Extending Conventional priors for Testing General Hypotheses in Linear Models

María Jesús Bayarri
University of Valencia
46100 Valencia, Spain

Gonzalo García-Donato *
University of Castilla-La Mancha
02071 Albacete, Spain

March 7, 2005

*M.J. Bayarri is Professor, Department of Statistics and O.R., University of Valencia, 46100 Valencia, Spain (email: susie.bayarri@uv.es); G. García-Donato is Assistant Professor, Department of Economy, University of Castilla-La Mancha, 02071 Albacete, Spain (email: gonzalo.garciadonato@uclm.es). This research was supported in part by the Spanish Ministery of Science and Education, under grant MTM2004-03290.

Abstract

In this paper, we consider that observations $\boldsymbol{Y}$ come from a general normal linear model and that it is desired to test a simplifying (null) hypothesis about the parameters. We approach this problem from an objective Bayesian, model selection perspective. Crucial ingredients for this approach are 'proper objective priors' to be used for deriving the Bayes factors. Jeffreys-Zellner-Siow priors have shown to have good properties for testing null hypotheses defined by specific values of the parameters in full rank linear models. We extend these priors to deal with general hypotheses in general linear models, not necessarily full rank. The resulting priors, which we call 'conventional priors', are expressed as a generalization of recently introduced 'partially informative distributions'. The corresponding Bayes factors are fully automatic, easy to compute and very reasonable. The methodology is illustrated for two popular problems: the change point problem and the equality of treatments effects problem. We compare the conventional priors derived for these problems with other objective Bayesian proposals like the intrinsic priors. It is concluded that both priors behave similarly although interesting subtle differences arise. Finally, we accommodate the conventional priors to deal with non nested model selection as well as multiple model comparison.

Key words and phrases: ANOVA models; Change Point problem; Model Selection; Objective Bayesian methods; Partially Informative Distributions; Regression models.

1. INTRODUCTION

In this paper we address the testing of general hypotheses in a general normal linear model as a Bayesian model selection problem. The Bayesian solution (the posterior probabilities of the hypotheses) is expressed in terms of the *Bayes factors*. In general, let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a random vector for which two competing models $M_1 : \boldsymbol{Y} \sim f_1(\boldsymbol{y} \mid \boldsymbol{\theta}_1)$ and $M_2 : \boldsymbol{Y} \sim f_2(\boldsymbol{y} \mid \boldsymbol{\theta}_2)$ are proposed. For the observed $\boldsymbol{y}$, the Bayes factor for M_1 (and against M_2) can be simply expressed as the ratio between the marginal distributions for $\boldsymbol{Y}$ evaluated at the observed $\boldsymbol{y}$, that is:

$$B_{12} = \frac{m_1(\boldsymbol{y})}{m_2(\boldsymbol{y})},$$

where

$$m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\, \pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i,$$

and $\pi_i(\boldsymbol{\theta}_i)$ is the prior distribution for $\boldsymbol{\theta}_i$ under model M_i , $i = 1, 2$. It is well known that the posterior odds (of M_1 to M_2) is given by the prior odds times the Bayes factor B_{12}. For the purposes of this paper, we will concentrate only on the derivation of B_{12}; we will not discuss assessment of prior probabilities for each model and consider them as given. Of course, if both models have equal prior probabilities (a common assessment in objective Bayesian testing), then B_{12} coincides with the posterior odds. A comprehensive survey on the interpretation and applicability of the Bayes factors can be found in Kass and Raftery (1995).

Often, it is desired to employ Bayesian methods that use only the information in $\boldsymbol{y}$ and the models (that is, no external subjective information is incorporated). We refer to them as *Objective Bayesian* methods, but they have also received other names in the literature: default, automatic, non informative, minimum informative, reference, etc. Objective Bayes methods are usually evaluated based on their behavior with respect to several criteria, both Bayesian and frequentist. A sensible and intuitive criteria consists in investigating whether an objective Bayes technique corresponds to a genuine Bayes technique with respect to a sensible 'objective' prior (Berger and Pericchi 2001). Several methods for objective model selection exist which directly derive objective Bayes factors; in this case, the previous criteria should be separately investigated. Another possibility is to propose sensible objective priors, derive the Bayesian answers and judge whether the proposed priors are 'good'.

In problems with no model uncertainty (estimation or prediction problems for a given model, to be called simply 'estimation problems' for short), much is known about good default priors; for instance, the "reference priors" of Berger and Bernardo (1992) have shown to have very good properties in a wide variety of problems. However, the question of which objective (default) prior to use in the presence of model uncertainty (model choice, hypothesis testing, model averaging) is still largely an open question; only partial answers are known. In general, familiar 'objective' priors appropriate for estimation problems are seriously inadequate for model selection problems. In particular, improper priors (reference priors and others) cannot be used in general since they produce arbitrary Bayes factors (see, for example, O'Hagan 1994). Notable exceptions to this general rule are investigated in Berger, Pericchi and Varshavsky (1998); see also Liang, Paulo, Molina, Clyde and Berger (2005).

There are several proposals for objective priors in Bayesian model selection, including the *intrinsic priors* (Berger and Pericchi 1996; Moreno, Bertolino and Racugno 1998); the *expected posterior priors* (Pérez 1998; Pérez and Berger 2002); the *unit information priors* (Kass and Wasserman 1995); in the context of generalized linear models, the *reference set of proper priors* (Raftery 1996); for covariate selection in regression models, the *Jeffreys-Zellner-Siow (JZS)* priors (Jeffreys 1961; Zellner and Siow 1980; Zellner and Siow 1984); also in this context, *g-priors* (Zellner 1986) are very often used, although they were originally derived for estimation problems.

The JZS prior distributions have a number of desirable properties making them reasonable priors to use in Bayesian variable selection (see Berger and Pericchi 2001). In fact, they are often used as a bench-mark for comparison with other objective Bayes proposals (expected posterior priors, intrinsic priors, etc). In particular, and under certain conditions, JZS priors are consistent (in the sense of asymptotically choosing the true model) whereas other default Bayesian methods are inconsistent (see Berger, Ghosh and Mukhopadhyay 2003 for details). Jeffreys (1961) first derived these priors to test $H_1 : \mu = 0$ versus $H_2 : \mu \neq 0$, where μ is

the mean of a normal distribution with unknown variance σ^2. Jeffreys extensively argued that a number of requirements should be met by any 'sensible' objective prior for testing in this scenario. Specifically, he proposed to use the "simplest function" which: i) is centered at zero (i.e. centered at H_1); ii) has scale σ; iii) is symmetric around zero and iv) has no moments. Zellner and Siow (1980, 1984) extended the idea to deal with covariate selection in full rank linear models. In this paper, we also work in the general linear model framework but do not require the design matrix to be of full rank. We also extend JZS priors to problems in which the simpler hypotheses (models) are given by general linear restrictions. Hence, in particular we provide a Bayesian alternative to the F classical tests (see Ravishanker and Dey 2002).

