions on model size and coefficients. In each of these classes, there is no direct reliance on using Bayes factors to compute posterior model probabilities. The review by Wasserman discussed this approach, which forms an additional class of model selection methods. We focus on the literature for this last class, since it most closely relates to the development of this dissertation.

Notably, model selection methods that rely on Bayes factors can seem, in many cases, unappealing because the Bayes factors can be difficult to calculate. This has lead to the development of estimating procedures such as a method in Spiegelhalter and Smith [28], fractional Bayes factors from O'Hagan [29] and intrinsic Bayes factors from Berger and Pericchi [30, 31]. Bayes factors can also be replaced by the Bayesian information criterion, which gives an approximation to the log of the Bayes factor.

The article by Liang et al. [23] investigated model selection in the normal linear model with a g prior. In a simulation model selection study with 15 predictors, they concluded that the Zellner and Siow method, the Empirical Bayes procedures, and the hyper-g prior perform nearly identically well, and outperform the other priors. No work has been published regarding the performance of UMPBTs in a model selection framework, but the tests could be easily modified to accommodate comparisons between multiple models and may possess desirable properties in model

selection such as Bayes factor coherence and consistency.

In close relationship to the model selection problem lies the model averaging problem, which attempts not to select one "best" model but to average the output from all models considered. This averaging is analogous to the averaging over nuisance parameters that is done in certain Bayesian calculations which marginalize parameters of interest. In this case, the models are weighted by their posterior probabilities. Bayesian model averaging is described in [32].

2.6 R Computing for Bayesian Hypothesis Testing and Model Selection

The development of R tools for Bayesian hypothesis testing and variable selection has resulted in the availability of several useful packages for download. We discuss packages for hypothesis testing first.

The 'BEST' package (for "Bayesian Estimation Supersedes the t-Test") offers an alternative to one- and two-sample t-tests by providing posterior estimates for group means and their differences [33]. The premise of the functions is not to facilitate hypothesis testing but to prevent it.

The package 'BayesFactor' offers several functions for computing Bayes factors in various testing scenarios, including one- and two-sample t tests, general ANOVA designs, and linear regression. Functions testing general linear models, regression models, and ANOVA models assume the JSZ prior on regression coefficients. The functions contain functionality for testing multiple models at once [34].

There are many packages which extend the functionality for testing multiple models to a general Bayesian variable selection or Bayesian model averaging capability. Some which do either Bayesian variable selection or model averaging can in fact do both, since the difference merely depends on the way the posterior distribution on the model space is summarized. We first discuss three packages written specifically

for Bayesian model averaging problems.

Clyde et al. [35, 36] described the package BAS (for "Bayesian Adaptive Sampling") as a set of functions designed to obtain a posterior distribution on the model space in the linear model variable selection problem. Available prior distributions on regression coefficients include many g priors and mixtures of g priors (e.g. JZS and hyper-g), but model selection criteria such as AIC and BIC are also available [37]. The main function, `bas`, can either search the model space exhaustively when there are less than 25 covariates or use adaptive sampling without replacement for larger model spaces.

The package BMS (for "Bayesian Model Selection") is written for performing Bayesian model averaging for linear models. The syntax of its main function, `bms`, requires the specification of a fixed value for g or the name of a fixed (e.g. UIP or HQ) or model-specific prior (e.g. RIC, BRIC, EBL, and hyper-g) [38, 39]. However, more flexible prior specifications are possible using other functions [37]. The BMS package can enumerate the model space when there are less than 15 covariates and search exhaustively or utilize various MCMC approaches to search stochastically for larger model spaces.

The package BMA (for "Bayesian Model Averaging") carries out averaging for linear models and certain nonlinear models [40]. For the linear models, it does not use g priors on regression coefficients, but instead employs the BIC approximation, which gives somewhat similar results to a g prior with g set by the UIP [37]. The package can either perform an exhaustive search of the model space using the leaps and bounds algorithm or utilize a Markov chain to search the model space stochastically [41].

These three packages were reviewed by Amini and Parmeter [37]. In a comparison study, they found that the BAS package is usually faster than its competitors, both

for small models and especially for large models. When trying to reproduce the results of two published data analyses that carried out Bayesian model averaging using handwritten code, they found that BMS gave results which were most consistent to those which were published [37].

There are also several packages nominally designed specifically for Bayesian variable selection. The BayesVarSel package was designed to perform variable selection using JZS, hyper-g, UIP, and BRIC priors, as well as the "Robust" prior from [42]. It is written to search the model space either exhaustively or using a Gibbs sampler [43]. The spikeSlabGAM package performs Bayesian variable selection for Gaussian and certain types of non-Gaussian responses in additive mixed regression models. The package is designed to fit spike and slab priors on regression coefficients, rather than g priors [44]. The modelSampler package, likewise, performs Bayesian variable selection using spike and slab priors [45]. The mombf package performs model selection when non-local priors are put on regression coefficients [46].

In addition, there exists functionality for implementing certain Bayesian variable selection methods in R through packages that connect R to independent MCMC engines such as WinBUGS, OpenBUGS, and JAGS.

3. RMPBTS FOR LINEAR MODELS

In Section 2 we reviewed Johnson's [2] definition of Uniformly Most Powerful Bayesian Tests (UMPBTs). UMPBTs exist in a relatively limited number of testing scenarios (e.g., one parameter exponential families), and in particular they cannot be defined for tests of parameters in the general linear model when variance parameters are not known *a priori*.

To remedy this situation, we define an extension of UMPBTs that we call restricted most powerful Bayesian tests. The extension is obtained by restricting the class of prior densities on $\boldsymbol{\theta}$ that define the hypotheses to a parametric class, say $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\psi})$.

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

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

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

In essence, we obtain RMPBTs by narrowing the search of alternative hypotheses to a class of prior densities on $\boldsymbol{\theta}$. We assume that this class either incorporates prior knowledge or provides computational convenience. The optimization within this class

produces a value for one or more hyperparameters $\boldsymbol{\psi}$ which maximize the probability that the Bayes factor exceeds γ over all possible values of $\boldsymbol{\psi}$ and over all $\boldsymbol{\theta}_t$.

The remainder of this section is organized as follows. In Subsection 3.1 we show that by restricting the class of prior densities to g priors in the general linear model, we are able to define an RMPBT, and that the value of g has a simple form when the test's rejection region is matched to a classical α-size test. We then specialize this result for ANOVA and t-testing scenarios. In Subsection 3.2, we present two simulation studies to compare the g prior-RMPBT to other Bayesian methods for setting g, and finally in Subsection 3.3 we illustrate an application of our method to a real data set.

3.1 g Prior-RMPBTs

We begin by considering the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n) \tag{3.1}$$

$$= \mathbb{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \tag{3.2}$$

We partition $\mathbf{X}$ and $\boldsymbol{\beta}$ so that tests of hypotheses on subsets of $\boldsymbol{\beta}$ are performed on the sub-vector $\boldsymbol{\beta}_1$. The prior density that we propose is based on Zellner's g prior [19], which in the general linear model leads to

$$\boldsymbol{\beta}_1|g, \sigma^2 \sim \mathcal{N}(\mathbf{0}, g\sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}), \quad \text{and} \quad \pi(\sigma^2,\ \beta_0,\ \boldsymbol{\beta}_2) \propto 1/\sigma^2.$$

If we restrict attention to prior densities of this form, and assume (without loss of generality) that the model has been parameterized in such as way that $\mathbb{1}_n^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} = \mathbf{0}$ and $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$, then the value of g that provides the RMPBT(γ) is provided by

the following theorem.

Theorem 3.1.1 *Suppose that* $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$*, and partition* $\mathbf{X}$ *and* $\boldsymbol{\beta}$ *according to* $\mathbf{X} = \begin{bmatrix} \mathbb{1}_n & \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$ *and* $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}_1^T & \boldsymbol{\beta}_2^T \end{bmatrix}^T$*, where* $\mathbf{X}_i$ *has* p_i *columns and* $p = p_1 + p_2$*. Assume* $n > p_2 - 1$*, and that the design matrix has been constructed so that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are of full-column rank,* $\mathbb{1}_n^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} = \mathbf{0}$*, and* $\mathbf{X}_1^T \mathbf{X}_2 = \mathbf{0}$*. Assume further that the joint prior distribution on* σ^2*,* β_0*, and* $\boldsymbol{\beta}_2$ *is proportional to* $1/\sigma^2$*. If the null hypothesis is* $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ *and the alternative hypothesis is restricted to take the form*

$$H_1 : \boldsymbol{\beta}_1 | g, \sigma^2 \sim \mathcal{N}(\mathbf{0}, g\sigma^2 (\mathbf{X}_1^T \mathbf{X}_1)^{-1}) \tag{3.3}$$

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

$$\arg\max_{g^*} \; (g^*)^{-1} \left[\gamma^{\frac{-2}{n-p_2-1}} (1 + g^*)^{\frac{n-p-1}{n-p_2-1}} - 1 \right]. \tag{3.4}$$

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

This theorem is important because it provides a default alternative hypothesis for constructing a Bayesian test of regression coefficients in a linear model. Although other objective methods for setting g have been proposed [23], the RMPBT computed from (3.4) provides greater probability that the Bayes factor exceeds the given threshold (i.e., has greater power) than any other alternative hypothesis taking the form (3.3). As we demonstrate in Subsection 3.2, the resulting difference in power can often be quite appreciable.

