TEST (2021) 30:103–132
https://doi.org/10.1007/s11749-020-00704-4

ORIGINAL PAPER

On the prevalence of information inconsistency in normal linear models

Joris Mulder[1,2] · **James O. Berger**[3] · **Víctor Peña**[4] · **M. J. Bayarri**[5]

Received: 12 May 2019 / Accepted: 5 February 2020 / Published online: 20 February 2020

Abstract
Informally, 'information inconsistency' is the property that has been observed in some Bayesian hypothesis testing and model selection scenarios whereby the Bayesian conclusion does not become definitive when the data seem to become definitive. An example is that, when performing a t test using standard conjugate priors, the Bayes factor of the alternative hypothesis to the null hypothesis remains bounded as the t statistic grows to infinity. The goal of this paper is to thoroughly investigate information inconsistency in various Bayesian testing problems. We consider precise hypothesis tests, one-sided hypothesis tests, and multiple hypothesis tests under normal linear models with dependent observations. Standard priors are considered, such as conjugate and semi-conjugate priors, as well as variations of Zellner's g prior (e.g., fixed g priors, mixtures of g priors, and adaptive (data-based) g priors). It is shown that information inconsistency is a widespread problem using standard priors while certain theoretically recommended priors, including scale mixtures of conjugate priors and adaptive priors, are information consistent.

Keywords Bayes factors · Conjugate priors · Information inconsistency · Regression models

Mathematics Subject Classification 62F03 · 62F15 · 62A01

✉ Joris Mulder
j.mulder3@tilburguniversity.edu

1 Tilburg University, Tilburg, The Netherlands

2 Jheronimus Academy of Data Science, 's Hertogenbosch, The Netherlands

3 Duke University, Durham, USA

4 Baruch College Zicklin School of Business, New York, NY, USA

5 Universitat de Valéncia, Valencia, Spain

1 Introduction

When testing a hypothesis H_0 against an alternative hypothesis H_1, a common Bayesian tool is the Bayes factor, B_{10}, which quantifies the relative evidence (or odds) from the data for H_1 against H_0. A Bayes factor is called *information inconsistent* if, when the evidence for the alternative hypothesis appears to be overwhelming (in the sense that the observed effect under the alternative hypothesis becomes arbitrarily large), the Bayes factor converges to a constant $B^* < \infty$. This conflicting behavior, which already dates back to Jeffreys (1961), is also referred to as the information paradox (Liang et al. 2008). Note that we utilize the language of Bayes factors simply for convenience; everything could equivalently be stated in terms of posterior probabilities, e.g., there is information inconsistency if the posterior probability of H_1 is bounded away from 1 as the evidence for H_1 appears to be overwhelming.

Example 1 A typical example of an information inconsistent Bayes factor is when using Zellner's (1986) g prior for testing the regression coefficients in a linear regression model $\mathbf{y} = \gamma \mathbf{1}_n + \mathbf{X}_1\boldsymbol{\theta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{y}$ is a vector containing the n responses, γ is the intercept, $\mathbf{X}_1$ is a $n \times r_1$ matrix containing the explanatory variables, $\boldsymbol{\theta}$ is a vector with the r_1 unknown coefficients that are tested, σ^2 is the unknown error variance, $\mathbf{1}_n$ is a vector of length n with ones, $\mathbf{0}$ is a vector of zeros,[1] $\mathbf{I}_n$ is the identity matrix of size n, and N_n denotes a n-dimensional normal (or Gaussian) distribution. When testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \neq \mathbf{0}$ with the g prior, $\pi_0(\gamma, \sigma^2) \propto \sigma^{-2}$ and $\pi_1(\boldsymbol{\theta} \mid \gamma, \sigma^2) = N_{r_1}(\boldsymbol{\theta}|\mathbf{0}, g\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1})$ and $\pi_1(\gamma, \sigma^2) \propto \sigma^{-2}$, for some fixed $g > 0$, the Bayes factor goes to $(1+g)^{(n-r_1-1)/2} < \infty$ as the evidence against H_0 accumulates in the sense that $|\hat{\boldsymbol{\theta}}| \to \infty$, where $\hat{\boldsymbol{\theta}}$ denotes the least squares estimate of $\boldsymbol{\theta}$ and $|\cdot|$ denotes Euclidean norm of a vector (see also, Berger and Pericchi 2001). Furthermore, it has also been reported that the g prior is information inconsistent when testing one-sided hypotheses (Mulder 2014a).

In comparison with large sample inconsistency, which occurs when the evidence for the true hypothesis against another hypothesis does not go to infinity as the sample size grows, information inconsistency has not received much attention in the literature. In our view, both types of inconsistency are undesirable and should be avoided in general testing procedures. The goal of this paper is therefore to explore information inconsistency in the general setting of testing in the normal linear model with unknown variance. We will consider improper as well as proper priors; conjugate priors, scale mixtures of conjugate priors, independent priors, and adaptive priors; and precise null hypothesis testing, one-sided hypothesis testing, and multiple hypothesis testing. Throughout the paper, we also consider variations of Zellner's g prior (e.g., fixed g priors, mixtures of g priors, and adaptive (data-based) g priors) as this class of priors is commonly observed in the literature. We show that information inconsistency typically results when using 'standard' conjugate or independent semi-conjugate priors, while information consistency typically results when using more sophisticated scale mixture or adaptive priors. We also explore the practical consequences of information

[1] Throughout this paper, the symbol for the vector of zeros, $\mathbf{0}$, only receives an index to reflect its length when its length is not directly clear from the context.

consistency, by investigating when information inconsistency starts to manifest itself and finding the limiting value of the Bayes factor. Note that having an unknown variance is crucial; we are not aware of any information inconsistency results for testing in the normal linear model with known variance.

The paper is organized as follows. First the linear regression model with dependent errors and some notation are introduced (Sect. 2). Subsequently, Sect. 3 explores information consistency when testing a precise hypothesis using various prior specifications, followed by one-sided hypothesis tests in Sect. 4 and a multiple hypothesis test in Sect. 5. We end the paper with some conclusions and recommendations in Sect. 6.

2 The linear regression model with dependent errors

Throughout this paper, the focus shall be on the linear regression model with dependent errors,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}), \tag{1}$$

where the vector $\mathbf{y}$ of length n contains the responses, $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_K]$ is an $n \times K$ matrix containing the K predictor variables which are regressed on the K unknown regression coefficients in $\boldsymbol{\beta}$ ($n > K$), $\boldsymbol{\epsilon}$ is a normally distributed error vector, σ^2 is an unknown common variance, and $\boldsymbol{\Sigma}$ is a known positive definite matrix.

Three different types of hypothesis tests will be considered. First, we consider the classical null hypothesis test of a set of linear restrictions on $\boldsymbol{\beta}$ against an unrestricted alternative, i.e., $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0}_{r_1}$ versus $H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{0}_{r_1}$, where $\mathbf{R}$ is an $r_1 \times K$ matrix with known constants ($r_1 \leq K$). Second, we consider the equivalent one-sided hypothesis test of $H_0 : \mathbf{R}\boldsymbol{\beta} \leq \mathbf{0}_{r_1}$ versus $H_1 : \mathbf{R}\boldsymbol{\beta} \not\leq \mathbf{0}_{r_1}$, where "$\not\leq$" implies that at least one inequality goes to the other direction. Third, we briefly consider the multiple hypothesis test $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0}_{r_1}$ versus $H_1 : \mathbf{R}\boldsymbol{\beta} \leq \mathbf{0}_{r_1}$ (with $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}_{r_1}$ excluded) versus $H_2 : \mathbf{R}\boldsymbol{\beta} \not\leq \mathbf{0}_{r_1}$. The precise Bayesian hypothesis test of a set of linear restrictions was also investigated by Bayarri and García-Donato (2007). A Bayesian hypothesis test with combinations of equality and one-sided constraints was, for instance, considered by Mulder et al. (2010).

The model is reparametrized so that the linear combination of the parameters of interest, i.e., $\boldsymbol{\theta} = \mathbf{R}\boldsymbol{\beta}$, is perpendicular to the nuisance parameters, i.e., $\boldsymbol{\gamma} = \mathbf{D}\boldsymbol{\beta}$, i.e.,

$$\begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{R} \\ \mathbf{D} \end{bmatrix} \boldsymbol{\beta} = \mathbf{T}\boldsymbol{\beta},$$

where the $r_2 \times K$ matrix $\mathbf{D}$ contains $r_2 = K - r_1$ independent rows of $\mathbf{P}_{\mathbf{R}}^{\perp}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$, where the orthogonal projection matrix is given by $\mathbf{P}_{\mathbf{R}}^{\perp} = \mathbf{I}_K - \mathbf{R}'\left(\mathbf{R}\mathbf{R}'\right)^{-1}\mathbf{R}$. Subsequently, the model can be written as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta} + \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\mathbf{X}_1$ contains the first r_1 columns of $\mathbf{X}\mathbf{T}^{-1}$ that are regressed on $\boldsymbol{\theta}$ and $\mathbf{X}_0$ contains the remaining r_2 columns of $\mathbf{X}\mathbf{T}^{-1}$ that are regressed on $\boldsymbol{\gamma}$. The null hypothesis can then be written as $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \in \mathbb{R}^{r_1}$, and the one-sided

hypothesis test can be written as $H_0 : \boldsymbol{\theta} \leq \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \not\leq \mathbf{0}$. Thus, the design matrix under the precise null hypothesis H_0 is denoted by $\mathbf{X}_0$, and under the unconstrained alternative hypothesis H_1 in the precise test, it is denoted by $[\mathbf{X}_0\ \mathbf{X}_1]$. Further note that the ML estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are independent because $\left([\mathbf{X}\mathbf{T}^{-1}]'\boldsymbol{\Sigma}^{-1}[\mathbf{X}\mathbf{T}^{-1}]\right)^{-1} = \text{diag}\left(\left(\mathbf{X}_1'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\right)^{-1}, \left(\mathbf{X}_0'\boldsymbol{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\right)$ which is a direct consequence of the choice of $\mathbf{D}$.

Throughout this paper, the free parameters under a hypothesis have a hypothesis index to make it explicit that the parameters under different hypotheses have different interpretations and therefore different priors. For example, the population variances under H_0 and H_1 are denoted by σ_0^2 and σ_1^2, respectively. Also, $\hat{\boldsymbol{\theta}}$ will denote the maximum likelihood estimate of $\boldsymbol{\theta}$.

3 Testing a precise hypothesis

The following definition will be used for information inconsistency when testing a precise hypothesis.

Definition 1 A Bayes factor, B_{10}, is called information inconsistent for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \neq \mathbf{0}$ if there exists a sequence $\{\hat{\boldsymbol{\theta}}_i, i = 1, 2, \ldots\}$ that satisfies $|\hat{\boldsymbol{\theta}}_i| \to \infty$ as $i \to \infty$, for which the Bayes factor $B_{10} \leq B_{10}^* < \infty$.

For normal linear models, this definition is equivalent to the more general formulation using the likelihood ratio Λ_{10}, as proposed by Bayarri et al. (2012). The definition implies that an information consistent Bayes factor and the classical likelihood ratio test (using the usual F or t statistic) result in identical conclusions as $\Lambda_{10} \to \infty$.