Following the terminology in Berger and Pericchi (2001), we use the term "conventional prior" distributions (CPD) to refer to priors that have been derived under the philosophy in JZS proposals. Bayes factors resulting from conventional priors will be called "conventional Bayes factor". We will show that, in the context of general linear models, the conventional Bayes factors are fully automatic, very easy to compute and, more important, very reasonable.

The main criticism of CPD's is that, since they are not derived from any general methodology, they are not applicable to scenarios other than the normal one. However, García-Donato (2003) shows that the CPD's can be seen to be a special case of a general set of prior distributions based on Jeffreys-Kullback-Leibler divergence between the competing models. These prior distributions, applied to model selection in non normal (and even irregular) problems have produced very promising results.

This paper is organized as follows: in Section 2, the original arguments of JZS are reviewed. In Section 3, the idea is extended to derive conventional Bayes factors for general lineal models (whether or not of full rank) and with the simpler model defined by any linear combination of the parameters. In Section 4 explicit expressions for the CPD's will be given. Interestingly, the conventional prior distributions can be expressed in terms of a generalization of the *the partially informative distributions* recently introduced for spatial and nonparametric settings (Sun, Tsutakawa and Speckman 1999; Speckman and Sun 2003; see also Ibrahim and Laud 1994). In Section 5 we apply the results to two interesting problems: the *change point problem* and the *equality of effects problem*. Formal extensions to the more general problem of multiple and non nested model selection are provided in Section 6. Finally, conclusions are given in Section 7.

2. JEFFREYS-ZELLNER-SIOW (JZS) PROPOSALS

Let $\boldsymbol{X}_1$, an $n \times k_1$ matrix (denoted $\boldsymbol{X}_1 : n \times k_1$) and $\boldsymbol{X}_e : n \times k_e$ be such that $\boldsymbol{X}_2 = (\boldsymbol{X}_1 \ , \ \boldsymbol{X}_e)$ has full column rank (i.e. $\text{rank}(\boldsymbol{X}_2) = r(\boldsymbol{X}_2) = k_1 + k_e$). In a variable selection problem, we

consider the two competing models (N_n denotes a n-variate normal density):

$$\begin{aligned} M_1 &: f_1(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n) \\ M_2 &: f_2(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_e\boldsymbol{\beta}_e, \sigma^2\boldsymbol{I}_n), \end{aligned} \tag{1}$$

where the parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_e$ have dimensions k_1 and k_e respectively. These models are often called 'regression' or 'full' rank models.

This problem is often alternatively expressed as that of testing the hypotheses $H_1 : \boldsymbol{\beta}_e = \mathbf{0}$ versus $H_2 : \boldsymbol{\beta}_e \neq \mathbf{0}$. Note that $(\boldsymbol{\beta}_1, \sigma)$ may have different meaning under M_1 than under M_2. Hence, in general and without extra conditions, using the same prior (informative or not) for $(\boldsymbol{\beta}_1, \sigma)$ in M_1 and M_2 is not reasonable (see Berger and Pericchi 2001 for a deeper discussion).

To clearly expose the ideas of JZS, let us begin by assuming that $\boldsymbol{X}_1^t\boldsymbol{X}_e = \mathbf{0}$. Under this assumption, the 'common' parameters $\boldsymbol{\beta}_1, \sigma$ are orthogonal (the Fisher information matrix is block-diagonal) to $\boldsymbol{\beta}_e$ in M_2. In this scenario, Zellner and Siow (1980, 1984), extending the pioneering work of Jeffreys (1961), propose for prior π_i under model $M_i,\ i = 1, 2$:

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$

and

$$\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1} Ca_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\boldsymbol{X}_e^t\boldsymbol{X}_e)^{-1}),$$

where Ca_{k_e} denotes a k_e-variate Cauchy density. Note that both π_1 and π_2 give the same, non-informative prior to both the standard deviation, and the coefficient of $\boldsymbol{X}_1$. The intuitive reasoning behind this assessment is based in two broadly accepted, intuitive arguments:

(i) Common orthogonal parameters have the same meaning across models. Hence, they can be given the *same* prior distribution.

(ii) The Bayes factor is not very sensitive to the (common) prior used for the common orthogonal parameters (see Jeffreys 1961; Kass and Vaidyanathan 1992).

Hence, from (i) and (ii), it seems fine to assess the *same* improper prior distributions for the common parameters. With that choice, the arbitrary constant multiplying these priors would be the same for both models, and since they occur in both, the numerator and denominator of Bayes factor, they cancel out, producing a well defined Bayes factor. A more rigorous argument based on invariance is given in Berger, Pericchi and Varshavsky (1998).

The 'non common' parameter, $\boldsymbol{\beta}_e$ appearing only in M_2 can not be given an improper prior (the arbitrary constant does not cancel in the Bayes factor). The proposal of JZS is to use a Cauchy prior, centered (and spiked) around the simpler model (the value $\mathbf{0}$ in this case) and with no moments (some advantages of priors with no moments for model selection are reviewed

in Liang et al. 2005). Moreover, this prior also has a 'right' type of scale, since it is oriented like the likelihood, but it is much wider.

In the more general case $\boldsymbol{X}_1^t\boldsymbol{X}_e \neq \boldsymbol{0}$, a linear reparameterization of the problem (1) is proposed by Zellner and Siow to reproduce the previous (orthogonal) situation, leading to the (conventional) priors

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$

and

$$\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1} Ca_{k_e}(\boldsymbol{\beta}_e \mid \boldsymbol{0}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}),$$

where

$$\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e, \quad \boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t. \tag{2}$$

$\boldsymbol{P}_1$ is the orthogonal projection operator onto $\mathcal{C}(\boldsymbol{X}_1)$, the space spanned by the columns of $\boldsymbol{X}_1$, so that $\boldsymbol{V}$ is the orthogonal projection of $\boldsymbol{X}_e$ onto $\mathcal{C}^{\perp}(\boldsymbol{X}_1)$, the orthogonal complement of $\mathcal{C}(\boldsymbol{X}_1)$.

Expressing the Cauchy distribution as an scale mixture of normal distributions, produces simple and intuitive expressions for the conventional Bayes factor in terms of the usual residual sums of squares for both models:

Proposition 1. *Let SSE_i, $i = 1, 2$, denote the residual sum of squares:*

$$SSE_i = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{P}_i)\boldsymbol{y}, \quad \boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i^t\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^t, \quad i = 1, 2,$$

where $\boldsymbol{X}_2 = (\boldsymbol{X}_1\ ,\ \boldsymbol{X}_e) : n \times k$, $(k = k_1 + k_e)$. Then the conventional Bayes factor depends on the data only through SSE_1 and SSE_2 and can be expressed in the following two alternative ways:

$$\begin{aligned} B_{21} &= \int \left(1 + t\, n\, \frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1 + t\, n)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt \\ &= \int \left(1 + t\, \frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1 + t)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{n}{2})\, dt, \end{aligned}$$

where $IGa(x|a, b)$ denotes an inverse gamma density at x, proportional to $x^{-(a+1)} \exp\{-b/x\}$.