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

$$BF_{10}(\mathbf{y}) = (1+g)^{(n-p-1)/2}\left[1 + g \cdot \frac{n-p-1}{\hat{F}p_1 + n - p - 1}\right]^{-(n-p_2-1)/2}. \tag{3.5}$$

The evidence threshold γ must be determined before g can be computed from (3.4). In classical terms, the evidence threshold plays a role that is similar to the size of a test; it specifies the value of the Bayes factor required to reject the null hypothesis in favor of the alternative. In the case of UMPBTs, Johnson [2] fixed evidence thresholds by equating the rejection regions of UMPBTs and frequentist tests possessing specified type-I error rates. We propose to extend this idea for application to RMPBTs; the next theorem provides a mechanism for doing this.

Theorem 3.1.2 *Under the conditions in Theorem 3.1.1, the value of g that produces a g prior-RMPBT that has the same rejection region as a size-α classical F-test is obtained by setting*

$$g = F_{1-\alpha} - 1, \tag{3.6}$$

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

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

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

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

There is an interesting similarity between the expression for g in Theorem 3.1.2 and the local empirical Bayes estimate for g described in Liang et al. [23],

$$\hat{g}^{\mathrm{EBL}} = \max\{\hat{F} - 1, 0\}. \tag{3.7}$$

The RMPBT value for g in (3.6) is obtained from (3.7) by substituting $F_{1-\alpha}$ for $\hat{F}$. The implications of this difference are explored in Subsection 3.2.

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

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

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

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

identifiable, various constraints can be used. One option is to eliminate one column of $\mathbf{X}$ (equivalent to setting one component of $\boldsymbol{\beta}_1$ equal to 0). The use of such constraints in the Bayesian setting has generated discussion in Gelman [9] and Rouder et al. [10]. Gelman recommended an alternative constraint $\mathbb{1}^T \boldsymbol{\beta}_1 = 0$, whereas Rouder et al. employed this constraint only for fixed factors. In the following corollaries we assume that an identifiable parameterization of the design matrix has been specified, although the particular parameterization used is not important as long as the

following conditions are satisfied:

1. the design matrix can be written as $\mathbf{X}^* = \begin{bmatrix} \mathbb{1}_n & \mathbf{X}_1^* \end{bmatrix}$ for some $n \times (J-1)$ matrix $\mathbf{X}_1^*$, and

2. the column space of $\mathbf{X}^*$ is the same as the column space of $\mathbf{X}$ (i.e., the column space of $\mathbf{X}_1^*$ is equivalent to $\mathbf{X}_1$).

The parameter vector constraints described by the functions `contr.treatment`, `contr.SAS`, `contr.sum`, `contr.helmert`, and `contr.poly` in R are all examples of parameterizations that satisfy these conditions. We define β_0^* and $\boldsymbol{\beta}_1^*$ as the corresponding regression parameters.

The principal problem in applying Theorem 3.1.1 to the one-way ANOVA setting is that the condition $\mathbb{1}^T\mathbf{X}_1^* = \mathbf{0}$ is not, in general, satisfied by $\mathbf{X}_1^*$. Wetzels et al. [47] resolved this problem by centering the columns in $\mathbf{X}_1^*$ so that the resulting model is

$$\mathbf{y} = \mathbb{1}_n\beta_0^* + (\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\,\mathbf{X}_1^*\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}, \tag{3.8}$$

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

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

and as a result, the test that $\boldsymbol{\beta}_1^* = \mathbf{0}$ in model (3.8) and the classical one-way ANOVA test have the same null hypothesis. For concreteness, we use the Wetzels et al. parameterization to state the g prior RMPBT for one-way ANOVA tests.

Corollary 3.1.3 *Assume that*

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

where y_{ij} is observation i under treatment j for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ and ϵ_{ij} are independent, mean-zero normally-distributed observational errors with constant variance σ^2. Under the parameterization in (3.8)*, assume that the prior density for (σ^2, β_0^*) is given by*

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

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

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

is obtained by setting g equal to

$$\arg\max_{g^*} \; (g^*)^{-1} \left[\gamma^{\frac{-2}{n-1}}(1+g^*)^{\frac{n-J}{n-1}} - 1\right].$$

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

$$g = F_{1-\alpha} - 1, \tag{3.9}$$

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

$$\gamma = \left[\frac{(J-1)F_{1-\alpha} + n - J}{F_{1-\alpha}^{(J-1)/(n-1)}(n-1)}\right]^{(n-1)/2}. \tag{3.10}$$

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

Bayes factor to the parameterization of the design matrix follows from its expression as

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

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

In addition to the UMPBTs developed in Johnson [1], approximate UMPBTs are given for one- and two-sample t-tests. However, these approximations fail for large values of the sample mean $\bar{y}$. As an alternative, Corollary 3.1.3 can be applied to obtain a g prior-RMPBT for the two-sample t-test as follows.

Corollary 3.1.4 *Assume that $y_{ij}|\beta_0,\beta_1,\beta_2,\sigma^2$ are conditionally independent normally distributed random variables with mean $\beta_0 + \beta_j$ and variance σ^2 for $i = 1,\ldots,n_j$ and $j = 1,\ 2$. Under model (3.8), let $\pi(\sigma^2,\beta_0^*) \propto 1/\sigma^2$ and suppose that the design matrix satisfies the two conditions stated above. For the test of*

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

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

$$\arg\max_{g^*}\ (g^*)^{-1}\left[\gamma^{\frac{-2}{n-1}}(1+g^*)^{\frac{n-2}{n-1}}-1\right].$$

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

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

where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a t distribution with $n-2$ degrees of freedom.

Moreover, the evidence threshold γ is given by

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

In this test, the Bayes factor can be written as a function of the classical t statistic $\hat{t}$ and a quantile from the t distribution as

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

The previous corollaries describe RMPBTs for the one-way ANOVA and two-sample t-tests. These corollaries follow directly from Theorem 3.1.1. However, one-sample t-tests are not a special case of Theorem 3.1.1 because these tests are tests of the intercept term (rather than the effect term) in that theorem. Instead, the following theorem describes the g prior-RMPBT for a one-sample t-test. Without loss of generality, we consider only the case of testing $H_0 : \beta_0 = 0$.

Theorem 3.1.5 *Assume that $y_i|\beta_0, \sigma^2$ are independent normally-distributed random variables with mean β_0 and variance σ^2 for $i = 1, \ldots, n$. Under the priors $\pi(\sigma^2) \propto 1/\sigma^2$ and $\beta_0|g, \sigma^2 \sim \mathcal{N}(0, g\sigma^2/n)$, the g prior-RMPBT that $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ for evidence threshold γ is obtained by setting g equal to*

$$\arg\max_{g^*} \; (g^*)^{-1} \left[(1+g^*)^{(n-1)/n} \gamma^{-2/n} - 1 \right]. \tag{3.11}$$

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

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

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

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

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

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

3.2 Numerical Comparisons of g Prior-based Bayes Factors in Linear Models

In this section, we compare the performance of some of the methods from the literature for setting g, discussed in Section 2, to that of the g prior-RMPBT in a numerical study. This study evaluates performance in terms of statistical power; we therefore estimate power functions for each method in a simulated testing problem. A second simulation study compares the power functions for the two-sample t-test under the g prior-RMPBT and the approximate UMPBT.

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

For simplicity, we restrict attention to a balanced one-way ANOVA test in which

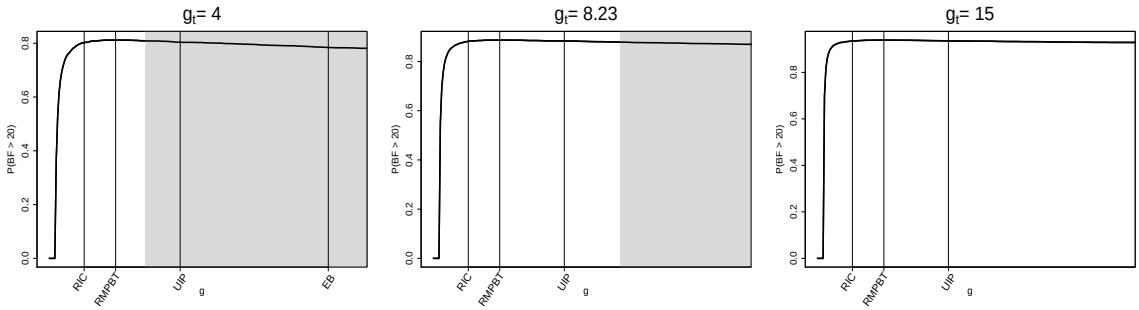

Figure 3.1: Numerical simulation results comparing the probability of exceeding the threshold $\gamma = 20$ for all values of g between 0 and 35. Vertical lines indicate the values of the RIC value of g; the RMPBT value of g; the UIP value of g; and the mean EBL value of g. The shaded region represents the middle 50% of EBL values of g. In this simulation, $J = 3$, $n = 15$, and $\sigma^2 = 5$.

the true model for the random effects $\boldsymbol{\beta}_1^*$ is given by

$$\boldsymbol{\beta}_1^* \sim \mathcal{N}\left(\mathbf{0}, g_t\sigma^2(\mathbf{X}_1^{*T}(\mathbf{I} - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*)\right). \tag{3.12}$$

Here, g_t is a fixed "true" value of g, $\sigma^2 = 5$ is the error variance, and $J = 3$ and $n = 15$. The elements of $\boldsymbol{\beta}_1^*$ are generated from a centered model so that the values of g_t are on the same scale as the RMPBT value of g. Figure 3.1 displays $\mathbf{P}_{g_t}\left[BF_{10}(\mathbf{y}) > \gamma\right]$ as a function of g from three different experiments where data were simulated using model (3.12) with g_t set to different values. The evidence threshold used in this plot was $\gamma = 20$, which is the minimum threshold for "strong" evidence according to the modified schedule in Kass and Raftery [7]. Vertical lines are drawn to indicate values of g corresponding to the RIC prior; the RMPBT prior; the UIP; and the mean EBL prior. Also shown is a shaded region which represents the center 50% of $\hat{g}^{\text{EBL}}$ values from the simulation. Values of g_t were selected to be equal to the RIC value $[(J-1)^2 = 4]$, the RMPBT value (8.23), and the UIP value $(n = 15)$.