3.1 Conjugate priors

In the conjugate case, the conditional prior of $\boldsymbol{\theta} \mid \sigma_1^2$ under H_1 has a multivariate normal distribution and the marginal prior of σ_t^2, for $t = 0$ or 1, has a scaled inverse Chi-squared distribution, resulting in

$$\begin{aligned} \pi_1(\boldsymbol{\theta}, \boldsymbol{\gamma}_1, \sigma_1^2) &= \pi_1(\boldsymbol{\theta} \mid \sigma_1^2) \times \pi_1(\boldsymbol{\gamma}_1) \times \pi_1(\sigma_1^2) \\ &\propto N_{r_1}(\boldsymbol{\theta}|\mathbf{0}, \sigma_1^2\boldsymbol{\Omega}) \times 1 \times \text{inv-}\chi^2(\sigma_1^2|s_1^2, \nu_1) \qquad (2) \\ \pi_0(\boldsymbol{\gamma}_0, \sigma_0^2) &= \pi_0(\boldsymbol{\gamma}_0) \times \pi_0(\sigma_0^2) \\ &\propto 1 \times \text{inv-}\chi^2(\sigma_0^2|s_0^2, \nu_0), \qquad (3) \end{aligned}$$

where s_0^2 and s_1^2 are prior scale parameters and ν_0 and ν_1 are prior degrees of freedom for the error variance under the two different hypotheses H_0 and H_1, respectively. The scaled inverse Chi-squared distribution is used (instead of the inverse gamma distribution) because of the natural relation between the prior degrees of freedom ν_t and the sample size n (Gelman et al. 2004). When setting the prior degrees of freedom equal to $\nu_t = 0$, we obtain the objective improper prior, $\pi_t(\sigma_t^2) \propto \sigma_t^{-2}$, for $t = 0$ or 1, and when additionally setting $\boldsymbol{\Omega} = g\left(\mathbf{X}_1'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\right)^{-1}$, we obtain Zellner's g prior.

The use of improper priors in testing for common "group invariant" parameters, such as the variances, is justified in Berger et al. (1998) and further discussed in the current testing problem in Bayarri et al. (2012). The conditional prior for $\boldsymbol{\theta}$ is centered at the null value of $\mathbf{0}$, as is common in testing and model uncertainty, but any other (fixed) centering could be used without affecting the results that follow.

Denoting the ML estimates by $\hat{\boldsymbol{\theta}} = \left(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1\right)^{-1} \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{y}$ and $\hat{\boldsymbol{\gamma}} = \left(\mathbf{X}_0' \boldsymbol{\Sigma}^{-1} \mathbf{X}_0\right)^{-1} \mathbf{X}_0' \boldsymbol{\Sigma}^{-1} \mathbf{y}$ and the sums of squares by $s_{\mathbf{y}}^2 = (\mathbf{y} - \mathbf{X}_1 \hat{\boldsymbol{\theta}} - \mathbf{X}_0 \hat{\boldsymbol{\gamma}})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}_1 \hat{\boldsymbol{\theta}} - \mathbf{X}_0 \hat{\boldsymbol{\gamma}})$, a standard calculation yields that the Bayes factor of H_1 against H_0, based on the conjugate priors in (2) and (3), is

$$B_{10} = C_1 \times \frac{\left(s_1^2 \nu_1 + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}' \left(\left(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1\right)^{-1} + \boldsymbol{\Omega}\right)^{-1} \hat{\boldsymbol{\theta}}\right)^{-(n+\nu_1-r_2)/2}}{\left(s_0^2 \nu_0 + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}' \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \hat{\boldsymbol{\theta}}\right)^{-(n+\nu_0-r_2)/2}}, \quad (4)$$

where the constant is

$$C_1 = \frac{(\nu_1/2)^{\nu_1/2} s_1^{\nu_1} \Gamma\left(\frac{\nu_0}{2}\right) \Gamma\left(\frac{n+\nu_1-r_2}{2}\right)}{(\nu_0/2)^{\nu_0/2} s_0^{\nu_0} \Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{n+\nu_0-r_2}{2}\right)} 2^{(\nu_1-\nu_0)/2} |\boldsymbol{\Omega} + (\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1}|^{-\frac{1}{2}} |\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1|^{-\frac{1}{2}}.$$

The Bayes factor depends on both $\hat{\boldsymbol{\theta}}$ and $s_{\mathbf{y}}^2$, which are independent. We will thus assume that $s_{\mathbf{y}}^2$ is fixed. The following result is immediate.

Lemma 1 *As* $|\hat{\boldsymbol{\theta}}| \to \infty$, *the Bayes factor in* (4) *satisfies* $B_{10} \to 0$ *if* $\nu_0 < \nu_1$; $B_{10} \to \infty$ *if* $\nu_0 > \nu_1$; *and if* $\nu_0 = \nu_1$,

$$B_{10} \leq C_1 \left(\limsup_{|\hat{\boldsymbol{\theta}}| \to \infty} \frac{\hat{\boldsymbol{\theta}}' \boldsymbol{X}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_1 \hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\theta}}' \left(\left(\boldsymbol{X}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_1\right)^{-1} + \boldsymbol{\Omega}\right)^{-1} \hat{\boldsymbol{\theta}}} \right)^{\frac{(n+\nu-r_2)}{2}}$$
$$= C_1 \ (1 + \lambda_{\max})^{(n+\nu-r_2)/2} < \infty,$$

where $\lambda_{\max}$ *is the largest eigenvalue of* $\boldsymbol{X}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_1 \boldsymbol{\Omega}$.

Remark 1 Setting $\nu_0 < \nu_1$ seems logical because it implies that the prior for σ_1^2 is more concentrated than the prior for σ_0^2 (consistent with a nonzero mean explaining some of the variation compared to a zero mean). This choice, however, results in a disastrously information inconsistent Bayes factor, with the conclusion being that the null hypothesis is certainly true when $|\hat{\boldsymbol{\theta}}| \to \infty$.

Remark 2 Setting $\nu_0 = \nu_1$ is the usual choice, which still results in an information inconsistent Bayes factor. Note that the prior degrees of freedom would be set to 0 in the objective Bayesian approach. The impact of this inconsistency will be discussed below for the special case of the univariate t test.

Remark 3 Setting $\nu_0 > \nu_1$ would not be a logical choice because the prior for σ_0^2 is then more concentrated in the tails than the prior for σ_1^2, even though the regression coefficient $\boldsymbol{\theta}$ under H_1 can explain some of the variation in the data. The resulting Bayes factor, however, is information consistent. A special case of this choice arises from setting the prior for the variance under H_0 to be proportional to the conditional prior of the variance given $\boldsymbol{\theta} = \mathbf{0}$ under H_1, i.e., $\pi_0(\sigma^2) = \pi_1(\sigma^2 \mid \boldsymbol{\theta} = \mathbf{0}) = \text{inv-}\chi^2(\sigma^2|\frac{\nu_1}{\nu_1+r_1}s_1^2, \nu_1 + r_1)$, so that $\nu_0 = \nu_1 + r_1$. The Bayes factor can then be expressed as the Savage–Dickey density ratio (Dickey 1971), $B_{10} = \frac{\pi_1(\boldsymbol{\theta}=\mathbf{0}|\mathbf{y})}{\pi_1(\boldsymbol{\theta}=\mathbf{0})}$, where the marginal prior and the posterior of $\boldsymbol{\theta}$ have a multivariate Student t distribution.

Remark 4 The definition of information inconsistency in this paper is a purely analytic definition; how does the function B_{10} behave as $|\hat{\boldsymbol{\theta}}| \to \infty$, while $s_{\mathbf{y}}^2 > 0$ remains fixed. The statistical scenario in which this will most commonly arise is when $\boldsymbol{\theta}$ itself grows increasingly large, with σ^2 staying constant, consistent with the notion that there should then be overwhelming evidence against H_0. Indeed, the definition of conditional Lindley's paradox in Som et al. (2016), which is closely related to information consistency, is formally based on the limiting behavior of parameters. We utilize the analytic version of information inconsistency because it captures the essential behavior without having to deal with probabilistic issues, and also because it is remarkably general in certain situations. For instance, with the standard objective prior having $\nu_0 = \nu_1 = 0$, one can divide through by $s_{\mathbf{y}}^2$ in (4), and state information inconsistency in terms of the statistic $|\hat{\boldsymbol{\theta}}|/s_{\mathbf{y}} \to \infty$, which covers many possible situations in terms of the true parameters.

3.1.1 Practical implications for a univariate test under dependence

The practical importance of information inconsistency is explored for the objective prior with $\nu_1 = \nu_0 = 0$ for a univariate t test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ with correlated data. Specifically, consider $r_1 = 1$, $r_2 = 0$, $\mathbf{X}_1 = \mathbf{1}_n$, and $\boldsymbol{\Omega} = 1$, with $\boldsymbol{\Sigma}$ being the correlation matrix with identical correlations ρ in the off-diagonal elements. The t-statistic, $t = \frac{\hat{\theta}\sqrt{\mathbf{1}_n'\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}}{s_{\mathbf{y}}/\sqrt{n-1}}$, then has a t-distribution with $n-1$ degrees of freedom under H_0. The Bayes factor in (4) can then be expressed as a function of the t-statistic, namely

$$B_{10}(\rho) = \left(1 + \frac{n}{1+(n-1)\rho}\right)^{-1/2} \left(1 - \frac{nt^2}{[t^2+n-1][n+1+(n-1)\rho]}\right)^{-n/2}. \tag{5}$$

The limiting value of the Bayes factor, as $|t|$ goes to infinity, is

$$\begin{aligned}\lim_{|t|\to\infty} B_{10}(\rho) &= \left(1 + \frac{n}{1+(n-1)\rho}\right)^{-1/2} \left(1 - \frac{n}{n+1+(n-1)\rho}\right)^{-n/2} \\ &= \begin{cases} (1+n)^{(n-1)/2}, & \text{if } \rho = 0; \\ \left(1+\frac{2n}{n+1}\right)^{-1/2}\left(\frac{3n+1}{n+1}\right)^{n/2} \approx 3^{(n-1)/2}, & \text{if } \rho = 0.5; \\ 2^{(n-1)/2}, & \text{if } \rho = 1. \end{cases}\end{aligned}$$

Table 1 Limiting values of the Bayes factor for a univariate t test as $|t| \to \infty$ for different choices of the sample size n and the correlation ρ

n	2	5	7	10	20
$\rho = 0$					
Limit	1.73	36	512	4.85×10^4	1.79×10^{11}
B_{10} for $t = 4$	1.55	6.36	12.21	23.61	66.20
$\rho = 0.5$					
Limit	1.53	7.10	20.8	106	2.01×10^4
B_{10} for $t = 4$	1.42	3.46	5.31	8.54	20.71
$\rho \approx 1$					
Limit	1.41	4	8	22.6	724
B_{10} for $t = 4$	1.34	2.76	3.44	4.86	9.47
p value for $t = 4$	0.156	0.016	0.0071	0.0031	0.00077
$1/[-ep \log p]$	2.25	7.81	13.47	24.40	72.01

Additionally, Bayes factors and two-sided p values are given when $t = 4$. The approximation $1/[-ep \log p]$ is an upper bound of the evidence against H_0 (Sellke et al. 2001)

Hence, the correlation can dramatically affect the situation. Table 1 provides the limiting value of the Bayes factor as $|t|$ goes to ∞ for different choices of the correlation ρ and different sample sizes varying from $n = 2$ to a sample size of $n = 20$. The table also provides the Bayes factor when $t = 4$ to check whether inconsistency starts coming into play for a large t value. As comparisons, the corresponding two-sided p values are also provided, as well as the upper bound $B_{10} < 1/[-ep \log p]$, which is a bound over a large nonparametric class of priors [derived in Sellke et al. (2001)].

When there is zero correlation, the limit $(n+1)^{(n-1)/2}$ is large for sample sizes larger than 6, so that information inconsistency is not problematical from a practical point of view. For large correlations on the other hand, and especially when ρ is close to 1, the limiting values can be quite small, arguing against the use of objective conjugate priors.

Figure 1 displays the logarithm of the Bayes factor as a function of $\log_{10}(t)$ when using conjugate priors (solid lines) and $n = 7$, $\rho = .5$, $s_{\mathbf{y}}^2 = n - 1 = 6$, $s_0^2 = s_1^2 = 1$, and different choices for the prior degrees of freedom, namely $(\nu_0, \nu_1) = (0, 0), (1, 2)$ or $(2, 1)$. As can be seen, if $\nu_0 = \nu_1 = 0$, the logarithm of the Bayes factor converges to $\log_{10}(20.8) = 1.32$ (Table 1). Furthermore, if $\nu_0 < \nu_1$ (or $\nu_0 > \nu_1$), the evidence goes to ∞ for H_0 (or H_1) as $t \to \infty$ implying information inconsistency (or information consistency). The results are qualitatively similar when using other values for the prior scales.