Proof. It is straightforward. Details can be found in García-Donato (2003). □

Thus, B_{21} can be trivially computed by just evaluating a one-dimensional integral. Alternatively, a Monte Carlo approximation is also trivial:

$$B_{21} \approx \frac{1}{N}\sum_{i=1}^{N} L(t_i), \qquad L(t) = \left(1 + t\, n\, \frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1 + t\, n)^{(n-k)/2},$$

and $t_1, t_2, \ldots, t_N$ are simulations from an $IGa(1/2, 1/2)$ distribution.

Liang et al. (2005), propose the following Laplace approximation of B_{21}:

$$B_{21} \approx \sqrt{2\pi}\,\tilde{v}\left(1 + n\,\hat{t}\,\frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2}(1 + n\,\hat{t})^{(n-k)/2}\,IGa(\hat{t} \mid \frac{1}{2}, \frac{1}{2}),$$

where $\hat{t}$ is the (real) positive solution of the cubic equation:

$$t^3\frac{SSE_2}{SSE_1}n^2\big(k_1 - k - 3\big) + t^2 n\big(-k + n - 3 + \frac{SSE_2}{SSE_1}(k_1 - 3)\big) + t\big(n - 3 + n\frac{SSE_2}{SSE_1}\big) + 1 = 0,$$

and

$$\tilde{v} = \Big(-\frac{d^2}{dt^2}\log L(t) IGa(t \mid \frac{1}{2}, \frac{1}{2})\Big|_{t=\hat{t}}\Big)^{-1/2}.$$

Liang et al. (2005) show, through an extensive simulation study, the accuracy of the approximation above, and its superiority over the approximation proposed by Zellner and Siow (1980).

In the next sections, we generalize JZS's proposals to deal with a variety of problems in the general lineal model context. These generalizations are based on convenient reparameterizations of the original problem and we will need some formal notation. We say that $f^*(\boldsymbol{y} \mid \boldsymbol{\nu})$, $\boldsymbol{\nu} \in \Theta_\nu \subset \mathcal{R}^m$ is a reparameterization of $f(\boldsymbol{y} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^k$, with $\boldsymbol{\nu} = g(\boldsymbol{\theta})$, $g : \mathcal{R}^k \to \mathcal{R}^m$ if

$$f^*(\boldsymbol{y} \mid g(\boldsymbol{\theta})) = f(\boldsymbol{y} \mid \boldsymbol{\theta}).$$

We will generally refer to the parameterization in f as the "original parameterization" and that in f^* as the "convenient parameterization".

3. CONVENTIONAL BAYES FACTORS

In this section, conventional Bayes factors are derived for situations not immediately covered in JZS's formulation. First, in Section 3.1 we consider the full rank case (regression models) in which the simpler model is defined by a specific linear combination of the parameters (and not merely by a specific value). Then, in Section 3.2 we allow for a non full rank design (ANOVA models) where the simpler model is also characterized by linear combinations of the parameters (including as a particular case the testing of a fix value for the parameter vector).

3.1 Regression models

Let $\boldsymbol{X} : n \times k$ and $\boldsymbol{C} : k \times k_e$, $(k_e \leq k)$ have full column rank. It is desired to choose between the following models:

$$\begin{aligned} M_1 &: f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) :\ \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} \\ M_2 &: f_2(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n). \end{aligned} \tag{3}$$

Note that there is no lost of generality in assuming $r(\boldsymbol{C}) = k_e$, since M_1 would otherwise have redundant restrictions. Alternatively, we can cast the problem as that of testing:

$$H_1 : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}, \quad \text{vs} \quad H_2 : \boldsymbol{C}^t\boldsymbol{\beta} \neq \boldsymbol{0}.$$

In the following result we explicitly provide an alternative reparameterization to which the JZS proposal directly applies.

Proposition 2. *Let $\boldsymbol{A} : k \times (k - k_e)$ be any matrix so that $\boldsymbol{R}^t = (\boldsymbol{A} \ , \ \boldsymbol{C})$ is non singular. Let similarly partition $\boldsymbol{R}^{-1} : k \times k$ as $\boldsymbol{R}^{-1} = (\boldsymbol{S} \ , \ \boldsymbol{T})$ where $\boldsymbol{S}$ has dimension $\boldsymbol{S} : k \times (k - k_e)$ and $\boldsymbol{T} : k \times k_e$. Call $k_1 = k - k_e$ and define $\boldsymbol{X}_e = \boldsymbol{XT}$, $\boldsymbol{X}_1 = \boldsymbol{XS}$. Then the model selection problem (3) is equivalent to choosing between the models:*

$$\begin{aligned} M_1^* &: f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n) \\ M_2^* &: f_2^*(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_e\boldsymbol{\beta}_e, \sigma^2\boldsymbol{I}_n). \end{aligned} \tag{4}$$

Proof. See the Appendix. □

For simplicity of the notation, and without loss of generality, we assume that $|det\boldsymbol{R}| = 1$ (note that the problem is unaffected if $\boldsymbol{C}$ and $\boldsymbol{A}$ are multiplied by the same non zero constant.)

From Proposition 2, it follows that the conventional Bayes factor for the original problem (3) is the same as that for the reparameterized problem (4), which has been expressed in the form analyzed by JZS. Hence, we can directly use Proposition 1 to derive the Bayes factor:

$$B_{21} = \int \left(1 + tn\,\frac{SSE_2^*}{SSE_1^*}\right)^{-(n-k+k_e)/2} (1+tn)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

where SSE_i^*, is the residual sum of squares for model M_i^*, for $i = 1, 2$. Note that SSE_i^* (and so the Bayes factor) can depend on the choice of the arbitrary matrix $\boldsymbol{A}$. However, in Proposition 3 below, we express B_{21} in terms of statistics of the original problem, thus showing that, in fact the Bayes factor is not affected by the specific choice of the (arbitrary) matrix $\boldsymbol{A}$.

Proposition 3. *Let SSE_f and SSE_r be the residual sum of squares for the full (M_2) and restricted (M_1) models respectively. Then,*

$$B_{21} = \int \left(1 + tn\,\frac{SSE_f}{SSE_r}\right)^{-(n-k+k_e)/2} (1+tn)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt. \tag{5}$$

Proof. See the Appendix. □

The same comments that we made for B_{21} in the previous section apply here: (5) is very easy to compute and trivial to simulate. Hence, we have provided a general strategy to test

any linear combination of the parameter vector in regression problems (that is, full column rank design matrix). The testing of a specific value (JZS scenarios) is, of course, a particular case.