The RMPBT, by definition, corresponds to the peak of the curve in each plot.

The RIC and UIP tests provide smaller probabilities that the Bayes factor exceeds the threshold $\gamma = 20$, although these differences in power are mitigated as g_t grows. The variability in results obtained with the local empirical value of g increases quickly as the true value of g_t becomes large.

Results from this simulation study were also used to compare Bayes factors from the RMPBT to classical p-values through the relation specified in (3.10) with $\alpha = 0.05$. To this end, g was set according to (3.9). The left panel of Figure 3.2 displays the resulting correspondence between p-values and Bayes factors for this experiment. This plot illustrates the tendency for the magnitude of p-values to exaggerate evidence against the null hypothesis, as was similarly found in [1]. This tendency argues in favor of requiring more stringent criteria for rejecting tested null hypotheses in frequentist testing.

A separate simulation study was used to compare the power curves of the g prior-RMPBT and the approximate UMPBT from Johnson [1] in two-sample t-tests. The right panel of Figure 3.2 displays the resulting power curves. The two-sided tests were simulated 5,000 times at an increasing sequence of β_2 values, where β_1 and β_0 were held fixed at 0. In this experiment, $n_1 = n_2 = 15$ and $\gamma = 20$. As expected, the approximate UMPBT outperforms the g prior-RMPBT in terms of power for small to moderate values of β_2; this occurs because the UMPBT alternatives are not restricted to the class of g priors. However, as evidence against the null hypothesis becomes strong, the quality of the approximation to the UMPBT decays and its power declines. The g prior-RMPBT does not suffer from this problem and actually provides higher power for large values of β_2.

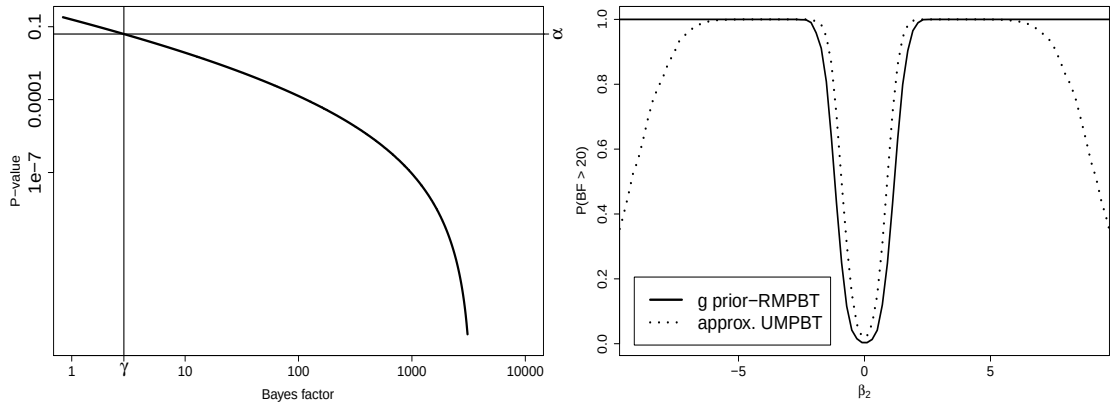

Figure 3.2: Left panel: p-values plotted against Bayes factors using the RMPBT in a one-way ANOVA. As before, $J = 3$, $n = 15$, $\sigma^2 = 5$, and $g_t = 4$. In these plots, $\gamma \approx 2.88$ from (3.10). Right panel: A two-sided two-sample t-test power curve for the g prior-RMPBT and the approximate UMPBT from Johnson [1]. In this simulation, $n_1 = n_2 = 15$, $\gamma = 20$, $\sigma^2 = 1$, and $\beta_0 = \beta_1 = 0$. The power of each test is plotted against a range of β_2 values.

3.3 Example

To illustrate the use of g prior-RMPBTs on real data, we re-analyzed the seaweed grazer data previously analyzed by Qian and Shen [48]. The experimental design in this study was a randomized complete block design with six treatments (grazers) in eight blocks (intertidal locations) with two replications. The response variable y_{ijk} was the logit of the percentage seaweed recovery in the kth experimental plot ($k =$ 1,2) in block j ($j =$ 1,...,8) under treatment i ($i =$ 1,...,6). An ANOVA model for the experiment can be written as

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

where $\boldsymbol{\beta}_1$ is the vector of treatment effects, $\boldsymbol{\beta}_2$ is the vector of block effects, and $\boldsymbol{\beta}_3$ is the vector of interactions. The elements in the vector $\boldsymbol{\epsilon}$ are assumed to be i.i.d.

Table 3.1: RMPBT and frequentist tests results of the seaweed grazers data

Effect	g	BF_{10}	γ	p-value
Treatment	1.41	1.7×10^7	3.0	4.5×10^{-20}
Block	1.21	1.4×10^6	3.2	5.4×10^{-17}
Interaction	0.67	1.76	3.8	0.1209

mean-zero normal random variables.

We begin by testing the interaction effect, $\boldsymbol{\beta}_3$. In the notation of Theorem 3.1.1, we have $n = 96$ and $p = p_1 + p_2 = 35 + (5 + 7)$, so that with the intercept β_0 there are 48 parameters in the model. If we set g and γ so that the RMPBT corresponds to a 5% classical test, then $g = 0.67$ and $\gamma = 3.83$. The Bayes factor for the resulting RMPBT test is 1.76 and the p-value is 0.12.

The main effects are tested next. The Bayes factor of the treatment effect is 1.7×10^7, and that of the blocking effect is 1.4×10^6. The corresponding p-values are 4.5×10^{-20} and 5.4×10^{-17}, respectively. These results are summarized in Table 3.1.

We emphasize that the Bayes factors cited in this example were obtained through straightforward calculations that were based only on the F statistics reported from standard ANOVA software. Indeed, an R function to compute these values is described in Section 5. RMPBT methodology thus provides a simple mechanism for converting classical test statistics and p-values into Bayes factors. This methodology also makes explicit the alternative hypothesis that is implicitly being tested in a significance test, and provides practitioners with an estimate of the posterior probability that both the null and alternative hypotheses are true, given the prior probabilities they assign to the truth of each hypothesis.

4. RMPBTS FOR MODEL SELECTION

The g prior-RMPBTs described in the previous section maximize the power of a test between sharp hypotheses over a class of priors. By testing multiple models against a common benchmark model, the RMPBT can choose which of them significantly fit the data better than the benchmark, using values of g that allow each model to perform most favorably against the benchmark. This rationale underlies the model selection method that is based on g prior-RMPBTs, which we describe in Subsection 1. Subsection 2 provides numerical studies to evaluate the performance of this model selection routine against other proposed methods.

4.1 g Prior-RMPBTs in Model Selection

As before, let $\mathbf{y}$ represent a sample of size n drawn from the population. Define an $n \times p$ matrix of p candidate predictors $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p \end{bmatrix}$ for $\mathbf{y}$ and let the model space Ω consist of the set of models determined by all 2^p possible combinations of the columns of $\mathbf{X}$. Without loss of generality, we assume $\mathbf{X}$ has been centered, so that $\mathbb{1}^T\mathbf{X} = \mathbf{0}$. As in Liang et al. [23], we index the model space with the vector $\boldsymbol{\gamma}$, which is of length p and contains 1s and 0s to indicate the membership status of each candidate predictor in the model. We denote the model as $\mathcal{M}_{\boldsymbol{\gamma}}$ that consists of an intercept term, a linear combination of the columns of $\mathbf{X}$ indicated by γ, and an unknown error term. We write $\mathbf{X}_{\boldsymbol{\gamma}}$ for the $n \times p_{\boldsymbol{\gamma}}$ matrix of predictors in model $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ for the vector of regression coefficients of length $p_{\boldsymbol{\gamma}}$.

It is assumed that the data vector $\mathbf{y}$ is generated as a linear combination of an intercept and a certain set of predictors $\boldsymbol{\gamma}_t$, with error, where $\mathcal{M}_{\boldsymbol{\gamma}_t} \in \Omega$. In detail,

$$\mathbf{y} = \mathbb{1}_n\beta_0 + \mathbf{X}_{\boldsymbol{\gamma}_t}\boldsymbol{\beta}_{\boldsymbol{\gamma}_t} + \boldsymbol{\epsilon}, \tag{4.1}$$

where β_0 is an intercept, and, once again, $\boldsymbol{\epsilon}$ is a normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}\sigma^2$.

Rather than formally testing each candidate model in Ω against a benchmark, which would result in a list of rejected and a list of accepted models, we compute their Bayes factors for straightforward comparison. Each Bayes factor is weighted by the prior probability of the model being tested. For candidate model $\mathcal{M}_{\boldsymbol{\gamma}_i}$ and benchmark model $\mathcal{M}_{\boldsymbol{\gamma}_k}$, the Bayes factor $BF_{\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_k}$ is weighted by $\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})$. By rescaling the weighted Bayes factors, we obtain posterior probabilities, i.e.

$$\frac{BF_{\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_k}\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})}{\sum_{j=1}^{2^p} BF_{\boldsymbol{\gamma}_j,\boldsymbol{\gamma}_k}\pi(\mathcal{M}_{\boldsymbol{\gamma}_j})}, = \frac{\pi(\mathbf{y}|\mathcal{M}_{\boldsymbol{\gamma}_i})\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})}{\sum_{j=1}^{2^p} \pi(\mathbf{y}|\mathcal{M}_{\boldsymbol{\gamma}_j})\pi(\mathcal{M}_{\boldsymbol{\gamma}_j})} \tag{4.2}$$

$$= \pi(\mathcal{M}_{\boldsymbol{\gamma}_i}|\mathbf{y}) \tag{4.3}$$

By approaching the model selection problem as a comparison of Bayes factors between each candidate model $\mathcal{M}_{\boldsymbol{\gamma}_i}$ (for $i \in \{1, \ldots, 2^p\}$) and some common benchmark model $\mathcal{M}_{\boldsymbol{\gamma}_k}$, we find a justification for the use of g prior-RMPBTs: each candidate model's acceptability is measured by its improvement in fitting $\mathbf{y}$ over the benchmark model, and therefore the Bayes factor that measures each candidate model's improvement should be calibrated to maximize that model's improvement. To do anything else would imply comparing models on unequal footing; those whose Bayes factor is enhanced by the prior implicit in the Bayes factor would enjoy artificially inflated posterior probabilities relative to those whose Bayes factor is not.