It is natural to ask if information inconsistency also occurs if ρ is unknown. The answer is yes, as shown in the following lemma.

Lemma 2 *If $\rho > 0$ is unknown with prior density $\pi(\rho)$, and the same priors are assumed for the other parameters, then, for $t^2 > n - 1$,*

$$B_{10}(\rho) \leq (1+n)^{-1/2} \left(1 - \frac{nt^2}{(t^2+n-1)(n+1)}\right)^{-n/2}$$

which converges to $(1+n)^{(n-1)/2}$ as $|t| \to \infty$, implying information inconsistency.

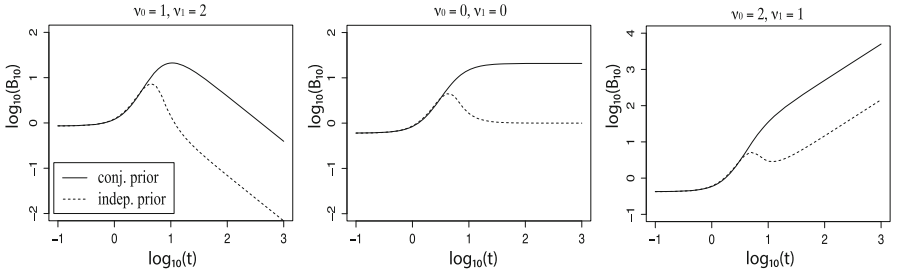

Fig. 1 The Bayes factor B_{10} based on the conjugate prior (solid line) and independence prior (dashed line) as a function of t values when $n = 7$, $\rho = .5$, $s^2_{\mathbf{y}} = n - 1 = 6$, $s^2_0 = s^2_1 = 1$, and different choices for the prior degrees of freedom ν_0 and ν_1

Proof Calculus shows that, for $t^2 > n - 1$, (5) is a decreasing function of ρ on $[0, 1]$ and hence is maximized at

$$B_{10}(0) = (1 + n)^{-1/2} \left(1 - \frac{nt^2}{(t^2 + n - 1)(n + 1)}\right)^{-n/2}.$$

We complete the proof by showing that

$$B_{10}(\rho) = \frac{\int p_1(\mathbf{y}|\rho)\pi(\rho)\mathrm{d}\rho}{\int p_0(\mathbf{y}|\rho)\pi(\rho)\mathrm{d}\rho} \leq B_{10}(0). \tag{6}$$

Indeed, (6) is equivalent to

$$\int [p_1(\mathbf{y}|\rho) - B_{10}(0)p_0(\mathbf{y}|\rho)]\pi(\rho)\mathrm{d}\rho \leq 0,$$

which is true because $[p_1(\mathbf{y}|\rho) - B_{10}(0)p_0(\mathbf{y}|\rho)] \leq 0$ is equivalent to $B_{10}(\rho) \leq B_{10}(0)$, ending the proof. □

The restriction to $\rho > 0$ is not necessary, but simplifies the proof.

3.2 Mixtures of conjugate priors

Although use of conjugate priors in testing is common, it has long been argued [starting with Jeffreys (1961)] that fatter-tailed prior distributions should be used. One such class that is increasingly popular is the class of scale mixtures of conjugate priors. This class results in information consistent Bayes factors if the prior on g is thick enough, as shown by the following lemmas which generalize the result in Liang et al. (2008) for $\nu_0 = \nu_1 = 0$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\Omega} = g(\mathbf{X}'_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_{\boldsymbol{\theta}})^{-1}$.

Lemma 3 *Let* $\boldsymbol{\theta} \mid g, \boldsymbol{\gamma}_1, \sigma^2_1 \sim N_{r_1}(\mathbf{0}, g\sigma^2_1\boldsymbol{\Omega})$, *where* σ^2_1 *has the prior specified in* (2) *and* g *has a prior with density* $\pi(g)$. *If* $\nu_0 > \nu_1$, *any* $\pi(g)$ *with positive support yields an information consistent* B_{10}. *The condition*

$$\int_0^\infty (g+1)^{(n-r_1-r_2+\nu_1)/2}\pi(g)\mathrm{d}g = \infty$$

is necessary and sufficient for information consistency whenever $\nu_0 = \nu_1$*, and necessary whenever* $\nu_0 < \nu_1$.

Proof See "Appendix A".

The maximum number of finite moments that the prior on g can have to achieve information consistency increases with the sample size n and decreases with the number of predictors $K = r_1 + r_2$. Lemma 3 gives us a complete description for all scale mixtures of conjugate priors whenever $\nu_0 \geq \nu_1$, but only gives us a necessary condition for information consistency for $\nu_0 < \nu_1$. The lemma below characterizes the behavior of polynomial-tailed priors on g in this latter case and provides partial results for priors with thinner- and thicker-than-polynomial priors on g. □

Lemma 4 *Suppose* $\nu_0 < \nu_1$ *and let* $\boldsymbol{\theta} \mid g, \boldsymbol{\gamma}_1, \sigma_1^2 \sim N_{r_1}(\mathbf{0}, g\sigma_1^2\boldsymbol{\Omega})$*, where* σ_1^2 *has the prior specified in* (2) *and* g *has a prior with density* $\pi(g)$*. Then, the following are true:*

1. *If there exist* $0 < M < \infty$ *and* $0 < K < \infty$ *such that for all* $g \geq M$, $\pi(g) \geq Kg^{-\alpha}$ *for* $\alpha > 1$, B_{10} *is information consistent whenever* $\alpha < (n - r_1 - r_2 + \nu_0)/2 + 1$.
2. *If there exist* $0 < M' < \infty$ *and* $0 < K' < \infty$ *such that for all* $g \geq M'$, $\pi(g) \leq K'g^{-\alpha}$ *for* $\alpha > 1$, B_{10} *is information inconsistent whenever* $\alpha \geq (n - r_1 - r_2 + \nu_0)/2 + 1$.

[NB: All of the priors on g considered in Liang et al. (2008) *satisfy both conditions.]*

Proof See "Appendix B".

Note that the Zellner and Siow prior (1980) (which was the first proposed information consistent prior for this situation) and the hyper-g prior (Liang et al. 2008) satisfy both conditions because they have polynomial tails. □

3.3 Independence priors

3.3.1 Semi-conjugate prior

A feature of the conjugate prior that is sometimes questioned is the dependence induced between $\boldsymbol{\theta}$ and σ^2; in objective Bayesian analysis, this is hard to avoid (only σ is available to provide an objective scale for $\boldsymbol{\theta}$), but it does seem rather arbitrary. For example, Moran et al. (2018) advocated the use of independent priors as dependent conjugate priors may result in severe underestimation of the error variance in variable selection problems. Hence, it is of interest to also investigate information consistency using independent semi-conjugate priors of the form

$$\begin{aligned}\pi_1(\boldsymbol{\theta}, \boldsymbol{\gamma}_1, \sigma_1^2) &= \pi_1(\boldsymbol{\theta}) \times \pi_1(\boldsymbol{\gamma}_1) \times \pi_1(\sigma_1^2) \\ &\propto N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega}) \times 1 \times \text{inv-}\chi^2(\sigma_1^2|s_1^2, \nu_1) \\ \pi_0(\boldsymbol{\gamma}_0, \sigma_0^2) &= \pi_0(\boldsymbol{\gamma}_0) \times \pi_0(\sigma_0^2)\end{aligned}$$

$$\propto 1 \times \text{inv-}\chi^2(\sigma_0^2|s_0^2, \nu_0).$$

With these semi-conjugate priors, the Bayes factor becomes

$$B_{10} = C_2 \times \frac{\int \left(\nu_1 s_1^2 + s_{\mathbf{y}}^2 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{X}_{\boldsymbol{\theta}}' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right)^{-\frac{n-r_2+\nu_1}{2}} N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega}) \mathrm{d}\boldsymbol{\theta}}{\left(\nu_0 s_0^2 + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}' \mathbf{X}_{\boldsymbol{\theta}}' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_0}{2}}}, \tag{7}$$

where

$$C_2 = \frac{(\nu_1/2)^{\nu_1/2} s_1^{\nu_1} \Gamma\left(\frac{\nu_0}{2}\right) \Gamma\left(\frac{n+\nu_1-r_2}{2}\right)}{(\nu_0/2)^{\nu_0/2} s_0^{\nu_0} \Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{n+\nu_0-r_2}{2}\right)} 2^{(\nu_1-\nu_0)/2}.$$

Lemma 5 *As* $|\hat{\boldsymbol{\theta}}| \to \infty$, *the Bayes factor in* (7), *based on the independent semi-conjugate prior, behaves as follows:*

$$B_{10} \to \begin{cases} 0 & \text{if } \nu_0 < \nu_1; \\ 1 & \text{if } \nu_0 = \nu_1; \\ \infty & \text{if } \nu_0 > \nu_1. \end{cases}$$

Proof See "Appendix C". □

Note that, in the typical case of $\nu_0 = \nu_1$, we observe an even worse case of information inconsistency than for the conjugate prior because the relative evidence between H_1 and H_0 goes to 1 when there appears to be overwhelming evidence for H_1; in contrast, for the conjugate prior case, the limiting Bayes factor—while nonzero—was at least exponentially small in n.

The intuition behind this result is that very large $\hat{\boldsymbol{\theta}}$ is equally unlikely under H_1 and H_0, due to the light-tailed normal prior for $\boldsymbol{\theta}$ under H_1. Furthermore, the limits are the same as in the conjugate case if $\nu_0 \neq \nu_1$. Hence, the choice of the prior degrees of freedom plays a crucial role in information inconsistency, even when the variance is a priori independent of $\boldsymbol{\theta}$.

Figure 1 also displays the Bayes factor, based on the independence prior, as a function of $\log_{10}(t)$ for the univariate t test when the data correlation is $\rho = .5$ (dashed line). As can be seen, the Bayes factor based on the independence prior and the conjugate prior with the same hyperparameters is approximately equal for absolute t values smaller than approximately $\log_{10}(.5)$. For larger t values, the flatter tails of the independence priors start to have an effect resulting in a decrease in the Bayes factor, relative to the Bayes factor based on the conjugate priors.

3.3.2 Fatter-tailed independence priors

It is somewhat unfair to use an independent normal prior for model comparison here since, from Jeffreys (1961), the use of fatter-tailed priors has been recommended. To

keep the discussion of fatter-tailed priors simple, we consider only the one-dimensional case (i.e., $r_2 = 0$) and restrict the prior $\pi_1(\theta)$ to be a t-distribution with mean 0, scale τ (fixed) and degrees of freedom ν, i.e.,

$$\pi_1(\theta) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)\tau}\left(1+\frac{\theta^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}}.$$

Then Theorem 3.3 in Fan and Berger (1992) shows that, as $|\hat{\theta}| \to \infty$,

$$B_{10} = C\,\frac{\frac{\Gamma((n^*+1)/2)}{\sqrt{n^*\pi}\,\Gamma(n^*/2)\sqrt{V}}\left(1+\frac{\hat{\theta}^2}{n^*V}\right)^{-\frac{n^*+1}{2}} + \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)\tau}\left(1+\frac{\hat{\theta}^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}}}{\left(\nu_0 s_0^2 + s_{\mathbf{y}}^2 + \hat{\theta}'\mathbf{X}_\theta'\boldsymbol{\Sigma}^{-1}\mathbf{X}_\theta\hat{\theta}\right)^{-\frac{n+\nu_0}{2}}} \times(1+o(1)),$$

where $n^* = n + \nu_1 - 1$, $V = (\nu_1 s_1^2 + s_{\mathbf{y}}^2)/[n^*\mathbf{X}_\theta'\boldsymbol{\Sigma}^{-1}\mathbf{X}_\theta]$ and

$$C = \frac{(\nu_1/2)^{\nu_1/2} s_1^{\nu_1}\sqrt{n^*\pi}\,\Gamma(n^*/2)\sqrt{V}}{\Gamma(\nu_1/2)\;(\nu_1 s_1^2 + s_{\mathbf{y}}^2)^{(n+\nu_1)/2}}.$$

Thus, as $|\hat{\theta}| \to \infty$,

$$B_{10} \to \begin{cases} 0 & \text{if} \quad n+\nu_0 < \min\{n-1+\nu_1, \nu+1\}; \\ \text{constant} & \text{if} \quad n+\nu_0 = \min\{n-1+\nu_1, \nu+1\}; \\ \infty & \text{if} \quad n+\nu_0 > \min\{n-1+\nu_1, \nu+1\}. \end{cases}$$

Since $n \geq 2$, if $0 < \nu < 1$ it will be true that $n + \nu_0 > \min\{n-1+\nu_1, \nu+1\}$ so that B_{10} will be information consistent. For the commonly used Cauchy prior ($\nu = 1$), information consistency also holds, except for the case when $n = 2$ and $\nu_0 = 0$ (this last corresponding to the objective prior for σ_0^2). It is interesting that information consistency does hold for this last case when $\pi_1(\theta)$ is chosen to be Cauchy$(0, \sigma_1)$ (cf. Liang et al. 2008) and $\nu_1 = 0$; thus, once again, insisting on prior independence of σ_1^2 and θ only appears to worsen the problem of information inconsistency.