3.2 Analysis of Variance models

In this section, we derive conventional Bayes factors in the spirit of JZS when the design matrix is not of full column rank (ANOVA models) and the null model is characterized by a linear combination of the parameters.

To be more precise, we address the problem of choosing between models:

$$\begin{aligned} M_1 &: f_1(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma) = \{N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n) : \ \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0}\} \\ M_2 &: f_2(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma) = N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n), \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{X}} : n \times k$ is a matrix of rank r, with $r < k$ and $\tilde{\boldsymbol{C}}^t : k_e \times k$ is of rank k_e.

This model selection problem is often expressed as that of testing

$$H_1 : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0} \quad \text{vs} \quad H_2 : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}.$$

In the context of regression models, the design matrix $\boldsymbol{X}_2 = (\boldsymbol{X}_1 \ , \ \boldsymbol{X}_e)$ has full column rank, ensuring the existence of $(\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1}$ and $(\boldsymbol{X}_e^t (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e)^{-1}$. We show in the next Result that, in the ANOVA case, at least one of these inverses do not exist, precluding the direct application of JZS proposals.

Proposition 4. *Let $(\tilde{\boldsymbol{X}}_1 \ , \ \tilde{\boldsymbol{X}}_e)$ be any partition of $\tilde{\boldsymbol{X}}$. If $\tilde{\boldsymbol{X}}_1^t \tilde{\boldsymbol{X}}_1$ is nonsingular, then the matrix $\tilde{\boldsymbol{X}}_e^t (\boldsymbol{I}_n - \ \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e$ is singular.*

Proof. See the Appendix. □

We now look for an alternative, full rank, expression of the problem to which the results in the previous sections can be applied. We show a most interesting result, namely that non degenerate conventional Bayes factors exist for testing null (reduced) models defined by testable hypotheses (see Rencher 2000; Ravishanker and Dey 2002). Recall that $\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is a testable hypothesis if $\tilde{\boldsymbol{C}}^t \boldsymbol{G} \tilde{\boldsymbol{X}}^t \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{C}}^t$, where $\boldsymbol{G}$ is a generalized inverse of $\tilde{\boldsymbol{X}}^t \tilde{\boldsymbol{X}}$. Also, $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{E}$, $\boldsymbol{X} : n \times r$, $\boldsymbol{E} : r \times k$, is a full rank factorization of $\tilde{\boldsymbol{X}}$ if $r(\boldsymbol{X}) = r(\boldsymbol{E}) = r$.

Proposition 5. *The hypothesis $\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is testable if and only if there exists a full rank factorization of $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{E}$ such that $\tilde{\boldsymbol{C}}^t = \boldsymbol{C}^t \boldsymbol{E}$, for some matrices $\boldsymbol{X}, \boldsymbol{E}$ and $\boldsymbol{C}$.*

Note that the factorization in Proposition 5 (when it exists) it is not unique. The following result explicitly produces an alternative full rank formulation of model (6).

Proposition 6. *If there is a full rank factorization as described in Proposition 5, the problem (6) is equivalent to that of choosing between the models:*

$$\begin{aligned} M_1^* &: f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\}, \\ M_2^* &: f_2^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n). \end{aligned} \tag{7}$$

Proof. See the Appendix. □

When Proposition 6 applies, Proposition 3 directly provides the expression of the conventional Bayes factor for the problem (7), which coincides with the original problem (6):

$$B_{21} = \int \left(1 + tn\,\frac{SSE_f^*}{SSE_r^*}\right)^{-(n-r+k_e)/2} (1+tn)^{(n-r)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

where SSE_f^*, and SSE_r^* are the residual sums of squares in the full f_2^* and reduced f_1^* reparameterized models, respectively. In Theorem 1, we produce a convenient form of the Bayes factor in terms of the residual sums of squares for the original problem which implicitly also proofs the uniqueness of conventional Bayes factor, regardless of the matrices $\boldsymbol{X}, \boldsymbol{C}$ and $\boldsymbol{E}$ in Proposition 5.

Theorem 1. *Assume that there exists a full rank factorization in the sense described in Proposition 5, then the conventional Bayes factor is*

$$B_{21} = \int \left(1 + tn\,\frac{SSE_f}{SSE_r}\right)^{-(n-r+k_e)/2} (1+tn)^{(n-r)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt, \tag{8}$$

where where SSE_f, and SSE_r are the residual sums of squares in the original models f_2 and f_1, respectively.

Proof. See the Appendix. □

3.3 Summary

The formulations and derivations in the previous sections can be summarized in a very attractive, unified way as follows: Consider the model selection problem:

$$\begin{aligned} M_1 &: f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} \\ M_2 &: f_2(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \end{aligned} \tag{9}$$

where $\boldsymbol{X} : n \times k$ has rank $r(\boldsymbol{X}) = r$, $r \leq k$, $\boldsymbol{C}^t : k_e \times k$ has rank k_e and $\boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}$ is a testable hypothesis. The conventional Bayes factor is

$$B_{21} = \int \left(1 + t\,n\,\frac{SSE_f}{SSE_r}\right)^{-(n-r+k_e)/2} (1+t\,n)^{(n-r)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt, \tag{10}$$

where SSE_f and SSE_r are the residual sums of squares in the full M_2 and restricted M_1 models, respectively. Expression (10) is very easy to evaluate and depends only on very standard statistics, available from any statistical package. MC and Laplace approximations are very easy to compute, as shown in Section 2.

4. CONVENTIONAL PRIOR DISTRIBUTIONS

We now consider derivation of the conventional priors for the original formulation in terms of f_i, $i = 1, 2$. That is, we want to explicitly derive the priors producing the conventional Bayes factors of the previous section. Our main motivation is to judge the adequacy of the derived Bayes factors by studying the priors producing them. This is in the spirit of Berger and Pericchi (2001) who claimed that "one of the best ways of studying any biases in a procedure is by examining the corresponding prior for biases". Of course we might also wish to have the priors available for further statistical analyses.

Although it seems natural for a Bayesian to judge the adequacy of a procedure by studying how sensible the corresponding prior is, this does not seem to be routinely done in objective Bayesian model selection.