There are several possible choices for the benchmark model $\mathcal{M}_{\boldsymbol{\gamma}_k}$, such as the full model $\mathcal{M}_{\mathbb{1}}$, the null model $\mathcal{M}_{\varnothing}$, and any other fixed model in Ω. To preserve coherency in the Bayes factors, several authors [23] have chosen the null model as the benchmark model; we also adopt this choice. Under this convention, the test of

model $\mathcal{M}_{\boldsymbol{\gamma}_i}$ can be expressed as follows:

$$\mathbf{y} = \mathbb{1}_n\beta_0 + \mathbf{X}_{\boldsymbol{\gamma}_i}\boldsymbol{\beta}_{\boldsymbol{\gamma}_i} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

$$H_1 : \boldsymbol{\beta}_{\boldsymbol{\gamma}_i}|g, \sigma^2 \sim \mathcal{N}(\mathbf{0}, g_{\boldsymbol{\gamma}_i}\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}_i}^T\mathbf{X}_{\boldsymbol{\gamma}_i})^{-1}), \quad \text{versus} \quad H_0 : \boldsymbol{\beta}_{\boldsymbol{\gamma}_i} = \mathbf{0},$$

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

Assuming that $\mathbf{X}_{\boldsymbol{\gamma}_i}$ is of full rank, and given some threshold γ, $g_{\boldsymbol{\gamma}_i}$ is given by Theorem 3.1.1 as

$$g_{\boldsymbol{\gamma}_i} = \arg\max_{g^*} \ (g^*)^{-1}\left[\gamma^{\frac{-2}{n-1}}(1+g^*)^{\frac{n-p_{\boldsymbol{\gamma}_i}-1}{n-1}} - 1\right]. \tag{4.4}$$

The resulting Bayes factor is given by (3.5):

$$BF_{\boldsymbol{\gamma}_i,\varnothing} = (1+g_{\boldsymbol{\gamma}_i})^{(n-p_{\boldsymbol{\gamma}_i}-1)/2}\left[1 + g_{\boldsymbol{\gamma}_i}\cdot\frac{n-p_{\boldsymbol{\gamma}_i}-1}{\hat{F}p_{\boldsymbol{\gamma}_i}+n-p_{\boldsymbol{\gamma}_i}-1}\right]^{-(n-1)/2}.$$

Specifying prior model probabilities $\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})$ requires some care due to the issue of multiplicity. Scott and Berger [49] describe the need for multiplicity correction to account for the false positives that result from comparing the posterior probabilities of all 2^p models and argue that such correction is best applied through the prior model probabilities. One fully Bayesian solution sets prior probabilities equal to

$$\pi(\mathcal{M}_{\boldsymbol{\gamma}_i}) = \frac{1}{p+1}\binom{p}{p_{\boldsymbol{\gamma}_i}}^{-1},$$

so that prior model mass is equally divided among the various sizes of model and among models of the same size.

Finally, the g prior-RMPBTs in this model selection method require the specifica-

tion of the evidence threshold γ. The values of $g_{\boldsymbol{\gamma}_i}$ in (4.4) are derived by evaluating the expression

$$\begin{aligned}
&\arg\max_{g^*} \mathbf{P}(BF_{\boldsymbol{\gamma}_i,\varnothing} > \gamma) \\
&\qquad = \arg\max_{g^*} \mathbf{P}\left(\frac{\pi(\mathcal{M}_{\boldsymbol{\gamma}_i}|\mathbf{y})}{\pi(\mathcal{M}_{\boldsymbol{\gamma}_\varnothing}|\mathbf{y})} > \tilde{\gamma}\right) \qquad \text{where } \tilde{\gamma} = \gamma \cdot \frac{\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})}{\pi(\mathcal{M}_{\boldsymbol{\gamma}_\varnothing})}.
\end{aligned}$$

Thus, the RMPBT can be defined either by setting an evidence threshold γ for the Bayes factor, or by setting an evidence threshold $\tilde{\gamma}$ for the posterior odds. In some cases it may be desirable to set the hyperparameter $g_{\boldsymbol{\gamma}_i}$ so as to maximize the probability that the posterior odds of a model exceed a threshold, after accounting for the prior odds against the model. For instance, if the prior odds against a non-null model $\mathcal{M}_{\boldsymbol{\gamma}_i}$ were 100:1, and if we wanted to set $g_{\boldsymbol{\gamma}_i}$ so as to maximize the probability that the posterior odds in favor of the model exceeded 4:1, this would imply setting γ at 400. That is, by fixing $\tilde{\gamma}$, the corresponding γ for this comparison will depend on $p_{\boldsymbol{\gamma}_i}$ through $\pi(\mathcal{M}_{\boldsymbol{\gamma}_i})$. As a result, a different Bayes factor evidence threshold γ is used for each model size. This strategy incorporates the multiplicity correction into the RMPBT calculation of the optimal γ.

Guidelines for $\tilde{\gamma}$ can be based on the schedule of evidence thresholds for γ given by Jeffreys [3] and modified by Kass and Raftery [7], which provide satisfactory demarcations between distinct weights of evidence in the posterior odds.

4.2 Numerical Comparisons of g Prior-based Model Selection Methods

To evaluate the performance of the g prior-RMPBT in a model selection problem, a simulation study was conducted. The parameters of the study closely follow those of the simulation study conducted by Liang et al. [23].

The candidate predictor matrix $\mathbf{X}$ was generated as an $n \times p$ matrix with $n = 100$ and $p = 50$ in such a way that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, the identity matrix, and $\mathbb{1}^T\mathbf{X} = \mathbf{0}$. In each independent simulation iteration, a "true" model $\mathcal{M}_{\boldsymbol{\gamma}_t}$ of size $p_{\boldsymbol{\gamma}_t}$ was randomly chosen, and values of $\boldsymbol{\beta}_{\boldsymbol{\gamma}_t}$ were generated from a $\mathcal{N}(\mathbf{0}, g_t\sigma^2\mathbf{X}_{\boldsymbol{\gamma}_t}^T\mathbf{X}_{\boldsymbol{\gamma}_t})$ distribution. Then data $\mathbf{y}$ were generated according to the model (4.1), where $\boldsymbol{\epsilon}$ was generated from a $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$ distribution, $\alpha = 2$, and $\sigma^2 = 1$. One-thousand simulation iterations were run for each of $p_{\boldsymbol{\gamma}_t} \in \{1, 5, 9, 13, 17, 21, 25\}$ and $g_t \in \{5, 25\}$.

Because the model space contains 2^{50} models, it was not practical to exhaustively search for the best model. We therefore sampled from the posterior model distribution using a simple birth-death Metropolis-Hastings sampler. This algorithm begins with a randomly-selected initial model $\mathcal{M}_{\boldsymbol{\gamma}_0}$. Then a new model $\mathcal{M}_{\boldsymbol{\gamma}'}$ is proposed through a birth/death process; i.e., one variable is randomly selected, and if it is already in the initial model, the proposed model omits it; if it is not in the initial model, the proposed model includes it. The proposed model is then accepted with probability $\min(1, a)$, where

$$a = \frac{BF_{\boldsymbol{\gamma}',\varnothing}}{BF_{\boldsymbol{\gamma}_0,\varnothing}} \cdot \frac{\pi(\boldsymbol{\gamma}')}{\pi(\boldsymbol{\gamma}_0)}.$$

The proposal density is symmetric, so that it does not contribute weight to a. If the proposal is not accepted, the model $\boldsymbol{\gamma}_0$ is accepted in its place. This sampler repeats 10,000 iterations plus 1,000 more for a burn-in period, which are discarded. Convergence was assessed using a trace plot of model size. Acceptance rates for all methods average roughly 10%.

If convergence is roughly achieved after the burn-in, the resulting draws approximately represent draws from the posterior distribution on model space. The empirical posterior probabilities were calculated on each model and the highest posterior

probability (HPP) model was recorded. The maximum *a posteriori* (MAP) estimate of $\boldsymbol{\beta}$ was taken from this model, which is denoted $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_{\text{HPP}}}$, and the selected model's estimation performance was evaluated under squared error loss:

$$\text{MSE}_{\boldsymbol{\gamma}_{\text{HPP}}} = ||\mathbf{X}_{\boldsymbol{\gamma}_{\text{t}}}\boldsymbol{\beta}_{\boldsymbol{\gamma}_{\text{t}}} - \mathbf{X}_{\boldsymbol{\gamma}_{\text{HPP}}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_{\text{HPP}}}||^2$$

The mean MSE was calculated over the 1,000 simulation iterations, and plotted as a function of p_{t} for each model selection method in Figure 4.1.