3.4 Adaptive priors

Another approach to Bayesian hypothesis testing is to let the prior under H_1 adapt to the likelihood, as in George and Foster (2000) and Hansen and Yu (2001).

Example 2 For the g prior in the t test, when the t-statistic $t = \sqrt{\frac{\hat{\boldsymbol{\theta}}'\mathbf{X}_1'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}}{s_{\mathbf{y}}^2/(n-1)}} > 1$, the marginal likelihood under H_1 is maximized for the choice $g = \frac{n-r_2-r_1}{r_1(n-1)}t^2 - 1$. The Bayes factor for this choice equals

$$B_{10} = \left(\frac{r_1(n-1)}{t^2(n-r_1-r_2)}\right)^{\frac{r_1}{2}} \left(\frac{(n-1+t^2)(n-r_1-r_2)}{(n-1)(n-r_2)}\right)^{\frac{n-r_2}{2}},$$

which is information consistent. For a univariate t test, with $r_1 = 1$ and $r_2 = 0$, the resulting Bayes factor can be expressed as $B_{10} = \frac{1}{|t|}\left(\frac{n-1+t^2}{n}\right)^{\frac{n}{2}}$.

The following lemma generalizes the result in Liang et al. (2008) for $\nu_0 = \nu_1 = 0$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\Omega} = g(\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_{\boldsymbol{\theta}})^{-1}$.

Lemma 6 *Let* $\boldsymbol{\theta} \mid g, \boldsymbol{\gamma}_1, \sigma_1^2 \sim N_{r_1}(\mathbf{0}, g\sigma_1^2\boldsymbol{\Omega})$*, where* σ_1^2 *has the prior specified in* (2)*. If* $g > 0$ *is set by maximizing* B_{10}*, information consistency holds.*

Proof See "Appendix D". □

Lemma 6 establishes information consistency for all ν_0 and ν_1. This is in contrast to the results in previous sections, where the behavior of B_{10} depends (sometimes rather strongly) on ν_0 and ν_1.

4 One-sided hypothesis testing

The following definition will be used for information consistency for a one-sided testing problem.

Definition 2 A Bayes factor is *information consistent*, for a one-sided hypothesis test of $H_0 : \boldsymbol{\theta} \leq \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \not\leq \mathbf{0}$, if $B_{10} \to \infty$ as $|\hat{\boldsymbol{\theta}}| \to \infty$ with at least one coordinate of $\hat{\boldsymbol{\theta}}$ going to ∞, and $B_{10} \to 0$, as all coordinates of $\hat{\boldsymbol{\theta}}$ go to $-\infty$. If this does not hold, the Bayes factor is called information inconsistent.

We shall denote the subspaces under H_0 and H_1 as $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \leq \mathbf{0}\}$ and $\boldsymbol{\Theta}_1 = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \not\leq \mathbf{0}\}$, respectively.

4.1 Conjugate prior

When testing nonnested hypotheses, it is common to formulate an encompassing prior π on the joint space $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$ and specify truncations of this prior under H_0 and H_1 (e.g., Berger and Mortera 1999; Klugkist and Hoijtink 2007). As in the null hypothesis test, the encompassing conjugate prior is centered on the boundary of the subspaces under investigation, i.e.,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2) \propto N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\boldsymbol{\Omega}) \times \text{inv-}\chi^2(\sigma^2|s^2, \nu), \tag{8}$$

with a flat improper prior for $\boldsymbol{\gamma}$. The priors under the nonnested hypotheses H_t, for $t = 0$ or 1, can then be expressed as

$$\pi_t(\boldsymbol{\theta} \mid \sigma^2) = \pi(\boldsymbol{\theta} \mid \sigma^2) I_{\boldsymbol{\Theta}_t}(\boldsymbol{\theta}) / P_\pi(\boldsymbol{\theta} \in \boldsymbol{\Theta}_t \mid \sigma^2), \tag{9}$$

$\pi_t(\sigma^2) = \pi(\sigma^2)$, and $\pi_t(\boldsymbol{\gamma}) = \pi(\boldsymbol{\gamma})$, with the denominator in (9) being equal to the conditional prior probability of $\boldsymbol{\Theta}_t$ under the joint prior on $\boldsymbol{\Theta}$, i.e., $P_\pi(\boldsymbol{\theta} \in \boldsymbol{\Theta}_t \mid \sigma^2) = \int_{\boldsymbol{\Theta}_t} N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\boldsymbol{\Omega})\mathrm{d}\boldsymbol{\theta} > 0$.

The Bayes factor for the one-sided hypothesis test based on the conjugate priors can then be expressed as

$$B_{10} = \left(P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \sigma^2 = 1)^{-1} - 1\right)^{-1} \left(P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y})^{-1} - 1\right). \tag{10}$$

The derivation is similar to that in Mulder (2014a). The prior and posterior probabilities that the constraints hold under the encompassing model can be computed as the proportion of draws satisfying the constraints. Also note that the conditional prior probability of $\boldsymbol{\theta} \leq \mathbf{0}$ is completely determined by the prior covariance matrix $\boldsymbol{\Omega}$ and is independent of σ^2 [therefore, we can set $\sigma^2 = 1$ in (10)]. This is a direct result of centering the encompassing prior on the point of interest $\mathbf{0}$. For example, if $\boldsymbol{\Omega} = \mathbf{I}_{r_1}$, then $P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \sigma^2) = 2^{-r_1}, \forall \sigma^2 > 0$. In the g prior with $\boldsymbol{\Omega} = g\sigma^2(\mathbf{X}_1\boldsymbol{\Sigma}^{-1}\mathbf{X}_1)^{-1}$, the prior probability is completely determined by the covariance structure of the predictors.

As can be concluded from (10), a Bayes factor for a one-sided hypothesis test is information consistent if $P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y}) \to 0$ as $|\hat{\boldsymbol{\theta}}| \to \infty$ with at least one coordinate of $\hat{\boldsymbol{\theta}}$ going to ∞, and $P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y}) \to 1$ as all coordinates of $\hat{\boldsymbol{\theta}}$ go to $-\infty$.

Lemma 7 *$P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y})$ is bounded away from 0 and 1 for all $\mathbf{y}$. Hence B_{10} is information inconsistent.*

If $\hat{\boldsymbol{\theta}} = c\boldsymbol{v}$ and $c \to \infty$, then

$$P_\pi(\boldsymbol{\theta} \leq \boldsymbol{0} \mid \boldsymbol{y}) \to P_\pi(\boldsymbol{\xi} \leq \boldsymbol{0} \mid \boldsymbol{y}),$$

where $\boldsymbol{\xi}$ has a multivariate t distribution with mean

$$\boldsymbol{v}^* = \frac{(\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1 + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1\boldsymbol{v}}{(n + \nu - r_2)^{-1/2}(\boldsymbol{v}'((\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1)^{-1} + \boldsymbol{\Omega})^{-1}\boldsymbol{v})^{1/2}},$$

scale matrix $(\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1 + \boldsymbol{\Omega}^{-1})^{-1}$, and $n + \nu - r_2$ degrees of freedom.

Proof See "Appendix E". The same result can be shown to hold (by essentially the same argument) if a proper conjugate prior is used for $\boldsymbol{\gamma}$. □

4.1.1 Practical implications for a univariate one-sided test under dependence

We investigate the practical importance of information inconsistency for a univariate one-sided t test under dependence of $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, with $\nu = 0, r_1 = 1, r_2 = 0, \mathbf{X}_1 = \mathbf{1}, \boldsymbol{\Omega} = 1$, and $\boldsymbol{\Sigma} = \rho\mathbf{J}_n + (1 - \rho)\mathbf{I}_n$, so that $P_\pi\left(\theta \leq 0 \mid \sigma^2\right) = \frac{1}{2}$. Based on Lemma 7, the Bayes factor is then given by

Table 2 Limiting values of the Bayes factor for a one-sided univariate t test as $t \to \infty$ for different choices of the sample size n and the correlation ρ

n	2	5	7	10	20
$\rho = 0$					
Limit	9.90	486	9.45×10^3	1.26×10^6	1.85×10^{14}
B_{10} for $t = 4$	8.62	78.9	199	510	2.40×10^3
$\rho = 0.5$					
Limit	7.19	57.2	199	1.21×10^3	4.02×10^5
B_{10} for $t = 4$	6.50	25.5	44.7	81.5	238
$\rho \approx 1$					
Limit	5.83	25.5	59.3	197	8.57×10^4
B_{10} for $t = 4$	5.37	14.7	22.4	35.2	80.9
One-sided					
p value for $t = 4$	0.078	0.008	0.0036	0.0016	0.0038

Additionally, Bayes factors and one-sided p values are given when $t = 4$

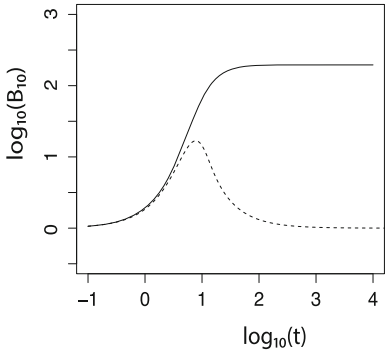

Fig. 2 The Bayes factor B_{10} for the one-sided hypothesis test based on the conjugate prior (solid line) and independence prior (dashed line) as a function of t values when $n = 7$, $\rho = .5$, $s^2_{\mathbf{y}} = n - 1 = 6$, and setting the objective prior to be improper via $\nu = 0$

$$B_{10} = T_n\left(-\sqrt{\frac{n^2}{1+(n-1)\rho+t^{-2}(n-1)(1+n+(n-1)\rho)}}\right)^{-1} - 1$$
$$\to T_n\left(-n(1+(n-1)\rho)^{-\frac{1}{2}}\right)^{-1} - 1, \tag{11}$$

as $t \to \infty$, where $T_\nu(\cdot)$ denotes the cdf of a univariate Student t distribution with ν degrees of freedom. Note that as $t \to -\infty$, B_{10} converges to the reciprocal of (11).

Table 2 provides the limiting values of the Bayes factors and Bayes factors in the case of a relatively large t value of 4 for different sample sizes and correlations. When comparing Table 2 with Table 1, we can conclude that the practical importance of information inconsistency for one-sided hypothesis testing is considerably less problematic in comparison with the null hypothesis test. Finally, Fig. 2 (solid line) displays the Bayes factor for the one-sided hypothesis test as a function of the t value based on $n = 7$, $\rho = .5$, $s^2_{\mathbf{y}} = n - 1 = 6$, and setting the objective improper based on $\nu = 0$.

4.2 Mixtures of conjugate priors

We provide the following necessary and sufficient condition for information consistency for a scale mixture of conjugate normal priors in a one-sided hypothesis test.