The general procedure to derive the conventional prior is simple: we typically know the conventional prior $\pi_i^*(\boldsymbol{\nu}_i)$ for the convenient reparameterization $f_i^*(\boldsymbol{y} \mid \boldsymbol{\nu}_i)$ of the original problem $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$, with $\boldsymbol{\nu}_i = g_i(\boldsymbol{\theta}_i)$, for $i = 1, 2$. It thus suffices to derive $\pi_i(\boldsymbol{\theta}_i)$ from $\pi_i^*(\boldsymbol{\nu}_i)$. Obviously, if g_i is a one to one transformation, then we merely use the usual transformation rule:

$$\pi_i(\boldsymbol{\theta}_i) = \pi_i^*(g_i(\boldsymbol{\theta}_i))\, |det\, \mathcal{J}_i(\boldsymbol{\theta}_i)|,$$

where $\mathcal{J}_i$ is the jacobian matrix of the transformation g_i. However, this is not always the case in the linear hypotheses studied previously, as when the dimension of $\boldsymbol{\nu}_i$ is less than the dimension of $\boldsymbol{\theta}_i$ (this happens for instance in the ANOVA problem.) When the reparameterization is not one-to-one, we derive π_i satisfying the (natural) condition that the predictive distributions in both the original and reparameterized models should be equal, that is:

$$\pi_i(\boldsymbol{\theta}_i): \qquad \int f_i^*(\boldsymbol{y} \mid \boldsymbol{\nu}_i)\pi_i^*(\boldsymbol{\nu}_i)d\boldsymbol{\nu}_i = \int f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \;\; i = 1, 2, \tag{11}$$

which also guarantees that B_{12} is unaffected by the reparameterization. Note that $\pi_i(\boldsymbol{\theta}_i)$ does not have to be unique.

Interestingly, we show that the conventional prior distributions (CPD) are closely related to the *Partially Informative Normal (PIN) Distributions*, originally introduced by Ibrahim and Laud (1994), and further studied (and named) by Sun, Tsutakawa and Speckman (1999) and Speckman and Sun (2003) (SS in what follows). A generalization of the PIN distributions will prove to be an attractive, unified way to express the CPD's. Furthermore, the term "Partially

Informative" reflects precisely the essence of the CPD's, which typically have, in the convenient reparameterization, improper distributions for the 'common' parameters, and proper distributions for the ones not occurring in the restricted model.

4.1 Partially Informative Distributions

We begin by introducing PIN distributions, we then generalize them and consider scale mixtures, which will be the representation we use for the CPD's.

Definition 1. *(SS) According to Speckman and Sun (2003) a random vector $\boldsymbol{Y} \in \mathcal{R}^n$ has a Partially Informative Normal distribution with parameter $\boldsymbol{A} : n \times n$, denoted by $PIN_n(\boldsymbol{A})$, if the joint density (possibly improper) of $\boldsymbol{Y}$ is of the form*

$$f(\boldsymbol{y} \mid \boldsymbol{A}) \propto det_+(\boldsymbol{A})^{1/2} \exp\{-\frac{1}{2}\, \boldsymbol{y}^t \boldsymbol{A} \boldsymbol{y}\},$$

where $\boldsymbol{A}$ is a symmetric nonnegative definite matrix, and $det_+(\boldsymbol{A})$ denotes the product of the positive eigenvalues of $\boldsymbol{A}$.

SS interpret PIN distributions as having two parts, a constant noninformative prior on the null space of $\boldsymbol{A}$, and a proper degenerate normal on the range of $\boldsymbol{A}$. This is precisely the type of priors one expects in conventional model selection of nested linear models, which spurred our interest on them. In SS formulation, PIN is only defined up to an arbitrary proportionality constant; for model selection purposes, however, we have to make explicit the constant for the 'proper' part. Moreover, it is easy to show that PIN's are not preserved under linear transformations, thus further limiting its practical usefulness in the conventional theory. However, we propose an easy generalization that overcomes both these difficulties:

Definition 2. *We say that the random vector $\boldsymbol{Y} \in \mathcal{R}^n$ has a Generalized Partially Informative Normal distribution with parameters $\boldsymbol{A} : n \times n$, and $\boldsymbol{C} : n \times n$, denoted by $GPIN_n(\boldsymbol{A}, \boldsymbol{C})$, if the joint density of $\boldsymbol{Y}$ (possibly improper) is of the form*

$$f(\boldsymbol{y} \mid \boldsymbol{A}, \boldsymbol{C}) = det_+\big(\frac{\boldsymbol{A}}{2\pi}\big)^{1/2} |det(\boldsymbol{C})| \exp\{-\frac{1}{2}\boldsymbol{y}^t(\boldsymbol{C}^t \boldsymbol{A} \boldsymbol{C})\boldsymbol{y}\},$$

where $\boldsymbol{A}$ is a symmetric nonnegative definite matrix and $\boldsymbol{C}$ is nonsingular.

In this definition the arbitrary integration constant was fixed so as to make GPIN's closed under linear transformations, and to reproduce the normal constant when GPIN reduces to a normal. GPIN's extend the definition by SS in the sense that, if $\boldsymbol{P}^t \boldsymbol{A} \boldsymbol{P} = \boldsymbol{D}$ is the *spectral decomposition* of $\boldsymbol{A}$, and $\boldsymbol{Y} \sim PIN_n(\boldsymbol{A})$, then $\boldsymbol{Y} \sim GPIN_n(\boldsymbol{D}, \boldsymbol{P}^t)$. We next establish some properties of GPIN distributions (the proofs are easy and hence omitted); the first one shows that GPIN's are close under linear transformations, and the second one the relationships between GPIN, PIN and normal distributions.

Result 1.

1. *If* $\boldsymbol{W} \sim GPIN_n(\boldsymbol{A}, \boldsymbol{C})$ *and* $\boldsymbol{U} : n \times n$ *is nonsingular, then* $\boldsymbol{Y} = \boldsymbol{U}^{-1}\boldsymbol{W} \sim GPIN_n(\boldsymbol{A}, \boldsymbol{CU})$.
2. *If* $\boldsymbol{Y} \sim GPIN_n(\boldsymbol{A}, \boldsymbol{I}_n)$ *then* $\boldsymbol{Y} \sim PIN_n(\boldsymbol{A})$. *If in addition* $\boldsymbol{A}$ *is positive definite, then* $GPIN_n(\boldsymbol{A}, \boldsymbol{I}_n) = PIN_n(\boldsymbol{A}) = N_n(\boldsymbol{0}, \boldsymbol{A}^{-1})$.

We next define the Cauchy-type counterpart of the GPIN distributions, which we simply call Partially Informative Cauchy distributions and represent by PIC:

Definition 3. *We say that* $\boldsymbol{Y} \in \mathcal{R}^n$ *has a Partially Informative Cauchy distribution with parameters* $\boldsymbol{A} : n \times n$, *and* $\boldsymbol{C} : n \times n$, *to be denoted by* $PIC_n(\boldsymbol{A}, \boldsymbol{C})$, *if the joint density (possibly improper) of* $\boldsymbol{Y}$ *is of the form*

$$f(\boldsymbol{y} \mid \boldsymbol{A}, \boldsymbol{C}) = \int GPIN_n(\boldsymbol{y} \mid \frac{\boldsymbol{A}}{t}, \boldsymbol{C})\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

where $\boldsymbol{A}$ *is a symmetric, nonnegative definite matrix and* $\boldsymbol{C}$ *is nonsingular.*

The next result, which we also present without proof, is the parallel for PIC of Result 1.