In addition to HPP, performance was also evaluated for median posterior probability model (MPP) and Bayesian model averaging (BMA). MPP selects all variables with a posterior probability greater than or equal to 0.5, where the posterior probability of variable i is equal to

$$\begin{aligned}\mathbf{P}(\boldsymbol{\gamma}^{(i)} = 1|\mathbf{Y}) &= \mathbf{E}(\mathbf{1}_{\boldsymbol{\gamma}^{(i)}=1}|\mathbf{Y}) \\ &= \sum_{j=1}^{2^p} \mathbf{1}_{\boldsymbol{\gamma}_j^{(i)}=1}\pi(\mathcal{M}_{\boldsymbol{\gamma}_j}|\mathbf{Y}),\end{aligned}$$

and the MAP estimate is likewise taken from the selected model. BMA provides the estimate of $\boldsymbol{\beta}$ given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_{\text{BMA}}} = \sum_{j=1}^{2^p} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_j}\pi(\mathcal{M}_{\boldsymbol{\gamma}_j}|\mathbf{Y}),$$

where within each model $\boldsymbol{\gamma}_j$, the estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_j}$ is the MAP.

The methods compared consist of the JZS, UIP, EBL, hyper-g, and RMPBT priors, together with the oracle model, which is based on the least-squares estimate taken from the data-generating model. The MAP estimate for JZS is obtained through a Laplace approximation, as described in Liang et al. [23] and implemented in the R package "BAS" [35], with hyperparameter a set equal to 3. The RMPBT

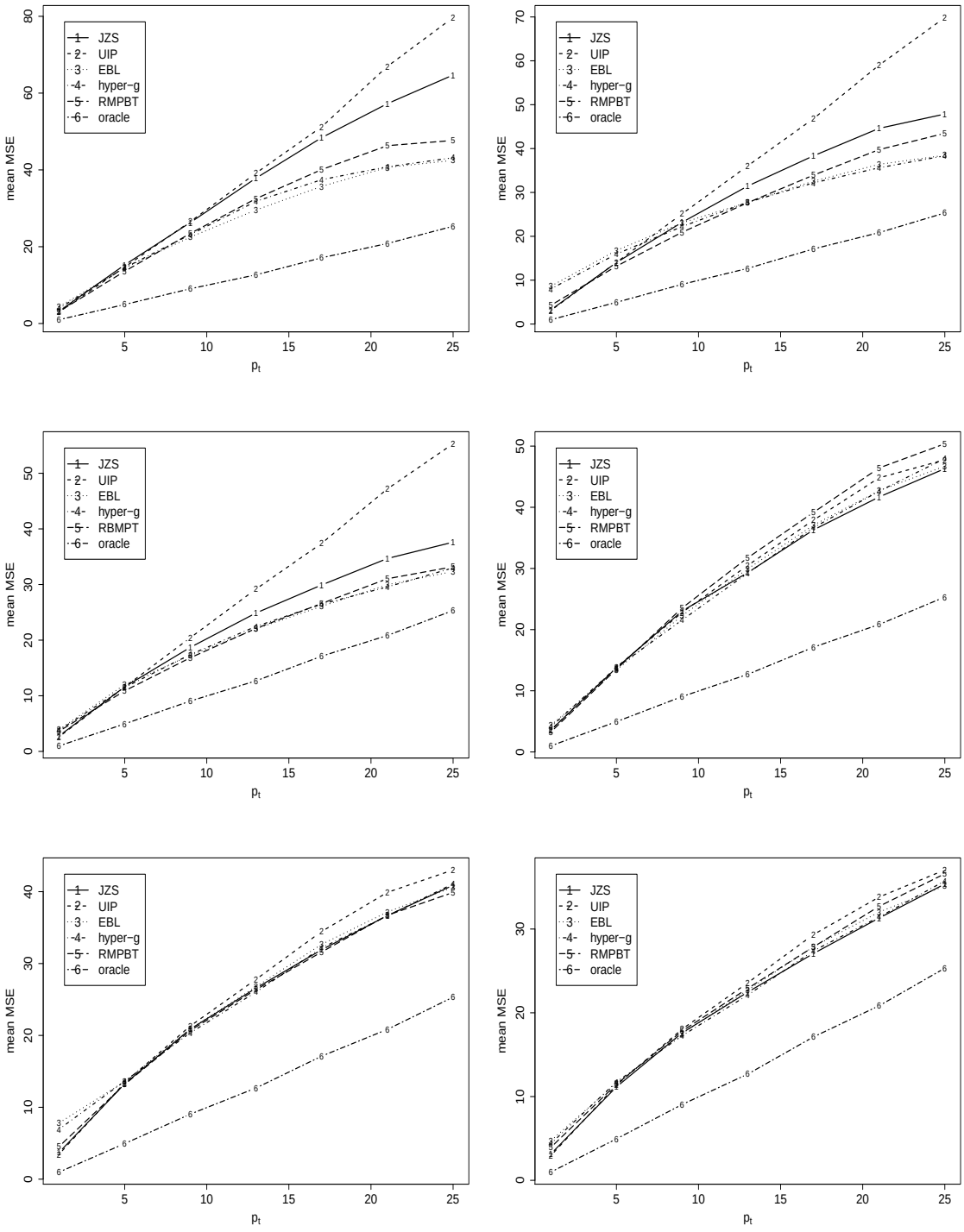

Figure 4.1: Simulation results of 6 methods. The y-axis is the mean MSE of the MAP estimate from the highest posterior model. The x-axis is the size of the data-generating model. Left panel: $g_t = 5$; Right panel: $g_t = 25$.

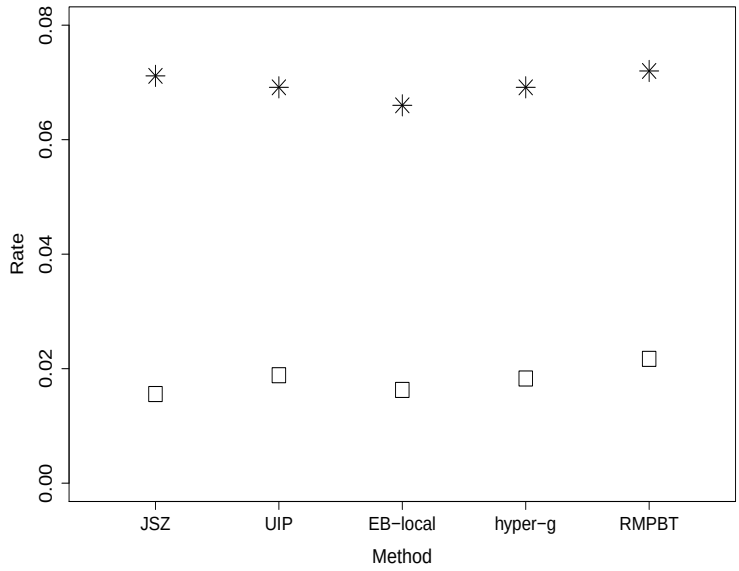

Figure 4.2: Simulation results of 5 methods (oracle not shown). Each point plots the percentage of the time that its method selected the correct model. Asterisks denote simulations with $g_{\mathrm{t}} = 25$, while squares denote simulations with $g_{\mathrm{t}} = 5$.

is derived using an evidence threshold $\tilde{\gamma}$ of 20, which is, once again, the minimum threshold for "strong" evidence of a Bayes factor according to Kass and Raftery [7], though now it is applied to posterior probabilities.

In terms of estimation error, Figure 4.1 shows that the RMPBT method performs slightly worse than model selection based on the the hyper-g prior and the EBL. It appears this difference in performance does not depend greatly on g_t. In contrast, model selection based on RMPBTs appears to do slightly better than selection based on the JZS and UIP priors. UIP is the worst-performing prior for most sizes of p_t.

We also examined the probability that each method correctly chose the true model. These probabilities are depicted in Figure 4.2. The procedures performed roughly similarly according to this criterion.

Although the RMPBT had slightly larger estimation errors than the hyper-g and

EBL methods in this simulation study, it appears to compete well in selecting the correct model. On the other hand, the RMPBT method is much easier to implement and requires an order of magnitude less computation than the hyper-g prior, and it has a much simpler and more direct interpretation than does the EBL method, while providing an explicit form of the alternative hypotheses that is being tested. In routine applications, these advantages may offset the slight increase in estimation error associated with the RMPBT, and make it an attractive alternative for practitioners.

5. AN R PACKAGE FOR MOST POWERFUL BAYESIAN TESTING

The tests described in [2, 1] and in Section 3 have been implemented in an R package for distribution through the CRAN repository. This package, entitled "MPBT", makes UMPBTs and RMPBTs easily accessible for R users. They are written to invoke syntax and arguments as similar as possible to functions that implement the corresponding classical tests, and provide both the Bayesian and classical test results as output. The Bayesian tests can be performed either by specifying the γ threshold, or in cases where an equivalence between the Bayesian test and the frequentist test exists, by specifying the frequentist threshold α. In this way, they facilitate the reporting of Bayes factors in conjunction with p-values.

In this section we describe the functions in the MPBT package, including their arguments, their output values, and important details. Subsection 1 describes the data sets that are included in this package, which are used to illustrate the package's functions. Subsection 2 treats functions designed for tests of one-parameter exponential family models. Subsection 3 describes functions that test the regression coefficients in the general linear model.

There are several commonalities between the various functions in this package. First, the output for the Bayesian test is always given as the Bayes factor in favor of the alternative hypothesis,

$$BF_{10} = \frac{m(\mathbf{y}|H_1)}{m(\mathbf{y}|H_0)}.$$

Given this quantity, the conversion to posterior odds is straightforward using the identity

$$\frac{\mathbf{P}(H_1|\mathbf{y})}{\mathbf{P}(H_0|\mathbf{y})} = BF_{10} \cdot \frac{\mathbf{P}(H_1)}{\mathbf{P}(H_0)}.$$

The advantage of reporting the Bayes factor over directly reporting the posterior odds is that each consumer of the test result is able to calculate custom posterior odds based on his or her subjective assessment of the prior odds.