Lemma 8 *Let* $\boldsymbol{\theta} \mid g, \sigma^2 \sim N_{r_1}(\mathbf{0}, g\sigma^2\boldsymbol{\Omega})$, *where* σ^2 *has the prior specified in* (8) *and* g *has a prior with density* $\pi(g)$, *and let* $\boldsymbol{w} = E(\boldsymbol{\theta} \mid g, \boldsymbol{y})$. *Assume that if there exists* i *such that* $\widehat{\theta}_i \to +\infty$, *there exists* j *such that* $w_j > 0$. *Alternatively, assume that if* $\widehat{\theta}_i \to -\infty$ *for all* i, *then* $w_i < 0$ *for all* i. *[For instance, this condition is satisfied if* $\boldsymbol{\theta}$ *is univariate or* $\boldsymbol{\Omega} \propto (\boldsymbol{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_{\boldsymbol{\theta}})^{-1}$*]. Then, the condition*

$$\int_0^\infty (g+1)^{(n-r_1-r_2+\nu)/2}\pi(g)\mathrm{d}g = \infty$$

is necessary and sufficient for information consistency.

Proof See "Appendix F". □

4.3 Independence prior

The independence semi-conjugate encompassing prior is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2) \propto N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega}) \times \text{inv-}\chi^2(\sigma^2|s^2, \nu). \tag{12}$$

The truncated priors of $\boldsymbol{\theta}$ under the nonnested hypotheses are as in (9), except that the normalizing constant $P_\pi(\boldsymbol{\theta} \in \boldsymbol{\Theta}_t)$ is the marginal prior probability of $\boldsymbol{\Theta}_t$.

The Bayes factor for the one-sided hypothesis test based on the independence prior can again be expressed as

$$B_{10} = \left(P_\pi(\boldsymbol{\theta} \le \mathbf{0})^{-1} - 1\right)^{-1}\left(P_\pi(\boldsymbol{\theta} \le \mathbf{0} \mid \mathbf{y})^{-1} - 1\right), \tag{13}$$

but note that the posterior probability is no longer available in closed form.

Lemma 9 *As* $|\hat{\boldsymbol{\theta}}| \to \infty$ *and at least one coordinate of* $\hat{\boldsymbol{\theta}}$ *goes to* ∞, *the Bayes factor of* $H_1 : \boldsymbol{\theta} \not\le \boldsymbol{0}$ *versus* $H_0 : \boldsymbol{\theta} \le \boldsymbol{0}$ *based on the independence encompassing prior in* (12) *satisfies*

$$B_{10} \to \left(P_\pi(\boldsymbol{\theta} \le \boldsymbol{0})^{-1} - 1\right)^{-1}.$$

Proof See "Appendix G". □

Thus, as in null hypothesis testing, the independence prior results in a serious violation of information consistency because the evidence in the data of H_1 relative to H_0 goes to 1 when the evidence against H_0 appears to be overwhelming. For completeness, the Bayes factor for the one-sided hypothesis test is also displayed in Fig. 2 (dashed line), illustrating the extreme form of information inconsistency.

4.4 Adaptive priors

An adaptive prior can be specified where the prior covariance matrix of $\boldsymbol{\theta}$ is adapted to the likelihood such that the Bayes factor is maximized for the hypothesis that is supported by the data (i.e., maximize B_{01} if $\hat{\boldsymbol{\theta}} \leq \mathbf{0}$, and maximize B_{10} elsewhere). Here we show that an adaptive g prior results in an information consistent Bayes factor.

Lemma 10 *The Bayes factor based on the g prior, with* $g_{\max} = \arg\max_g\{B_{01}\}$ *if* $\hat{\boldsymbol{\theta}} \leq \mathbf{0}$ *and* $g_{\max} = \arg\max_g\{B_{10}\}$ *if* $\hat{\boldsymbol{\theta}} \not\leq \mathbf{0}$*, is information consistent for one-sided hypothesis testing.*

Proof A proof is given in "Appendix H". □

As shown in the proof, the choice for g that maximizes the Bayes factor is obtained by letting g go to ∞ (see also, Mulder 2014a). As a result of letting the prior variances go to infinity, the posterior is not shrunk toward the prior mean, which is sufficient to establish information consistency. Therefore, the methods of Mulder (2014b) and Gu et al. (2014) are also information consistent. A potential issue of letting g go to infinity is that the marginal likelihoods under H_0 and H_1 go to 0 in the limit. However because the Bayes factor in (10) converges to a limit where the posterior probabilities are computed using flat priors and the prior probabilities are based on the prior covariance structure, the outcome seems a reasonable default quantification of the relative evidence for a one-sided test.

5 Multiple hypothesis testing

Below we consider the definition for information (in)consistency in a multiple testing problem. The definition implies that a Bayes factor needs to be information consistent for both a precise test and a one-sided test. A graphical representation for the bivariate case can be found in Fig. 3.

Definition 3 A Bayes factor is *information consistent*, for a multiple hypothesis test of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \leq \mathbf{0} \text{ and } \boldsymbol{\theta} \neq \mathbf{0}\}$ versus $H_2 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_2 = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \not\leq \mathbf{0}\}$, if $B_{20}, B_{21} \to \infty$ as $|\hat{\boldsymbol{\theta}}| \to \infty$ with at least one coordinate of $\hat{\boldsymbol{\theta}}$ going to ∞, and $B_{10}, B_{12} \to \infty$, as all coordinates of $\hat{\boldsymbol{\theta}}$ go to $-\infty$. If this does not hold, the Bayes factor is called information inconsistent.

As the conjugate and independent semi-conjugate priors resulted in information inconsistent Bayes factors for the one-sided hypothesis test, this automatically implies that these priors result in information inconsistency for the multiple hypothesis test. A specific case when using conjugate priors that is interesting to mention is when setting the prior degrees of freedom for σ^2 under H_0 larger than the prior degrees of freedom for σ^2 under the encompassing prior to construct truncated priors under H_1 and H_2, i.e., $\nu_0 > \nu$. This results in information consistency for the precise hypothesis test (a consequence of Lemma 1) and information inconsistency for the one-sided test (a consequence of Lemma 7). To see that this results in undesirable behavior consider a

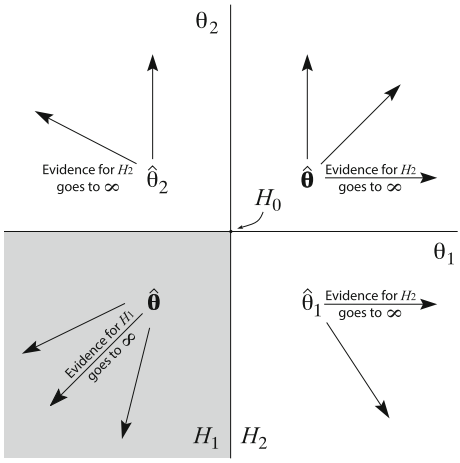

Fig. 3 Graphical representation of the definition of an information consistent Bayes factor in a multiple testing problem of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \leq \mathbf{0}$ and $\boldsymbol{\theta} \neq \mathbf{0}$ (gray quadrant) versus $H_2 : \boldsymbol{\theta} \not\leq \mathbf{0}$ (white quadrants). The directions of the arrows reflect directions of the limits. The evidence for H_1 against H_0 and H_2 should go to ∞ for limits in the lower left quadrant, and the evidence for H_2 against H_0 and H_1 should go to ∞ for the limits in the white quadrants, in order for the Bayes factor to be information consistent

univariate multiple t test of $H_0 : \theta = 0$ versus $H_1 : \theta < 0$ versus $H_2 : \theta > 0$. If we let $t \to \infty$, the support for H_1 against H_0 would go to ∞. Thus as the effect goes to plus infinity, the evidence for the existence of a negative effect against no effect diverges.

Finally, note that Lemmas 3 and 8 give the necessary and sufficient conditions for the mixing distribution of the scale mixture of conjugate priors to be information consistent in the multiple testing problem.

6 Conclusions

This paper explored the existence of information inconsistency when using conjugate priors, mixtures of g priors, independence priors, and adaptive g priors for precise testing, one-sided testing, and multiple hypothesis testing. An overview of our findings can be found in Table 3.

Table 3 Severity of information inconsistency of various priors for different hypothesis tests

	Prior properties	Precise testing	One-sided testing	Multiple testing
Conjugate priors	$\nu_0 < \nu_1$	Disastrous		Disastrous
	$\nu_0 = \nu_1$	Normal	Normal	Normal
	$\nu_0 > \nu_1$	No		Normal
Mixtures of g priors	Thick-tailed mixture	No	No	No
Independence priors	$\nu_0 < \nu_1$	Disastrous		Disastrous
	$\nu_0 = \nu_1$	Severe	Severe	Severe
	$\nu_0 > \nu_1$	No		Severe
Adaptive g prior	Any ν_0, ν_1	No	No	

"Normal" information inconsistency refers to a (typically large) limiting bound B of the evidence against the null (i.e., $B_{10} \to B$)
"Severe" refers to a limiting bound that is close to 1 (i.e., $B_{10} \leftrightsquigarrow 1$). "Disastrous" refers to infinite evidence in the opposite direction (i.e., $B_{10} \to 0$). "No" refers to no information inconsistency; thus, information consistency (i.e., $B_{10} \to \infty$)

The first major conclusion is that information inconsistency is ubiquitous when typical conjugate priors are used in hypothesis testing and model selection in the normal linear model with unknown variance. (Again, the problem does not seem to arise in normal linear models with known variance.) It happens in standard null hypothesis testing and one-sided testing; it happens with proper and improper conjugate priors; and it happens with almost all independence conjugate priors. The practical importance of the problem varies over different situations; it will primarily be a practical problem when the sample is small relative to the number of free parameters and there is high correlation between the observations. But, even in other cases, we consider information inconsistency to be highlighting a logical flaw that might have other serious consequences and is, hence, something to be avoided.

The second major conclusion is that use of either fatter-tailed priors (including appropriate mixtures of g-priors) or adaptive priors typically results in information consistency. This is not as surprising as the almost complete lack of information consistency for conjugate priors, in that previous particular fatter-tailed priors (such as the Zellner–Siow prior) had been shown to be information consistent. Still, the generality in which such priors can be shown to be information consistent is highly comforting.

It should be noted that, when proper priors yield information inconsistency, a logical flaw in Bayesian analysis is not being discovered; if one truly believed the priors were correct, then one should behave in an information inconsistent manner. But one rarely accurately knows features of the priors—such as their tail behaviors—that determine information inconsistency. Thus the intuitive appeal of information consistency can be used as a significant aid to selection of such prior features.

Finally, information inconsistency is not limited to the normal linear model with unknown variance, as shown in the following example.

Example 3 Let $y \mid \theta \sim \text{Cauchy}(\theta, 1)$ and suppose that we want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Under H_1, assume that $\theta \sim \text{Cauchy}(0, \psi)$. Then, the Bayes factor in favor of H_1 to H_0 is

$$\text{BF}_{10} = \frac{(1+\psi)(1+y^2)}{(1+\psi)^2 + y^2}.$$

As $y \to \infty$, $\text{BF}_{10} \to \psi(1+\psi) < \infty$, so the Bayes factor is information inconsistent. This example also shows that information consistency is not dependent, in general, on having an unknown scale parameter; here the scale parameter of the observation is known.

Funding The first author was funded by the Netherlands Organization for Scientific Research (NWO Veni Grant No. 451-13-011).