Result 2.

1. *If* $\boldsymbol{W} \sim PIC_n(\boldsymbol{A}, \boldsymbol{C})$, *and* $\boldsymbol{U} : n \times n$ *is nonsingular, then* $\boldsymbol{Y} = \boldsymbol{U}^{-1}\boldsymbol{W} \sim PIC_n(\boldsymbol{A}, \boldsymbol{CU})$.
2. *If* $\boldsymbol{A}$ *is positive definite, then* $PIC_n(\boldsymbol{A}, \boldsymbol{I}_n) = Ca_n(\boldsymbol{0}, \boldsymbol{A}^{-1})$.

The next example is crucial, since it expresses the JZS prior for the full rank, covariate selection problem as a PIC prior. This, along with the previous reparameterizations and the properties of the PIC distributions, will be the basis to derive the CPD for both regression (full rank) and ANOVA (not full rank) model selection problems with more general null models (defined by general linear combinations of the parameters). We deferred specific implementations till next Section.

Example 1. *In the general full rank, covariate selection problem (1), we recall that the priors proposed by Zellner and Siow were:*

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1}$$

and

$$\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1} Ca_{k_e}(\boldsymbol{\beta}_e \mid \boldsymbol{0}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}),$$

where $\boldsymbol{V}$ *was defined in (2),* $\boldsymbol{\beta}_1$ *(*$\boldsymbol{\beta}_e$*) is of dimension* k_1 *(*k_e*).*

It is straightforward to show that π_2 *can be expressed as*

$$\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \pi_{2.1}(\sigma)\pi_{2.2}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e \mid \sigma),$$

where

$$\pi_{2.1}(\sigma) = \sigma^{-1}, \quad \pi_{2.2}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e \mid \sigma) = PIC_k((\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_e^t)^t \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{I}_k),$$

$k = k_1 + k_e$, *and* $\boldsymbol{H} = \mathbf{0}_{k_1 \times k_1} \oplus (\boldsymbol{V}^t\boldsymbol{V})$ *is the direct sum of* $\mathbf{0}_{k_1 \times k_1}$ *and* $\boldsymbol{V}^t\boldsymbol{V}$. *(Recall, that* $\boldsymbol{A} \oplus \boldsymbol{B}$ *is the block diagonal matrix diag*$(\boldsymbol{A}, \boldsymbol{B})$ *;* $\mathbf{0}_{k \times k}$ *denotes the null matrix of dimension* $k \times k$*).*

4.2 Regression models

We now turn to regression (full rank) models in which the null model is characterized by linear functions of the parameters. We showed in Proposition 2 that the regression model selection problem (3):

$$M_1 : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \mathbf{0}\} \quad \text{vs} \quad M_2 : N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n),$$

can be reparameterized as the problem (4):

$$M_1^* : N_n(\boldsymbol{y} \mid \ \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n) \quad \text{vs} \quad M_2^* : N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_e\boldsymbol{\beta}_e, \sigma^2\boldsymbol{I}_n),$$

where (see proof of Proposition 2 in the Appendix) for M_1^*: $(\boldsymbol{\beta}_1, \sigma) = g_1(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \sigma)$, and for M_2^*: $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = g_2(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \boldsymbol{C}^t\boldsymbol{\beta}, \sigma)$, where $\boldsymbol{A}$ is defined in Proposition 2.

In what follows we let $1_k(\boldsymbol{\Psi} = \mathbf{0})$ denote the k dimensional density for $\boldsymbol{\Psi}$, degenerated at $\mathbf{0}$. The following proposition gives an explicit expression of the CPD's for the original parameterization.

Proposition 7. *Let* $\boldsymbol{X}_1, \boldsymbol{X}_e$ *and* $\boldsymbol{R}$ *be defined as in Proposition 2. The CPD's for problem (3) are given by:*

$$\pi_1(\boldsymbol{\beta}, \sigma) = \sigma^{-1}\, 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \mathbf{0}), \tag{12}$$

and

$$\pi_2(\boldsymbol{\beta}, \sigma) = \sigma^{-1} PIC_k(\boldsymbol{\beta} \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{R}),$$

where $\boldsymbol{H} = \mathbf{0}_{k_1 \times k_1} \oplus (\boldsymbol{V}^t\boldsymbol{V})$, $\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e$.

Proof. See the Appendix. □

Note that the CPD's distributions depend on the arbitrary matrix $\boldsymbol{A}$. However, the Conventional Bayes Factor does not, as shown in Section 3. We delay illustration till Section 5.

4.3 Analysis of Variance model

We next consider linear models with design matrices $\tilde{\boldsymbol{X}} : n \times k$ with rank $r < k$. We showed in Section 3.2, that if $\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \mathbf{0}$ is testable, the model selection problem (6)

$$M_1 : \{N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{I}_n) : \tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \mathbf{0}\} \quad \text{vs} \quad M_2 : N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{I}_n)$$

can be reparameterized as the full rank problem (7)

$$M_1^* : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} \quad \text{vs} \quad M_2^* : N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

where $\boldsymbol{X} : n \times r$ is of full column rank and, (see proof of Proposition 6 in the Appendix) $(\boldsymbol{\beta}, \sigma) = g(\tilde{\boldsymbol{\beta}}, \sigma) = (\boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma)$ for both M_1^* and M_2^*.

The following Proposition gives the explicit expression of the CPD for the original parameterization.

Proposition 8. *Assume that, in terms of matrices $\boldsymbol{X}, \boldsymbol{E}$ and $\boldsymbol{C}$, there exists a full rank factorization as described in Proposition 5. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_e$ be defined as in Proposition 2 (using the previous $\boldsymbol{X}$ and $\boldsymbol{C}$) and let $\boldsymbol{Q}_2 : k \times (k-r)$ be any matrix such that $\boldsymbol{Q} = (\boldsymbol{E}^t \ , \ \boldsymbol{Q}_2)$ is non singular. Then, the CPD's for problem (6) are*

$$\pi_1(\tilde{\boldsymbol{\beta}}, \sigma) = \sigma^{-1} \ 1_{k_e}(\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0}) \ h_{k-r}^1(\boldsymbol{Q}_2^t \, \tilde{\boldsymbol{\beta}}) \, |det\boldsymbol{Q}|,$$

and

$$\pi_2(\tilde{\boldsymbol{\beta}}, \sigma) = \sigma^{-1} \ PIC_r(\boldsymbol{E}\tilde{\boldsymbol{\beta}} \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{R}) \ h_{k-r}^2(\boldsymbol{Q}_2^t \tilde{\boldsymbol{\beta}}) \, |det\boldsymbol{Q}|,$$

where $\boldsymbol{H} = \boldsymbol{0}_{k_1 \times k_1} \oplus (\boldsymbol{V}^t\boldsymbol{V})$, $\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e$ and h_i^j, $i \in \mathcal{N}$, $j = 1, 2$ are arbitrary probability densities in $\mathcal{R}^i$.