Second, in all cases the two-sided UMPBTs or RMPBTs have not been derived, if they exist at all. Hence, a call to any function with a two-sided test will only return results for the frequentist test.

Third, in many cases the UMPBT or RMPBT is not available in closed form, so that a numerical algorithm must be used to estimate it. In such cases, the results of the numerical search are included among the output of the test, although they are not printed when using the `print` method defined for the class of objects that these functions return. These numerical search results are viewable through accessing the list of objects returned by each function through the `names` function.

Fourth, in the testing situations represented by some of these functions it is not obvious what the default frequentist test would be. In these cases, we will describe the particular frequentist test implemented.

5.1 MPBT Package Data Sets

The MPBT package includes seven data sets for use in illustrating the use of the various functions. The first of these is the `batteries` data set from [50]. It consists of data from a CRD experiment on the effect of ambient temperature on battery life. The data set contains 5 observations of a response variable `life`, which represents the battery life in hours, at each of 6 levels of the explanatory variable `temp`, which represents the ambient temperature of the battery's environment in degrees Fahrenheit.

The second data set is the `rubber` data set, also taken from [50]. This data results from a $3 \times 3 \times 4$ crossed factorial designed experiment. The first treatment, `lab`,

has 4 levels and indicates at which of 4 laboratories the measurements were made. The second treatment, `temp`, has 3 levels and represents the target temperature in degrees Fahrenheit of batches of rubber. The last treatment `mix`, has 3 levels and indicates which of three different mixing procedures were used on a rubber batch. The response variable, `time`, represents the time, in minutes, for the batch of rubber to solidify. There are 2 replicates of this experiment, so that there are 72 different `time` measurements total.

The third data set is the `rainfall` data set, taken from [51]. It provides the maximum daily rainfall (in mm) in each of 47 consecutive years for Turramurra, Sydney, Australia. This data is assigned to the variable `rain`.

The fourth data set is the `London` data set, also taken from [51]. It provides the number of times 576 different grid squares in South London, each 1/16 km^2 in area, were hit by bombs during World War II. This data is assigned to the variables `hits`.

The fifth data set is the `bearings` data set, also taken from [51]. It provides the measured diameter in microns of 10 randomly selected ball bearings from each of two production lines, `line1` and `line2`.

The sixth data set is the `health` data set, taken from [52]. This data was taken from 26 randomly selected males ages 25-30. Researchers measured each subject's weight in lbs. (`weight`) and his systolic blood pressure in mm Hg (`systolic`).

The last data set is the `pressure` data set, also taken from [52]. This data reports an experiment done on a bubble column with a screen plate. The response variable, `drop`, reports a dimensionless factor for the pressure drop through a bubble cap. The explanatory variables are `velocity`, the superficial fluid velocity of the gas in cm/s, `viscosity`, the kinematic viscosity, `mesh`, the mesh opening in cm, and `relationship`, a dimensionless measure of the relationship between the superficial fluid velocity of the gas and the superficial fluid velocity of the liquid.

5.2 Tests of One-Parameter Exponential Family Models

In [2, 1], Johnson derives the UMPBTs for several special cases of the one-parameter exponential family model. These are: the test of a binomial probability (when the sample size is fixed and known), the test of an exponential distribution scale parameter, the test of a negative binomial probability (when the target number of successes is fixed and known), the test of a Poisson rate parameter, and the test of a normal variance (when the mean is known). In addition, although it is not an exponential family model, we will also discuss here the test of a χ^2_1 noncentrality parameter.

5.2.1 Tests of a Binomial Probability

The function `binom_mpb` tests the probability parameter p in a sequence of Bernoulli trials against an alternative hypothesis when the sample size is fixed and known.

The proper syntax for the `binom_mpb` function is

```
binom_mpb(x, n, p0, gamma,
          alternative=c("two.sided", "less", "greater"))
```

where `x` is the observed number of successes, `n` is the sample size, `p0` is the value of p under H_0, and the γ threshold `gamma` must be user-specified for the Bayesian test. A call to the two-sided test only returns the frequentist test results. The frequentist p-value is the sum of discrete probability masses.

For an example we test the one-sided claim that a coin is fair after observing a string of 10 consecutive heads, with $\gamma = 20$.

```
> binom_mpb(x=10, n=10, p0=0.5, gamma=20, alternative='greater')
       BF      p.value
 243.4408 0.0009765625
```

```
The user-specified gamma is 20.
```

5.2.2 Tests of an Exponential Distribution Scale Parameter

The function `exp_mpb` tests the scale parameter μ in an exponential distribution against an alternative hypothesis.

The proper syntax for the `exp_mpb` function is

```
exp_mpb(x, mu0, gamma, alternative=c("two.sided", "less", "greater"))
```

where `x` is a numeric vector of observed data, `mu0` is the value of μ under H_0, and the γ threshold `gamma` must be user-specified for the Bayesian test. A call to the two-sided test only returns the frequentist test results. The frequentist test is performed using the fact that the pivot

$$\frac{(\bar{x} - \mu_0)}{\mu_0/\sqrt{n}}$$

converges to a standard normal distribution under H_0 as n increases.

For an example we test the claim that the scale parameter for the exponential distribution that best fits the `rainfall` data set is less than 1,500, with $\gamma = 20$.

```
> exp_mpb(x=rainfall$rain, mu0=1500, gamma=20, alternative='less')
        BF   p.value
 0.1401655 0.2748396


 The user-specified gamma is 20.
```

5.2.3 Tests of a Negative Binomial Probability

The function `negbinom_mpb` tests the probability parameter p in a sequence of Bernoulli trials against an alternative hypothesis when the target number of successes

is fixed and known. The negative binomial parameterization used matches that used by the `dnbinom` function in R.

The proper syntax for the `negbinom_mpb` function is

```
negbinom_mpb(k, r, p0, gamma,
             alternative=c("two.sided", "less", "greater"))
```

where `k` is the observed number of failures, `r` is the target number of successes, `p0` is the value of p under H_0, and the γ threshold `gamma` must be user-specified for the Bayesian test. A call to the two-sided test only returns the frequentist test results. The frequentist p-value is the sum of discrete probability masses.

For an example we test the one-sided claim that a coin is fair after observing 10 heads before a single tail, with $\gamma = 20$.

```
> negbinom_mpb(k=0, r=10, p0=0.5, gamma=20, alternative='greater')
       BF      p.value
 176.3025 0.0009765625


 The user-specified gamma is 20.
```

5.2.4 Tests of a Poisson Rate Parameter

The function `poisson_mpb` tests the rate parameter λ in a poisson distribution against an alternative hypothesis.

The proper syntax for the `poisson_mpb` function is

```
poisson_mpb <- function(x, lambda0, gamma,
                        alternative=c("two.sided", "less", "greater"))
```

where `x` is a numeric vector of observed data, `lambda0` is the value of λ under H_0, and the γ threshold `gamma` must be user-specified for the Bayesian test. A call to the two-sided test only returns the frequentist test results. The frequentist test is performed using the fact that the pivot

$$\frac{(\bar{x} - \lambda_0)}{\sqrt{\lambda_0/n}}$$

converges to a standard normal distribution under H_0 as n increases.

For an example we test the one-sided claim that the rate parameter for the poisson distribution that best fits the `London` data set is 1, with $\gamma = 20$.

```
> poisson_mpb(x=London$hits, lambda0=1, gamma=20, alternative="less")
       BF    p.value
 2.762978 0.05208128


 The user-specified gamma is 20.
```

5.2.5 Tests of a Normal Variance Parameter

The function `normalvar_mpb` tests the variance parameter σ^2 in a normal distribution against an alternative hypothesis when the mean is known.

The proper syntax for the `normalvar_mpb` function is

```
normalvar_mpb <- function(x, mu, s0, gamma, alternative=
c("two.sided", "less", "greater"))
```

where `x` is a numeric vector of observed data, `mu` is the mean of the population, `s0` is the value of σ^2 under H_0, and the γ threshold `gamma` must be user-specified for the

Bayesian test. A call to the two-sided test only returns the frequentist results. The frequentist test is performed using the fact that the pivot

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2} \sim \chi_n^2$$

under H_0.

For an example we test the one-sided claim that the variance parameter for the normal distribution with mean 7 that best fits the logarithm of the `rainfall` data set is 0.25, with $\gamma = 20$.

```
> normalvar_mpb(x=log(rainfall$rain), mu=7, s0=0.25, gamma=20,
>               alternative="less")
         BF   p.value
 0.01629677 0.5036909


 The user-specified gamma is 20.
```

5.2.6 Tests of a χ_1^2 Noncentrality Parameter

The function `chisq_mpb` tests the noncentrality parameter λ in a χ_1^2 distribution against a one-sided alternative hypothesis. Although this test is not a special case of testing a one-parameter exponential family model, we include it in this subsection because it was described among certain of the other tests mentioned here in [1].

The proper syntax for the `chisq_mpb` function is

```
chisq_mpb(x, gamma)
```

where `x` is a single numeric observation and the γ threshold `gamma` must be user-specified for the Bayesian test. The frequentist p-value is the area to the right of x under a χ_1^2 density curve.

For an example we perform a χ^2 test of independence on a 2×2 contingency table, where the test statistic equals 6, with $\gamma = 20$.

```
> chisq_mpb(x=6, gamma=20)
       BF    p.value
 9.859353 0.01430588

 The user-specified gamma is 20.
```

5.3 Tests of Regression Coefficients in the General Linear Model

The functions in this section perform tests of all or some of the regression coefficients in the general linear model or some special case of it. We begin by discussing the test of one or two normal means, followed by general ANOVA tests. Finally, we provide a function for testing in the general linear model setting. The tests in these functions were described variously in [2, 1] and Section 3.