A Proof of Lemma 3

Denote:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \left(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1\right)^{-1} \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{y} \\
s_{\mathbf{y}}^2 &= (\mathbf{y} - \mathbf{X}_1 \widehat{\boldsymbol{\theta}} - \mathbf{X}_0 \widehat{\boldsymbol{\gamma}})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}_1 \widehat{\boldsymbol{\theta}} - \mathbf{X}_0 \widehat{\boldsymbol{\gamma}}) \\
\mathsf{SSE}_0 &= s_0^2 \nu_0 + s_{\mathbf{y}}^2 \\
\mathsf{SSE}_1 &= s_1^2 \nu_1 + s_{\mathbf{y}}^2 \\
\mathsf{SSR} &= \widehat{\boldsymbol{\theta}}' \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \widehat{\boldsymbol{\theta}} \\
\mathcal{I}_\theta &= \mathbf{X}_\theta' \boldsymbol{\Sigma}^{-1} \mathbf{X}_\theta \\
p_0 &= r_2 - \nu_0 \\
p_1 &= r_2 - \nu_1
\end{aligned}$$

Throughout, we use the following notation for functions a, b:

- $a(g, \widehat{\boldsymbol{\theta}}) \lesssim b(g, \widehat{\boldsymbol{\theta}})$ if and only if there exists $0 < M < \infty$ which does not depend on g or $\widehat{\boldsymbol{\theta}}$ such that $a(g, \widehat{\boldsymbol{\theta}}) \leq M b(g, \widehat{\boldsymbol{\theta}})$.
- $a(g, \widehat{\boldsymbol{\theta}}) \gtrsim b(g, \widehat{\boldsymbol{\theta}})$ if and only if there exists $0 < M < \infty$ which doesn't depend on g or $\widehat{\boldsymbol{\theta}}$ such that $a(g, \widehat{\boldsymbol{\theta}}) \geq M b(g, \widehat{\boldsymbol{\theta}})$.
- $a(g, \widehat{\boldsymbol{\theta}}) \asymp b(g, \widehat{\boldsymbol{\theta}})$ if and only if $a(g, \widehat{\boldsymbol{\theta}}) \lesssim b(g, \widehat{\boldsymbol{\theta}})$ and $a(g, \widehat{\boldsymbol{\theta}}) \gtrsim b(g, \widehat{\boldsymbol{\theta}})$.

Before we prove Lemma 3, we prove an auxiliary result

Lemma 11 *Let*

$$h(g) = |g\boldsymbol{\Omega} + \mathcal{I}_\theta^{-1}|^{-1/2} [SSE_1 + \widehat{\boldsymbol{\theta}}' (g\boldsymbol{\Omega} + \mathcal{I}_\theta^{-1})^{-1} \widehat{\boldsymbol{\theta}}]^{-(n-p_1)/2},$$

then, there exist $0 < d_l < d_u < \infty$ *such that*

$$\frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)SSE_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \lesssim h(g) \lesssim \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)SSE_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}$$

Proof Consider the matrix factorization

$$\mathcal{I}_\theta^{-1} + g\boldsymbol{\Omega} = \boldsymbol{\Omega}^{1/2} [\boldsymbol{\Omega}^{-1/2} \mathcal{I}_\theta^{-1} \boldsymbol{\Omega}^{-1/2} + g I_{r_1}] \boldsymbol{\Omega}^{1/2},$$

and take the eigendecomposition $\boldsymbol{\Omega}^{-1/2} \mathcal{I}_\theta^{-1} \boldsymbol{\Omega}^{-1/2} = \boldsymbol{O} \boldsymbol{D} \boldsymbol{O}'$, where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{D}$ diagonal with elements $0 < d_l < d_i < d_u < \infty$. Then, we can rewrite

$$\mathcal{I}_\theta^{-1} + g\boldsymbol{\Omega} = \boldsymbol{\Omega}^{1/2} \boldsymbol{O} [\boldsymbol{D} + g I_{r_1}] \boldsymbol{O}' \boldsymbol{\Omega}^{1/2}.$$

□

We can bound

$$\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}/(d_u+g) \leq \widehat{\boldsymbol{\theta}}'(g\boldsymbol{\Omega}+\mathcal{I}_\theta^{-1})^{-1}\widehat{\boldsymbol{\theta}} \leq \widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}/(d_l+g)$$

and

$$|g\boldsymbol{\Omega}+\mathcal{I}_\theta^{-1}|^{-1/2} \propto |\boldsymbol{D}+gI_{r_1}|^{-1/2} \in [(g+d_u)^{-r_1/2},(g+d_l)^{-r_1/2}],$$

so

$$h(g) \lesssim \frac{(d_u+g)^{(n-p_1)/2}(d_l+g)^{-r_1/2}}{[(d_u+g)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \lesssim \frac{(d_u+g)^{(n-p_1-r_1)/2}}{[(d_u+g)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}.$$

Similarly, we can find the lower bound

$$h(g) \gtrsim \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}.$$

Now, we prove Lemma 3 arguing by cases.
Case $\nu_0 > \nu_1$ Applying the lower bound in Lemma 11,

$$B_{10} \gtrsim \frac{[\mathsf{SSE}_0+\mathsf{SSR}]^{(n-p_0)/2}}{(\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}})^{(n-p_1)/2}} \int_0^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g).$$

Since $p_0 < p_1$, the term outside the integral goes to infinity as $\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty$, and by Fatou's lemma,

$$\begin{aligned}&\lim \inf_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \int_0^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2}}{[(g+d_l)\frac{\mathsf{SSE}_1}{\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}}+1]^{(n-p_1)/2}}\pi(\mathrm{d}g)\\ &\geq \int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g),\end{aligned}$$

which is clearly bounded away from 0 for any prior on g with positive support, so any such prior yields an information consistent B_{10} whenever $\nu_0 > \nu_1$.

Case $\nu_0 = \nu_1$ Applying the lower bound in Lemma 11 and Fatou's lemma as we did for the case $\nu_0 > \nu_1$:

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} B_{10} \gtrsim \int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\mathsf{SSE}_0+\mathsf{SSR}]^{(n-p_0)/2}}{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}.$$

The limit is $O(1)$, so a sufficient condition for information consistency is

$$\int_0^\infty (g+d_l)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) \asymp \int_0^\infty (g+1)^{(n-p_1-r_1)/2}\pi(\mathrm{d}g) = \infty,$$

as required.

Case $\nu_0 < \nu_1$ In this case, we apply the upper bound in Lemma 11:

$$B_{10} \lesssim \frac{[\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2}}{(\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}})^{(n-p_1)/2}} \int_0^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}} + 1]^{(n-p_1)/2}} \pi(\mathrm{d}g).$$

The term outside the integral goes to 0, so a necessary condition for information consistency is that the integral be infinite. We can bound the integral:

$$\int_0^\infty \frac{(g+d_u)^{(n-p_1-r_1)/2}}{[(g+d_u)\frac{\mathsf{SSE}_1}{\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}} + 1]^{(n-p_1)/2}} \pi(\mathrm{d}g) \leq \int_0^\infty (g+d_u)^{(n-p_1-r_1)/2} \pi(\mathrm{d}g),$$

so a necessary condition for information consistency is

$$\int_0^\infty (g+d_u)^{(n-p_1-r_1)/2} \pi(\mathrm{d}g) \asymp \int_0^\infty (g+1)^{(n-p_1-r_1)/2} \pi(\mathrm{d}g) = \infty,$$

as required.

B Proof of Lemma 4

Throughout, we use the notation in "Appendix A".

Case 1. Suppose there exists $M < \infty$ such that for all $g \geq M$, $\pi(g) \gtrsim g^{-\alpha}$ for $\alpha > 1$ and $p_0 > p_1$. Then, we apply the lower bound in Lemma 11:

$$B_{10} \gtrsim [\mathsf{SSE}_0 + \mathsf{SSR}]^{(n-p_0)/2} \int_M^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \mathrm{d}g.$$

Now, note that for any $K, d > 0$ with $1 - d < K$,

$$\begin{aligned} 0 &\leq \lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} \int_{\min(0,1-d)}^K \frac{(g+d)^{(n-p_1-r_1)/2-\alpha}}{[(g+d)\mathsf{SSE}_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \mathrm{d}g \\ &\lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} [\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{-(n-p_1)/2} = 0, \end{aligned}$$

so

$$\begin{aligned} &\lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} \int_M^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \mathrm{d}g \\ &\quad = \lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} \int_{1-d_l}^\infty \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1 + \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \mathrm{d}g. \end{aligned}$$

Plugging in:

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} B_{10} \gtrsim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} [\mathsf{SSE}_0+\mathsf{SSR}]^{(n-p_0)/2} \int_{1-d_l}^{\infty} \frac{(g+d_l)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_l)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}\mathrm{d}g$$
$$\propto \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{(\mathsf{SSE}_0+\mathsf{SSR})^{(n-p_0)/2}}{\mathsf{SSE}_1^{(n-p_1)/2}} {}_2F_1\left(\frac{n-p_1}{2}, \frac{r_1}{2}+\alpha-1; \frac{r_1}{2}+\alpha; \frac{-\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}}{\mathsf{SSE}_1}\right).$$

Using the identity

$${}_2F_1(a,b;c;z) = (1-z)^{-b}\,{}_2F_1\left(b, c-a; c; \tfrac{z}{z-1}\right),$$

we have

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} B_{10} \gtrsim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{(\mathsf{SSE}_0+\mathsf{SSR})^{(n-p_0)/2}\,{}_2F_1\left(\frac{r_1}{2}+\alpha-1, \frac{r_1-(n-p_1)}{2}+\alpha; \frac{r_1}{2}+\alpha; R^2\right)}{\mathsf{SSE}_1^{(n-p_1)/2}\left[1+\frac{\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}}{\mathsf{SSE}_1}\right]^{(r_1/2)+\alpha-1}},$$

where $R^2 = \widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}/(\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+\mathsf{SSE}_1) \to 1$ as $\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty$. If $\alpha < (n-p_1-r_1)/2+1$ (which is satisfied because $\alpha < (n-p_0-r_1)/2$ and $p_0 > p_1$ by assumption), the limit of the hypergeometric function as $R^2 \to 1$ is a constant (by Gauss' theorem). From here, it is immediate to conclude that B_{10} is information consistent whenever the lower bound is infinite, which occurs for $\alpha < (n-p_0-r_1)/2+1$, as required.

Case 2. Suppose there exists $M' < \infty$ such that for all $g \geq M'$, $\pi(g) \lesssim g^{-\alpha}$ for $\alpha > 1$ and $p_0 > p_1$. Then, by Lemma 11:

$$B_{10} \lesssim [\mathsf{SSE}_0+\mathsf{SSR}]^{(n-p_0)/2} \int_M^{\infty} \frac{(g+d_u)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_u)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}\mathrm{d}g.$$

As argued in Case 1,

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \int_M^{\infty} \frac{(g+d_u)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_u)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}\mathrm{d}g$$
$$= \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \int_{1-d_u}^{\infty} \frac{(g+d_u)^{(n-p_1-r_1)/2-\alpha}}{[(g+d_u)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}}\mathrm{d}g,$$

and carrying out the same computations as in Case 1:

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} B_{10} \lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{(\mathsf{SSE}_0+\mathsf{SSR})^{(n-p_0)/2}\,{}_2F_1\left(\frac{r_1}{2}+\alpha-1, \frac{r_1-(n-p_1)}{2}+\alpha; \frac{r_1}{2}+\alpha; R^2\right)}{\mathsf{SSE}_1^{(n-p_1)/2}\left[1+\frac{\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}}{\mathsf{SSE}_1}\right]^{(r_1/2)+\alpha-1}}.$$

If $(n-p_0-r_1)/2+1 \leq \alpha < (n-p_1-r_1)/2+1$, the limit of the hypergeometric function is $O(1)$ and B_{10} is information inconsistent. If $\alpha \geq (n-p_1-r_1)/2+1$, the necessary condition of Lemma 3 implies that B_{10} is information inconsistent.

Therefore, B_{10} is information inconsistent whenever $\alpha \geq (n - p_0 - r_1)/2 + 1$, as required.