Proof. See the Appendix. □

Note that the CPD's distributions depend on the arbitrary matrices defining the convenient parameterization. However, the Conventional Bayes Factor does not, as shown in Section 3. Also, note that π_i depends on arbitrary densities h^i. The use of proper densities which "complete", in some sense, the parametric space of the full-rank model is not inusual in Bayesian statistics. A good example are the MCMC strategies proposed in Carlin and Chib (1995), Han and Carlin (2001) and Dellaportas, Foster and Ntzoufras (2002). In that context, the functions h are called "pseudopriors", and as in here, choice of the pseudopriors does not affect the results, although good choices can lead to numerical improvements. We delay illustration till Section 5.

5. SPECIFIC APPLICATIONS

In this section, we derive the conventional Bayes factor and conventional prior distributions (CPD's) for two standard problems: the "change point problem" and the "equality of treatment effects problem". The conventional Bayes factors have straightforward expressions, while those for the CPD's are more involved. As we have seen, this is common to virtually any conventional model selection problem. However, the CPD's will provide interesting insights into the conventional methodology for these problems.

5.1 Change point problem

Let

$$\boldsymbol{Y}_a = \boldsymbol{X}_a\boldsymbol{\beta}_a + \boldsymbol{\epsilon}_a, \qquad \boldsymbol{\epsilon}_a \sim N_{n_a}(\boldsymbol{0}, \sigma_a^2\boldsymbol{I}_{n_a}),$$

and

$$\boldsymbol{Y}_b = \boldsymbol{X}_b\boldsymbol{\beta}_b + \boldsymbol{\epsilon}_b, \qquad \boldsymbol{\epsilon}_b \sim N_{n_b}(\boldsymbol{0}, \sigma_b^2\boldsymbol{I}_{n_b}),$$

where $\boldsymbol{Y}_i$ is of dimension n_i, $\boldsymbol{\beta}_i$ of dimension k, and $\boldsymbol{X}_i : n_i \times k$, has full column rank, for $i = a, b$. We assume homoscedasticity, i.e. $\sigma_a = \sigma_b = \sigma$. We want to test:

$$H_1 : \boldsymbol{\beta}_a = \boldsymbol{\beta}_b \quad \text{vs} \quad H_2 : \boldsymbol{\beta}_a \neq \boldsymbol{\beta}_b.$$

This testing problem is usually known as the *change point problem*, and has been widely treated in the statistical literature. Frequentist solutions are usually based on the F statistic (Chow's Test). A Bayesian solution from the objective model selection point of view is also given in Moreno, Torres and Casella (2005); their solution is not in terms of the conventional theory as in here, but in terms of the *intrinsic prior* (Berger and Pericchi 1996; Moreno, Bertolino and Racugno 1998).

For observed $\boldsymbol{y}_a$, $\boldsymbol{y}_b$, let $\boldsymbol{y}^t = (\boldsymbol{y}_a^t, \boldsymbol{y}_b^t)$. The change point problem can be expressed as the following model selection problem:

$$M_1 : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} \quad \text{vs} \quad M_2 : N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n),$$

where $\boldsymbol{X} = \boldsymbol{X}_a \oplus \boldsymbol{X}_b$, $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_a^t, \boldsymbol{\beta}_b^t)$, and $\boldsymbol{C}^t = (\boldsymbol{I}_k \ , \ -\boldsymbol{I}_k) : k \times 2k$. Note that $\boldsymbol{X} : n \times 2k$ with $n = n_a + n_b$, is of full column rank. The conventional Bayes factor is given by the usual expression

$$B_{21} = \int \left(1 + tn\,\frac{SSE_f}{SSE_r}\right)^{-(n-k)/2} (1+tn)^{(n-2k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

where in this case $SSE_f = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t)\boldsymbol{y}$ and

$$SSE_r = SSE_f + \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C}(\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}.$$

To derive the CPD's, a matrix $\boldsymbol{A}$ is needed, so that $\boldsymbol{R}^t = (\boldsymbol{A} \ , \ \boldsymbol{C})$ is non singular. We choose, for example, $\boldsymbol{A}^t = (\boldsymbol{I}_k \ , \ \boldsymbol{0}_{k\times k})$, with $|det(\boldsymbol{R}^t)| = 1$. Note that

$$\boldsymbol{R}^{-1} = (\boldsymbol{S} \ , \ \boldsymbol{T}) = \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0}_{k\times k} \\ \boldsymbol{I}_k & -\boldsymbol{I}_k \end{pmatrix},$$

and, by definition, $\boldsymbol{X}_e = \boldsymbol{X}\boldsymbol{T}$ and $\boldsymbol{X}_1 = \boldsymbol{X}\boldsymbol{S}$. Finally, following Proposition 7, the CPD's are:

$$\pi_1(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) = \sigma^{-1} 1_k(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b = \mathbf{0}),$$

and

$$\pi_2(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) = \sigma^{-1}\, PIC_{2k}((\boldsymbol{\beta}_a^t, \boldsymbol{\beta}_b^t)^t \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{R}) \tag{13}$$

where $\boldsymbol{H} = \mathbf{0}_{k\times k} \oplus (\boldsymbol{V}^t\boldsymbol{V})$, $\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e$ and $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t$. A little algebra produces a more intuitive, nicer expression for (13):

$$\pi_2(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) = \sigma^{-1} \frac{\Gamma(\frac{k+1}{2})}{\pi^{(k+1)/2}}\; det\Big(\frac{\boldsymbol{\Sigma}_c^{-1}}{n\sigma^2}\Big)^{1/2} \Big[1 + (\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)^t \frac{\boldsymbol{\Sigma}_c^{-1}}{n\sigma^2}(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)\Big]^{-\frac{k+1}{2}}, \tag{14}$$

where $\boldsymbol{\Sigma}_c = (\boldsymbol{X}_a^t\boldsymbol{X}_a)^{-1} + (\boldsymbol{X}_b^t\boldsymbol{X}_b)^{-1}$. These priors have very attractive forms for model selection: variances ('common' parameters) receive the invariant non-informative prior under each model. Also, one of the regression coefficients, say $\boldsymbol{\beta}_a$, can be argued to be 'common' to both models, and the CPD is, as intuitively expected, a uniform distribution under both priors. The conditional distribution of $\boldsymbol{\beta}_b$ given the other two parameters $(\boldsymbol{\beta}_a, \sigma)$ varies: under the null model (no change point), it is degenerate on $\boldsymbol{\beta}_a = \boldsymbol{\beta}_b$, while under the full model, is a Cauchy with scale $n\sigma^2\boldsymbol{\Sigma}_c$ Furthermore, this density depends on the $\boldsymbol{\beta}$'s and the design matrices only through a quantity of the same functional form as the the usual F-statistic for this problem, which is:

$$F = (\hat{\boldsymbol{\beta}}_a - \hat{\boldsymbol{\beta}}_b)^t \frac{\boldsymbol{\Sigma}_c^{-1}}{k\,\hat{\sigma}^2}(\hat{\boldsymbol{\beta}}_a - \hat{\boldsymbol{\beta}}_b),$$

where $(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b, \hat{\sigma}^2)$ denotes the least squares estimate (under M_2) of $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma^2)$.