5.3.1 Tests of One- and Two-sample Normal Means

The function `normalmean_mpb` performs one- and two-sample z and t tests, i.e. tests of normal mean(s) where the variance σ^2 either is known *a priori* or is not. The one- and two-sample z tests are described in both [2, 1]. Both of these sources also provide an approximate UMPBT for the one-sample t-test, albeit their tests are slightly different. For the package we implement the approximate UMPBT found in [1], which is given for both one- and two-sample tests. We also implement the RMPBTs for the one- and two-sample RMPBTs t tests found in Section 3.

The proper syntax for the `normalmean_mpb` function is

```
normalmean_mpb(x, y=NULL, sigma2=NULL, mu0=0, gamma, alpha=NULL,
 paired=FALSE, var.equal=FALSE,
```

```
 method=c("restricted","approximate"),
 alternative=c("two.sided", "less", "greater"))
```

where `x` is a numeric vector of observed data, `y` is an optional numeric vector of data, `sigma2` is the known variance (for z tests only), and `mu0` is the value of μ under H_0 (for one-sample tests only). The γ threshold may either be user-specified for `gamma`, or calculated based on an α threshold by specifying `alpha` and leaving `gamma` unspecified (which prompts the function to match the frequentist and Bayesian tests' rejection regions). In addition, the `paired` option can be used to perform paired two-sample tests and the `var.equal` option can be used to perform either the two-sample z-test with unequal variances or the Welch-Satterthwaite approximation to the two-sample t-test, although there will be no corresponding Bayesian test performed. A call to the two-sided test, likewise, only returns the frequentist results. The `method` argument selects between the RMPBT (`"restricted"`) and the approximate UMPBT (`"approximate"`). All frequentist tests are computed in the usual way.

We provide four examples of using the `normalmean_mpb` function. The first illustrates the test of a normal mean when the variance σ^2 is known (the one-sample z-test). We do this by testing the one-sided claim that the mean of the normal distribution with variance 0.08 which best fits the `bearings$line1` data set is 1, with $\gamma = 20$.

```
> normalmean_mpb(x=bearings$line1, sigma2=0.08, mu0=1, gamma=20,
>                alternative="greater")
      BF    p.value
 10.1087 0.01504188
```

```
The user-specified gamma, 20, corresponds to an alpha of 0.0072.
```

The second example illustrates the test of the equality of two means when the common variance σ^2 is known (the two-sample z-test). We do this by testing the one-sided claim that the difference of the means of the normal distributions with variance 0.10 which best fit the `bearings` data set is 0, with γ set by matching the Bayesian test's rejection region to that of an $\alpha = 0.05$ level frequentist test. Note that, consistent with the two-sample z-test described in [1], the argument `x` should correspond to the sample with the lower mean under H_1 and the argument `y` should correspond to the sample with the higher mean under H_1, since the right-sided test is being performed.

```
> normalmean_mpb(x=bearings$line1, y=bearings$line2, sigma2=0.10,
>                alpha=0.05, alternative="greater")
       BF    p.value
 3.043393 0.06692821

 A gamma of 3.8681 corresponds to the user-specified alpha 0.05.
```

The third example illustrates the test of a normal mean when the variance is unknown (the one-sample t-test). We do this by testing the one-sided claim that the mean of the normal distribution which best fits the `bearings$line1` data set is 1, with $\gamma = 20$. We opt to do an approximate UMPBT.

```
> normalmean_mpb(x=bearings$line1, mu0=1, gamma=20,
>                method="approximate", alternative="greater")
      BF    p.value
 7.07438 0.03163275
```

```
The user-specified gamma, 20, approximately corresponds to an alpha
of 0.0119.
```

The last example illustrate the test of the equality of two means when the common variance σ^2 is unknown (the two-sample t-test). We do this by testing the one-sided claim that the difference of the means of the normal distributions with common variance σ^2 which best fit the `bearings` data set is 0, with γ set by matching the Bayesian test's rejection region to that of an $\alpha = 0.05$ level frequentist test. We opt to use the `"restricted"` method. Again, the argument `x` should correspond to the sample with the lower mean under H_1 and the argument `y` should correspond to the sample with the higher mean under H_1, since the right-sided test is being performed.

```
> normalmean_mpb(x=bearings$line1, y=bearings$line2, alpha=0.05,
>                alternative="greater")
      BF    p.value
 0.9111004 0.1066991


 A gamma of 2.2874 corresponds to the user-specified alpha 0.05.
```

5.3.2 General ANOVA Tests

The function `aov_mpb` is a subsidiary function for the more general function `lm_mpb` (described below) in analogy to the wrapper function `aov` which exists for the more general function `lm`. It performs ANOVA tests using the RMPBTs described in Section 3.

The proper syntax for the `aov_mpb` function is

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

where the `formula` argument is an object of the R `formula` class and the `data` argument is an object of the R `data.frame` class. The γ threshold may either be user-specified for `gamma`, or calculated based on an α threshold by specifying `alpha` and leaving `gamma` unspecified (which prompts the function to match the frequentist and Bayesian tests' rejection regions).

We provide two examples of using the `aov_mpb` function. First we use the `batteries` data set to test that the mean lifetimes for the five temperature groups are equal. We test at the $\alpha = 0.05$ level and the function provides the corresponding γ threshold for the factor `temp`.

```
> aov_mpb(life~temp,data=batteries,alpha=0.05)
           BF    gamma      p.value
temp 46083.43 3.064042 3.146185e-13

 The user-specified alpha is 0.05.
```

As another example, we use the `rubber` data set. We test the main effects, two-way interactions, and three-way interaction of the factors `temp`, `lab`, and `mix`. Again we test at the $\alpha = 0.05$ level and the function provides the corresponding γ threshold for each test.

```
> aov_mpb(time~temp*lab*mix,data=rubber,alpha=0.05)
                        BF    gamma      p.value
temp          1.643114e+09 2.594327 8.143344e-54
lab           1.409900e+06 2.818816 6.208338e-15
mix           9.583649e+04 2.594327 3.569071e-10
temp:lab      3.649044e+03 3.180965 4.722377e-08
```

```
temp:mix     1.357029e+03 2.972724 1.150316e-06
lab:mix      1.164397e+02 3.180965 1.394283e-04
temp:lab:mix 6.779397e+01 3.510330 3.675729e-04

 The user-specified alpha is 0.05.
```

5.3.3 Tests of Coefficients in the General Linear Model

Moving toward greater generality, we finally consider the function `lm_mpb`, which implements the RMPBT of linear regression coefficients with unknown variance developed in detail in Section 3.

The proper syntax for the `lm_mpb` function is

```
lm_mpb(formula, data=NULL, gamma, alpha=NULL)
```

where the `formula` argument is an object of the R `formula` class and the `data` argument is an object of the R `data.frame` class. The γ threshold may either be user-specified for `gamma`, or calculated based on an α threshold by specifying `alpha` and leaving `gamma` unspecified (which prompts the function to match the frequentist and Bayesian tests' rejection regions).

We provide two examples of using the `lm_mpb` function. In the first example, we use the `health` data set to test for the significance of the slope for `weight` in the simple linear regression of `systolic` on `weight`. We specify a γ threshold for `gamma` of 20.

```
> lm_mpb(systolic~weight, data=health, gamma=20)
             BF      p.value      alpha
weight 5490.473 3.591105e-06 0.00306274
```

```
 The user-specified gamma is 20.
```

The second example illustrates multiple linear regression with the `pressure` data set. We test for the significance of the slopes of the four explanatory variables in modeling `drop`, the response. We specify an α threshold for `alpha` of 0.05.

```
> lm_mpb(drop~velocity+viscosity+mesh+relationship, data=pressure,
>        alpha=0.05)
                         BF    gamma      p.value
velocity     1.045076e+00 2.165836 1.652960e-01
viscosity    2.533954e+08 2.165836 3.779436e-15
mesh         1.653112e+01 2.165836 2.316482e-03
relationship 2.183894e+00 2.165836 4.935047e-02

 The user-specified alpha is 0.05.
```

6. CONCLUSION AND DISCUSSION

The conditions in Theorem 3.1.1 encompass many of the ANOVA, ANCOVA, and linear regression tests performed in practice, as well as the ubiquitous t tests. The RMPBTs described by this theorem apply to a large set of testing situations. These tests' principal virtue lies in three salient features: the objective and default alternative hypotheses they define, the supplemental information they provide to frequentist test results, and the power-maximizing criterion that motivates their definition. We discuss each feature briefly.

First, Section 2 reviewed the literature on Bayesian hypothesis testing and model selection, finding that in the former area there exists no objective method for defining alternative hypotheses in general, while in the latter area there exists no default method for specifying g in Zellner's g prior. In a testing situation, once the decision is made to apply a π-RMPBT and γ is specified, the alternative hypothesis is completely determined. Likewise in the model selection situation, once the decision is made to utilize a g prior-RMPBT and γ is chosen, g is given by a formula. In both cases, the principal source of subjectivity is the selection of a general method. We argue that the other two features of RMPBTs, described proximately, largely obviate this subjectivity by putting forward additional rationale supporting their usage.

Second, by calibrating γ to provide an α-level test, the g prior-RMPBT provides an alternative quantification of the evidence against the null hypothesis, as well as a description of the weight of evidence in favor of it. For this reason, we view the g prior-RMPBT as a supplement to the classical F-test. Under an assumption of equipoise, the g prior-RMPBT provides an objective estimate of the posterior probabilities of the null and alternative hypotheses, quantities that in many cases

are of primary interest to practitioners and are more interpretable to consumers. Additionally, it may of value to some users that the RMPBT explicitly declares the alternative hypothesis being tested.