C Proof of Lemma 5

Break the integral in B_{10} into the two regions $R_1 = \{\boldsymbol{\theta} : |\boldsymbol{\theta}|^2 \leq |\hat{\boldsymbol{\theta}}|\}$ and $R_2 = \{\boldsymbol{\theta} : |\boldsymbol{\theta}|^2 > |\hat{\boldsymbol{\theta}}|\}$. It is easy to see that, for any fixed $\epsilon > 0$, there is a K_ϵ such that, for $|\hat{\boldsymbol{\theta}}| > K_\epsilon$ and $\boldsymbol{\theta} \in R_1$,

$$\begin{aligned}(1-\epsilon)\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}} &< \left(\nu_1 s_1^2 + s_{\mathbf{y}}^2 + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)^{-\frac{n-r_2+\nu_1}{2}} \\ &< (1+\epsilon)\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}}.\end{aligned}$$

Thus, letting $P(R_1)$ denote the probability of R_1 under the $N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega})$ density, it follows that, for $|\hat{\boldsymbol{\theta}}| > K_\epsilon$,

$$\begin{aligned}&(1-\epsilon)\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}P(\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}} P(R_1) \\ &< \int_{R_1}\left(\nu_1 s_1^2 + s_{\mathbf{y}}^2 + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)^{-\frac{n-r_2+\nu_1}{2}} N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega})\mathrm{d}\boldsymbol{\theta} \\ &< (1+\epsilon)\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}} P(R_1).\end{aligned}$$

As $|\hat{\boldsymbol{\theta}}| \to \infty$, the integral over R_2 is clearly going to zero exponentially fast, while $P(R_1) \to 1$. Since ϵ can be chosen arbitrarily small, it follows that, as $|\hat{\boldsymbol{\theta}}| \to \infty$,

$$\frac{\int\left(\nu_1 s_1^2 + s_{\mathbf{y}}^2 + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)^{-\frac{n-r_2+\nu_1}{2}} N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega})d\boldsymbol{\theta}}{\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}}} \to 1.$$

Thus, as $|\hat{\boldsymbol{\theta}}| \to \infty$,

$$B_{10} \to \lim_{|\hat{\boldsymbol{\theta}}|\to\infty} \frac{C_2\left(\hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_1}{2}}}{\left(\nu_0 s_0^2 + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}'\mathbf{X}_{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1\hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu_0}{2}}},$$

from which the results stated in the lemma follow directly.

D Proof of Lemma 6

Using the notation in "Appendix A" and applying Lemma 11:

$$B_{10} \gtrsim \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{n-p_1/2}} \frac{(g + d_l)^{(n-p_1-r_1)/2}}{[(g + d_l)\mathsf{SSE}_1 / \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}} + 1]^{(n-p_1)/2}}$$

For $g > 0$, the right-hand side is maximized at $\widehat{g} = \max(0, (n - p_1 - r_1)\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}} / (r_1 \mathsf{SSE}) - d_l)$. Then,

$$\begin{aligned} \lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} \max_{g \geq 0} B_{10} &\gtrsim \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{(n-p_1)/2}} \frac{(\widehat{g} + d_l)^{(n-p_1-r_1)/2}}{[(\widehat{g} + d_l)\mathsf{SSE}_1 / \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}} + 1]^{(n-p_1)/2}} \\ &\propto \lim_{\|\widehat{\boldsymbol{\theta}}\|^2 \to \infty} \frac{(\mathsf{SSE}_0 + \mathsf{SSR})^{(n-p_0)/2}}{[\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{r_1/2} \mathsf{SSE}^{(n-p_1-r_1)/2}} \\ &= \infty, \end{aligned}$$

so the adaptive prior is information consistent.

E Proof of Lemma 7

The marginal posterior of $\boldsymbol{\theta}$ in the joint space has a multivariate Student t distribution with mean $(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \hat{\boldsymbol{\theta}}$, scale matrix $(n + \nu - r_2)^{-1}(s^2\nu + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}'((\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1} + \boldsymbol{\Omega})^{-1} \hat{\boldsymbol{\theta}})(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 + \boldsymbol{\Omega}^{-1})^{-1}$, and $n + \nu - r_2$ degrees of freedom. Change variables to

$$\boldsymbol{\xi} = (n + \nu - r_2)^{1/2}(s^2\nu + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}'((\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1} + \boldsymbol{\Omega})^{-1} \hat{\boldsymbol{\theta}})^{-1/2} \boldsymbol{\theta},$$

which has a multivariate Student t distribution with mean

$$\boldsymbol{\xi}^* = \frac{(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \hat{\boldsymbol{\theta}}}{(n + \nu - r_2)^{-1/2}(s^2\nu + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}'((\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1} + \boldsymbol{\Omega})^{-1} \hat{\boldsymbol{\theta}})^{1/2}},$$

scale matrix $(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 + \boldsymbol{\Omega}^{-1})^{-1}$, and $n + \nu - r_2$ degrees of freedom. Note that

$$P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y}) = P_\pi(\boldsymbol{\xi} \leq \mathbf{0} \mid \mathbf{y}).$$

It is easy to see that $\boldsymbol{\xi}^*$ lies in a fixed compact set C for any $\hat{\boldsymbol{\theta}}$, from which it is immediate that $P_\pi(\boldsymbol{\xi} \leq \mathbf{0} \mid \mathbf{y})$ is bounded away from 0 and 1.

The second part of the lemma follows immediately from letting $c \to \infty$ in the expression for $\boldsymbol{\xi}^*$.

F Proof of Lemma 8

Throughout, we use the notation in "Appendix A".

Sufficient condition:

We start with the case where there exists $\widehat{\theta}_i \to +\infty$; we treat the case where all $\widehat{\theta}_i \to -\infty$ later.

We can write:

$$\begin{aligned}\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \le \mathbf{0} \mid \boldsymbol{y}) &= \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \int_0^\infty P(\boldsymbol{\theta} \le \mathbf{0} \mid g, \boldsymbol{y}) p(g \mid \boldsymbol{y}) \mathrm{d}g \\ &= \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{1}{p(\boldsymbol{y})} \int_0^\infty P(\boldsymbol{\theta} \le \mathbf{0} \mid g, \boldsymbol{y}) p(\boldsymbol{y} \mid g) \pi(\mathrm{d}g) \\ &\propto \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{1}{p(\boldsymbol{y})} \int_0^\infty P(\boldsymbol{\theta} \le \mathbf{0} \mid g, \boldsymbol{y}) h(g)\, \pi(\mathrm{d}g),\end{aligned}$$

with h as defined in Lemma 11 (but noting that, in this case, the notation is $\nu_1 = \nu$). Letting $p = \nu - r_2$ and using the upper bound in Lemma 11, we obtain

$$\begin{aligned}&\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \le \mathbf{0} \mid \boldsymbol{y}) \\ &\lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})} \int_0^\infty \frac{P(\boldsymbol{\theta} \le \mathbf{0} \mid g, \boldsymbol{y})\, (g + d_u)^{(n-p-r_1)/2}}{[(g + d_u)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega}^{-1} \widehat{\boldsymbol{\theta}} + 1]^{(n-p)/2}} \pi(\mathrm{d}g)\end{aligned}$$

From Lemma 7, we know that

$$P(\boldsymbol{\theta} \le \mathbf{0} \mid g, \boldsymbol{y}) = P(\boldsymbol{\xi} \le \mathbf{0} \mid g, \boldsymbol{y}),$$

where $\boldsymbol{\xi}$ has a multivariate Student t distribution, with location and scale

$$\begin{aligned}\boldsymbol{m} &= \frac{(n+\nu-r_2)^{1/2}\, \boldsymbol{w}}{[\mathsf{SSE}_1 + \widehat{\boldsymbol{\theta}}' (\mathcal{I}_\theta^{-1} + g\boldsymbol{\Omega})^{-1} \widehat{\boldsymbol{\theta}}]^{1/2}} \\ \boldsymbol{S} &= (\mathcal{I}_\theta + \boldsymbol{\Omega}^{-1}/g)^{-1},\end{aligned}$$

where

$$\boldsymbol{w} = (\mathcal{I}_\theta + \boldsymbol{\Omega}^{-1}/g)^{-1} \mathcal{I}_\theta \widehat{\boldsymbol{\theta}}.$$

We factor

$$\boldsymbol{S} = \boldsymbol{\Omega}^{1/2} (\boldsymbol{\Omega}^{1/2} \mathcal{I}_\theta \boldsymbol{\Omega}^{1/2} + I_{r_1}/g)^{-1} \boldsymbol{\Omega}^{1/2} = \boldsymbol{\Omega}^{1/2} \boldsymbol{O}' (\boldsymbol{D}^{-1} + I_{r_1}/g)^{-1} \boldsymbol{O} \boldsymbol{\Omega}^{1/2},$$

where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{D}$ is diagonal (with positive entries) as defined in Lemma 11. Therefore, for a fixed coordinate j,

$$\boldsymbol{S}_{jj} \in \left[\frac{g}{g/d_l + 1} \boldsymbol{\Omega}_{jj}, \frac{g}{g/d_u + 1} \boldsymbol{\Omega}_{jj} \right],$$

so $0 < \boldsymbol{S}_{jj} < \infty$ for $g > 0$. Using the same factorizations, we obtain $\|\boldsymbol{w}\|^2 \propto \widehat{\boldsymbol{\theta}}' \boldsymbol{\Omega} \boldsymbol{\Omega} \widehat{\boldsymbol{\theta}}$ for $g > 0$. Plugging this in and factorizing the denominator in $\boldsymbol{m}$ in a similar manner, we obtain

$$\begin{aligned} \boldsymbol{m} &= \frac{(n+\nu-r_2)^{1/2}\,\|\boldsymbol{w}\|}{[\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'(\mathcal{I}_\theta^{-1}+g\boldsymbol{\Omega})^{-1}\widehat{\boldsymbol{\theta}}]^{1/2}}\,\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \\ &\propto \frac{(\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}\boldsymbol{\Omega}\widehat{\boldsymbol{\theta}})^{1/2}}{[\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}'(\mathcal{I}_\theta^{-1}+g\boldsymbol{\Omega})^{-1}\widehat{\boldsymbol{\theta}}]^{1/2}}\,\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}. \end{aligned}$$

If we choose a coordinate j such that $w_j > 0$ (which exists by assumption), using the lower bound in Lemma 11, we obtain

$$\boldsymbol{m}_j \gtrsim \frac{(g+d_l)^{1/2}(\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}\boldsymbol{\Omega}\widehat{\boldsymbol{\theta}})^{1/2}}{\left[(g+d_l)\mathsf{SSE}_1+\widehat{\boldsymbol{\theta}}\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}\right]^{1/2}} \gtrsim \frac{(g+d_l)^{1/2}}{\left[(g+d_l)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1\right]^{1/2}}$$

Now,

$$\begin{aligned} \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \le \mathbf{0} \mid \boldsymbol{y}) &\lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})} \\ &\int_0^\infty \frac{P(T_{n-p} \ge \boldsymbol{m}_j/\sqrt{\boldsymbol{S}_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g) \end{aligned}$$

where T_{n-p} is a central Student t with $n-p$ degrees of freedom. Let $\varepsilon > 0$, then

$$\int_0^\varepsilon \frac{P(T_{n-p} \ge \boldsymbol{m}_j/\sqrt{\boldsymbol{S}_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g) \le (\varepsilon+d_u)^{(n-p-r_1)/2},$$

so

$$\begin{aligned} \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \le \mathbf{0} \mid \boldsymbol{y}) &\lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})} \\ &\int_\varepsilon^\infty \frac{P(T_{n-p} \ge \boldsymbol{m}_j/\sqrt{\boldsymbol{S}_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g). \end{aligned}$$

Therefore, we can plug in our bounds for $\boldsymbol{m}_j$ and $\boldsymbol{S}_{jj}$, which are bounded away from 0 whenever $g > 0$. Using the tail bound

$$P(T_{n-p} \ge x) \lesssim \frac{1}{x(1+x^2/\nu)^{(n-p-1)/2}} \lesssim x^{-(n-p)}$$

and our previous work, we obtain

$$\begin{aligned} \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \le \mathbf{0} \mid \boldsymbol{y}) &\lesssim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})} \int_\varepsilon^\infty (g+d_u)^{-r_1/2}\,\pi(\mathrm{d}g) \\ &\propto \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})}. \end{aligned}$$

Clearly

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}\frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})}=0\Leftrightarrow\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}\,p(\boldsymbol{y})=\infty$$

and

$$\begin{aligned}
\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}\,p(\boldsymbol{y}) &= \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}\int_0^\infty p(\boldsymbol{y}\mid g)\,\pi(\mathrm{d}g)\\
&\propto \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}\int_0^\infty h(g)\,\pi(\mathrm{d}g)\\
&\gtrsim \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}\int_0^\infty\frac{(g+d_l)^{(n-p-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\pi(\mathrm{d}g)\\
&\gtrsim \int_0^\infty\liminf_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}\frac{(g+d_l)^{(n-p-r_1)/2}}{[(g+d_l)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\pi(\mathrm{d}g)\\
&= \int_0^\infty(g+d_l)^{(n-p-r_1)/2}\pi(\mathrm{d}g)\\
&\asymp \int_0^\infty(g+1)^{(n-p-r_1)/2}\pi(\mathrm{d}g).
\end{aligned}$$

Therefore, if the integral above is infinite, $\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}P(\boldsymbol{\theta}\le\mathbf{0}\mid\boldsymbol{y})=0$, as required.

Now we turn to the case where $\widehat{\theta}_i\to-\infty$ for all i, in which case we assume that $w_i<0$ for all i. Then, a Fréchet bound ensures that

$$P(\boldsymbol{\theta}\le\mathbf{0}\mid\boldsymbol{y})=P(\theta_1\le0,\theta_2\le0,\ \ldots,\theta_{r_1}\le0\mid\boldsymbol{y})\ge\sum_{i=1}^{r_1}P(\theta_i\le0\mid\boldsymbol{y})-(r_1-1).$$

Therefore,

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}P(\theta_i\ge0\mid\boldsymbol{y})=0,\ 1\le i\le r_1\Rightarrow\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}P(\boldsymbol{\theta}\le\mathbf{0}\mid\boldsymbol{y})=1.$$

Then, we can work with the conditional probabilities exactly as we did for the previous case:

$$\begin{aligned}
&\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}P(\theta_i\ge0\mid\boldsymbol{y})\\
&=\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}\int_0^\infty P(\theta_i\ge0\mid g,\boldsymbol{y})p(g\mid\boldsymbol{y})\,\mathrm{d}g\\
&\lesssim\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}\frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}}{p(\boldsymbol{y})}\int_\varepsilon^\infty\frac{P(T_{n-p}\ge-\boldsymbol{m}_j/\sqrt{\boldsymbol{S}_{jj}})\,(g+d_u)^{(n-p-r_1)/2}}{[(g+d_u)\mathsf{SSE}_1/\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}+1]^{(n-p)/2}}\,\pi(\mathrm{d}g).
\end{aligned}$$

Since $-\boldsymbol{m}_j$ is positive, the subsequent steps in the proof for the previous case allow us to conclude that $\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}P(\theta_i\ge0\mid\boldsymbol{y})=0$, as required.

Necessary condition:

In the sequel, we assume that there is at least one i such that $\widehat{\theta}_i \to +\infty$. The case where all coordinates go to $-\infty$ can be dealt with the same way we did for the sufficient condition. We can write:

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y}) = \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}\int_0^\infty P(\boldsymbol{\theta}\leq\mathbf{0}|g,\mathbf{y})p(\mathbf{y}|g)\pi(\mathrm{d}g)}{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}p(\mathbf{y})}.$$

First, we show that the limit of the numerator is bounded away from 0. Applying Fatou's lemma and one of the bounds in Lemma 11, we obtain

$$\begin{aligned}
&\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \frac{\int_0^\infty P(\boldsymbol{\theta} \leq \mathbf{0} \mid g, \mathbf{y})p(\mathbf{y} \mid g)\pi(\mathrm{d}g)}{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{-(n-p)/2}} \\
&\geq \int_0^\infty \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \inf \frac{P(\boldsymbol{\theta} \leq \mathbf{0} \mid g, \mathbf{y})\, h(g)}{[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}}\,\pi(\mathrm{d}g) \\
&\gtrsim \int_0^\infty (g+d_l)^{(n-p-r_1)/2} \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \inf P(\boldsymbol{\theta} \leq \mathbf{0} \mid g, \mathbf{y})\pi(\mathrm{d}g),
\end{aligned}$$

and for any g,

$$\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \inf P(\boldsymbol{\theta} \leq \mathbf{0} \mid g, \mathbf{y}) = \lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} \inf P(\boldsymbol{\xi} \leq \mathbf{0} \mid g, \mathbf{y})$$

where $\boldsymbol{\xi}$ is a multivariate Student t as in Lemma 7. Lemma 7 shows that $P(\boldsymbol{\xi} \leq \mathbf{0} \mid g, \mathbf{y})$ is bounded away from 0, which implies that the numerator is bounded away from 0, as claimed. A necessary condition for $\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty} P(\boldsymbol{\theta} \leq \mathbf{0} \mid y) = 0$ is that $\lim_{\|\widehat{\boldsymbol{\theta}}\|^2\to\infty}[\widehat{\boldsymbol{\theta}}'\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\theta}}]^{(n-p)/2}p(\mathbf{y}) = \infty$ which, as we saw in the proof of the sufficient condition, is equivalent to

$$\int_0^\infty (g+1)^{(n-p-r_1)/2}\pi(\mathrm{d}g) = \infty,$$

as required.

G Proof of Lemma 9

The second part of the Bayes factor in (13) can be expressed as $\left(P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y})^{-1} - 1\right) = \frac{k(\boldsymbol{\Theta}_1)}{k(\boldsymbol{\Theta}_0)}$, where

$$k(\boldsymbol{\Theta}_t) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_t} \left(\nu s^2 + s_{\mathbf{y}}^2 + \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)' \mathbf{X}_1' \boldsymbol{\Sigma}^{-1}\mathbf{X}_1 \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right)^{-\frac{n-r_2+\nu}{2}} N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega}, \boldsymbol{\Theta}_t)\mathrm{d}\boldsymbol{\theta},$$

and $N(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Omega}, \boldsymbol{\Theta}_t)$ denotes a truncated multivariate normal density for $\boldsymbol{\theta}$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$, truncated in the subspace $\boldsymbol{\Theta}_t$ for $t = 0$ or 1. Exactly as in the proof of Lemma 5 it can be shown that $k(\boldsymbol{\Theta}_t) =$

$\left(\nu s^2 + s_{\mathbf{y}}^2 + \hat{\boldsymbol{\theta}}' \mathbf{X}_{\boldsymbol{\theta}}' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1 \hat{\boldsymbol{\theta}}\right)^{-\frac{n-r_2+\nu}{2}} (1 + o(1))$ in the limit, so that $\left(P_\pi(\boldsymbol{\theta} \leq \mathbf{0} \mid \mathbf{y})^{-1} - 1\right) \to 1$.

H Proof of Lemma 10

The marginal posterior of $\boldsymbol{\theta}$ in the joint space has a multivariate Student t distribution with mean $\frac{g}{g+1}\hat{\boldsymbol{\theta}}$, scale matrix $(n-r_2)^{-1}(s_{\mathbf{y}}^2 + (g+1)^{-1}\hat{\boldsymbol{\theta}}'(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)\hat{\boldsymbol{\theta}}) \frac{g}{g+1} (\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1}$, and $n - r_2$ degrees of freedom. A change of variables to $\boldsymbol{\xi} = \frac{g+1}{g}\boldsymbol{\theta}$ results in a multivariate Student t distribution with mean $\hat{\boldsymbol{\theta}}$, scale matrix $(n - r_2)^{-1}((1 + g^{-1})s_{\mathbf{y}}^2 + g^{-1}\hat{\boldsymbol{\theta}}'(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)\hat{\boldsymbol{\theta}})(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1}$, and degrees of freedom $n - r_2$. Note that the posterior probability is invariant under this transformation, i.e., $P_\pi(\boldsymbol{\theta} \leq \mathbf{0}|\mathbf{y}) = P_\pi(\boldsymbol{\xi} \leq \mathbf{0}|\mathbf{y})$. Furthermore, it is important to note that the factor $(1 + g^{-1})s_{\mathbf{y}}^2 + g^{-1}\hat{\boldsymbol{\theta}}'(\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)\hat{\boldsymbol{\theta}}$ in the scale matrix of $\boldsymbol{\xi}$ is a monotonically decreasing function of g. Now it is easy to see that if $\hat{\boldsymbol{\theta}} \leq \mathbf{0}$, $P_\pi(\boldsymbol{\xi} \leq \mathbf{0}|\mathbf{y})$ monotonically increases as the scales decrease, and if $\hat{\boldsymbol{\theta}} \not\leq \mathbf{0}$, $P_\pi(\boldsymbol{\xi} \leq \mathbf{0}|\mathbf{y})$ monotonically decreases as the scales decrease. Thus, in order to maximize B_{01} if $\hat{\boldsymbol{\theta}} \leq \mathbf{0}$, and maximize B_{10} if $\hat{\boldsymbol{\theta}} \not\leq \mathbf{0}$, we have to let g go to ∞. For completeness, note the marginal posterior of $\boldsymbol{\theta}$ in the joint space with a multivariate Student t distribution with mean $\hat{\boldsymbol{\theta}}$, scale matrix $(n - r_2)^{-1} s_{\mathbf{y}}^2 (\mathbf{X}_1' \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1}$, and $n - r_2$ degrees of freedom, in the limit as $g \to \infty$. Thus, even though a (data-based) adaptive prior is considered, the choice of g that maximizes the Bayes factor does not depend on the data. Note that taking the limit $g \to \infty$ was already considered by Mulder (2014a) but not in the context of an adaptive prior.

References

Bayarri MJ, García-Donato G (2007) Extending conventional priors for testing general hypotheses in linear models. Biometrika 95:135–152

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

Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison (with discussion). In: Lahiri P (ed) Model selection, monograph series, vol 38. Institute of mathematical statistics lecture notes edn. Institute of Mathematical Statistics, Beachwood Ohio, pp 135–207

Berger JO, Mortera J (1999) Default Bayes factors for nonnested hypothesis testing. J Am Stat Assoc 94:542–554

Berger JO, Pericchi LR, Varshavsky JA (1998) Bayes factors and marginal distributions in invariant situations. Sankhyā Ser A 60:307–321

Dickey J (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann Stat 42:204–223

Fan T, Berger JO (1992) Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. Stat Decis 10:99–120

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman & Hall, London

George E, Foster DP (2000) Calibration and empirical bayes variable selection. Biometrika 87(4):731–747

Gu X, Mulder J, Decovic M, Hoijtink H (2014) Bayesian evaluation of inequality constrained hypotheses. Psychol Methods 19(4):511

Hansen MH, Yu B (2001) Model selection and the principle of minimum description length. J Am Stat Assoc 96(454):746–774

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

Klugkist I, Hoijtink H (2007) The Bayes factor for inequality and about equality constrained models. Comput Stat Data Anal 51:6367–6379

Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of *g* priors for Bayesian variable selection. J Am Stat Assoc 103(481):410–423

Moran GE, Ročková V, George EI (2018) Variance prior forms for high-dimensional Bayesian variable selection. Bayesian Anal 14:1091–1119

Mulder J (2014a) Bayes factors for testing inequality constrained hypotheses: issues with prior specification. Br J Math Stat Psychol 67:153–171

Mulder J (2014b) Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. Comput Stat Data Anal 71:448–463

Mulder J, Hoijtink H, Klugkist I (2010) Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. J Stat Plan Inference 140:887–906

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of *p*-values for testing precise null hypotheses. Am Stat 55:62–71

Som A, Hans CM, MacEachern SN (2016) A conditional lindley paradox in Bayesian linear models. Biometrika 103(4):993–999

Zellner A, Siow A (1980) Posterior odds ratios for selected regression hypotheses. University Press, Valencia, pp 585–603

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g prior distributions. In: Goel PK, Zellner A (eds) Bayesian inference and decision techniques-essays in honor of Bruno de Finetti. Elsevier, Amsterdam, North-Holland, pp 233–243