Third, RMPBTs and UMPBTs optimize over all possible values of $\boldsymbol{\theta}_t$, the data-generating value of the parameter being tested. For a fixed γ these tests maximize statistical power, in analogy to the frequentist uniformly most powerful tests. This optimization results in a maximization of the probability that the Bayes factor exceeds γ for values of $\boldsymbol{\theta}_t$ that satisfy H_0. Philosophical objections to this facet of these tests should be balanced against the fact that the probability of a false rejection is still controllable through the specification of γ.

The upshot of these three features is that RMPBTs provide a broadly applicable, defensible, and coherent methodology for performing Bayesian hypothesis tests and model selection. The software package described in Section 5 makes this methodology easily accessible.

We conclude with some observations on potential research directions in subsequent work. Although the model selection method described in Section 4 did not outperform other common methods, its computation burden was lighter than that of the two highest-performing methods (hyper-g and local empirical Bayes). It has a simpler justification than the empirical Bayes method and does not depend on the use of additional prior specifications like the hyper-g method. Additional research may be able uncover improvements to the method which further increase its competitiveness in simulation tests.

The results described in this article depend on the use of the Normal-Gamma g prior on model coefficients, which restricts their applicability. Although the g prior has found wide and extensive application in Bayesian model averaging [32] and model selection methods, RMPBTs may be sought for other classes of priors, including non-

local priors [53]. Finally, the extension of RMPBTs to non-linear models is currently under investigation.

REFERENCES

[1] Johnson VE (2013b) Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. USA* 110:19313–19317.

[2] Johnson VE (2013a) Uniformly most powerful Bayesian tests. *Ann. Statist.* 41:1716–1741.

[3] Jeffreys H (1961) *Theory of Probability* (Oxford University Press, Oxford, U.K.), 3rd edition.

[4] Jeffreys H (1946) An invariant form for the prior probability in estimation problems. *P. Roy. Soc. Lond. A Mat.* 186:453–461.

[5] Laplace PS (1995) *Pierre-Simon Laplace Philosophical Essay on Probabilities* (Springer-Verlag New York, Inc., New York, NY).

[6] Bernardo J (1979) Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 41:113–147.

[7] Kass R, Raftery A (1995) Bayes factors. *J. Amer. Statist. Assoc.* 90:773–795.

[8] Smith AFM, Spiegelhalter DJ (1980) Bayes factors and choice criteria for linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 42:213–220.

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

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

[11] Solari F, Liseo B, Sun D (2008) Some remarks on Bayesian inference for one-way ANOVA models. *Ann. Inst. Stat. Math.* 60:483–498.

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

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

[14] Bernardo J (1980) A Bayesian analysis of classical hypothesis testing. *Trabajos de Estadística Y de Investigación Operativa* 31:605–647.

[15] Dickey JM (1977) Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* 72:138–142.

[16] Berger JO, Sellke T (1987) Testing a point null hypothesis: irreconcilability of p-values and evidence. *J. Amer. Statist. Assoc.* 82:112–122.

[17] Berger JO, Delampady M (1987) Testing precise hypotheses. *Statist. Sci.* 2:317–335.

[18] Berger J, Boukai B, Wang Y (1997) Unified frequentist and Bayesian testing of a precise hypothesis. *Statist. Sci.* 12:133–160.

[19] Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds Goes PK, Zellner A (North-Holland/Elsevier, Amsterdam), pp 233–243.

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

[21] Fernández C, Ley E, Steel MF (2001) Benchmark priors for Bayesian model averaging. *J. Econometrics* 100:381–427.

[22] Hannan EJ, Quinn BG (1979) The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* pp 190–195.

[23] Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* 103:410–423.

[24] Zellner A, Siow A (1980) Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, eds Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (University of Valencia Press, Valencia), pp 585–603.

[25] Guo R, Speckman PL (2009) Bayes factor consistency in linear models. In *The 2009 International Workshop on Objective Bayes Methodology.*

[26] Wasserman L (2000) Bayesian model selection and model averaging. *J. Math. Psych.* 44:92–107.

[27] O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how, and which. *Bayesian Analysis* 4:85–118.

[28] Spiegelhalter DJ, Smith AFM (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 44:377–387.

[29] O'Hagan A (1995) Fractional Bayes factors for model comparison. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 57:99–138.

[30] Berger JO, Pericchi LR (1996a) The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* 91:109–122.

[31] Berger JO, Pericchi LR (1996b) The intrinsic Bayes factor for linear models (with discussion). In *Bayesian Statistics 5. Proceedings of the Fifth Valencia*

International Meeting, June 5-9, 1994, eds Bernardo JM et al. (Oxford University Press, London).

[32] Feldkircher M, Zeugner S (2009) Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging., (IMF), Working paper WP/09/202.

[33] Kruschke JK, Meredith M (2015) *Package 'BEST'.* Available at http://cran.r-project.org/web/packages/BEST/BEST.pdf.

[34] Morey RD, Rouder JN, Jamil T (2015) *Package 'BayesFactor'.* Available at http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf.

[35] Clyde MA, Ghosh J, Littman ML (2011) Bayesian adaptive sampling for variable selection and model averaging. *J. Comput. Graph. Statist.* 20:80–101.

[36] Clyde M, Littman M (2005) *Package 'BAS'.* Available at http://cran.r-project.org/web/packages/BAS/BAS.pdf.

[37] Amini SM, Parmeter CF (2011) Bayesian model averaging in R. *J. Econ. Soc. Meas.* 36:253–287.

[38] Zeugner S (2011) Bayesian model averaging with BMS. Available at http://cran.rproject.org/web/packages/BMS/vignettes/bms.pdf.

[39] Feldkircher M, Zeugner S (2013) *Package 'BMS'.* Available at http://cran.r-project.org/web/packages/BMS/BMS.pdf.

[40] Raftery AE, Hoeting JA, Volinsky CT, Painter I, Yeung KY (2014) *Package 'BMA'.* Available at http://cran.r-project.org/web/packages/BMA/BMA.pdf.

[41] Raftery AE, Hoeting JA, Volinsky CT, Painter I, Yeung KY (2005) BMA: an R package for Bayesian model averaging. *R News* 5:2–7.

[42] Bayarri MJ, Berger JO, Forte A, García-Donato G (2012) Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* 40:1550–1577.

[43] Garcia-Donata G, Forte A (2015) *Package 'BayesVarSel'.* Available at http://cran.r-project.org/web/packages/BayesVarSel/BayesVarSel.pdf.

[44] Scheipl F (2015) *Package 'spikeSlabGAM'.* Available at http://cran.r-project.org/web/packages/spikeSlabGAM/spikeSlabGAM.pdf.

[45] Dey T (2013) modelSampler: An R tool for variable selection and model exploration in linear regression. *Journal of Data Science* 11:343–370.

[46] Rossell D, Cook JD, Telesca D, Roebuck P (2015) *Package 'mombf'.* Available at http://cran.r-project.org/web/packages/mombf/mombf.pdf.

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

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

[49] Scott JG, Berger JO (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* 38:2587–2619.

[50] Hicks C, Turner K (1999) *Fundamental Concepts in the Design of Experiments* (Oxford University Press, New York City, NY), 5th edition.

[51] Tamhane AC, Dunlop DD (2000) *Statistics and Data Analysis* (Prentice Hall, Upper Saddle River, NJ).

[52] Montgomery DC, Peck EA, Vining GG (2001) *Introduction to Linear Regression Analysis* (John Wiley & Sons, Inc., New York, NY), 3rd edition.

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

APPENDIX A

PROOFS OF THEOREMS

Proof of Theorem 3.1.1 Under the alternative hypothesis, $m_1(\mathbf{y})$, the marginal density is given by

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

where $\mathbf{P}_{\mathbf{X}_i} = \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T$. Under the null hypothesis, the marginal density is

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

Therefore, the Bayes factor in favor of the alternative is

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

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

$$\mathbf{P}\left(\frac{1-R_1^2}{1-R_0^2} < \frac{\gamma^{\frac{-2}{n-p_2-1}}(1+g)^{\frac{n-p-1}{n-p_2-1}}-1}{g}\right).$$

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

Proof of Theorem 3.1.2 The rejection region for the frequentist test is

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

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

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

which can be expressed as

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

where the constant c equals

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

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

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

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

By differentiating and equating to 0 the expression in (3.4), we obtain another expression for γ in terms of g:

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

Solving for g and γ completes the proof.

Proof of Corollary 3.1.3 We will show that the conditions given in this corollary satisfy the conditions in Theorem 3.1.1. Letting $\mathbf{X}_2 = \mathbf{0}$, $\boldsymbol{\beta}_2 = \mathbf{0}$, and $p_2 = 0$, define $p = p_1 = J - 1$. It is easily seen that $\mathbb{1}^T(\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^* = \mathbf{0}$. It only remains to show that $(\mathbf{I}_n - \mathbf{P}_{\mathbb{1}})\mathbf{X}_1^*$ is of full-column rank, or rank $J - 1$. A rearrangement of the rank-nullity theorem gives

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

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

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

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

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

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

The proof to the corollary follows from Theorems 3.1.1 and 3.1.2.

Proof of Corollary 3.1.4 The proof follows from Corollary 3.1.3 using $J = 2$ and the fact that the $1 - \alpha$ quantile from an F distribution with 1 and $n - 2$ degrees of freedom is equivalent to the square of the $1 - \alpha/2$ quantile of a t distribution with $n - 2$ degrees of freedom.

Proof of Theorem 3.1.5 Under the alternative hypothesis, the marginal density is given by

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

Under the null hypothesis, the marginal density is

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

Therefore, the Bayes factor in favor of the alternative is

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

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

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

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

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

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

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

which is equivalent to

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

where the constant c equals

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

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

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

Differentiating (3.11) and setting the result to zero leads to

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

Solving for g and γ completes the proof.