We now compare the CPD above with an alternative objective prior proposal: the *intrinsic prior*, derived for this problem by Moreno, Torres and Casella (2005). For comparative purposes, we further elaborate its derivation. The intrinsic prior under model M_2 (derived from the non-informative prior $\pi^N(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) \propto \sigma^{-2}$) can be seen to be:

$$\begin{aligned}
\pi_2^I(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) &= \int_0^\infty \int_{\mathcal{R}^k} (u^2+\sigma^2)^{-3/2} \prod_{i=a,b} N_k(\boldsymbol{\beta}_i \mid \boldsymbol{s}, (u^2+\sigma^2)(\boldsymbol{Z}_i^t\boldsymbol{Z}_i)^{-1}) d\boldsymbol{s}\, du \\
&= \int_0^\infty (u^2+\sigma^2)^{-3/2} N_k(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b \mid \mathbf{0}, (u^2+\sigma^2)\boldsymbol{\Sigma}_I)\, du,
\end{aligned} \tag{15}$$

where

$$(\boldsymbol{Z}_i^t\boldsymbol{Z}_i)^{-1} = \frac{1}{L_i}\sum_{l=1}^{L_i} (\boldsymbol{X}_i^t(l)\boldsymbol{X}_i(l))^{-1},$$

and $\boldsymbol{X}_i(l)$ are submatrices of $\boldsymbol{X}_i$ of dimension $(k+1)\times k$, $i = a, b$ and $\boldsymbol{\Sigma}_I = (\boldsymbol{Z}_a^t\boldsymbol{Z}_a)^{-1} +$

$(\boldsymbol{Z}_b^t \boldsymbol{Z}_b)^{-1}$. The intrinsic prior under M_1 is $\pi_1^I(\boldsymbol{\beta}_a, \sigma) = \sigma^{-2}$.

The striking similarities between the general shapes of the conventional (14) and the intrisic (15) priors under M_2 can readily be seen, with only some minor differences. First, we had $\pi_2(\boldsymbol{\beta}_a, \sigma) = \sigma^{-1}$, whereas for the intrinsic prior, and given the choice for π^N in its derivation, it is natural to take $\pi_2^I(\boldsymbol{\beta}_a, \sigma) = \sigma^{-2}$. However, the main difference between both objective priors is in the proper conditional distributions of $\boldsymbol{\beta}_b$ given $(\boldsymbol{\beta}_a, \sigma)$, which in both cases are scale mixtures of normal distributions. For the conventional prior we have:

$$\pi_2(\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a, \sigma) = \int_0^\infty N_k(\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a,\, n\, t\, \sigma^2\, \boldsymbol{\Sigma}_c)\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

and for the intrinsic:

$$\begin{aligned} \pi_2^I(\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a, \sigma) &= \int_0^\infty (u^2+\sigma^2)^{-3/2} \sigma^2 N_k(\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a, (u^2+\sigma^2)\, \boldsymbol{\Sigma}_I) du = \\ &= \int_1^\infty N_k(\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a,\, t\, \sigma^2\, \boldsymbol{\Sigma}_I)\, m(t)\, dt, \end{aligned}$$

where, in this case the mixing function is

$$m(t) = \frac{1}{2}\, t^{-3/2} (t-1)^{-1/2}, \qquad t > 1.$$

Note that, somehow surprisingly, $m(t)$ is a proper density:

$$\int_1^\infty m(t)\, dt = 1.$$

The differences are in the scale of the normal (which is multiplied by n for the conventional prior, but it uses the full $\boldsymbol{X}_i$ matrices instead of the submatrices $\boldsymbol{Z}_i$ in the intrinsic), and in the mixing distributions, with noticeably heavier tails for the conventional mixing (note that, in the tails, $\text{IGa}(t \mid 1/2, 1/2) = O(t^{-3/2})$ while $m(t) = O(t^{-2})$). Neither the intrinsic nor the conventional mixing densities have moments. Figure 1, shows the mixing densities for both priors.

The result is that the conditional distributions for $\boldsymbol{\beta}_b \mid \boldsymbol{\beta}_a, \sigma$ have very similar shapes, the conventional prior being more disperse. These conditional distributions are displayed in Figure 2 when $\boldsymbol{X}_a = \boldsymbol{X}_b = \boldsymbol{1}_m$, where $\boldsymbol{1}_m^t = (1, 1, \overset{m}{\ldots}, 1)$.

In this problem, it turns out that after some algebraic manipulations, it is possible to integrate out σ (and t) in both π_2 and π_2^I, resulting in simple expressions for the (improper) joint marginal prior of $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$, which provides further insights into the behavior of these two

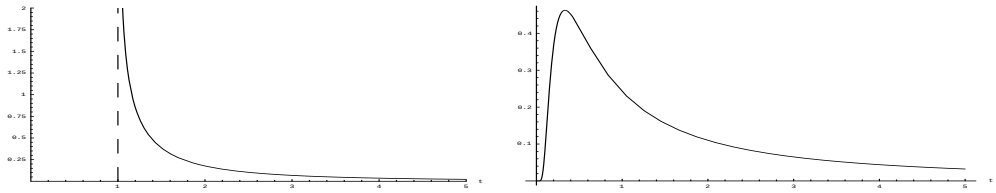

Figure 1: Mixing densities for the intrinsic prior (left) and the conventional prior (right)

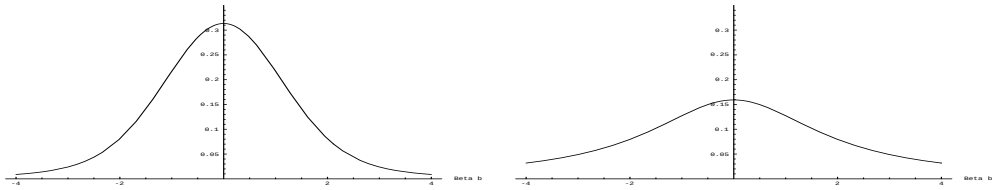

Figure 2: Intrinsic $\pi_2^I(\beta_b \mid \beta_a = 0, \sigma = 1)$ (left) and Conventional $\pi_2(\beta_b \mid \beta_a = 0, \sigma = 1)$ (right) with $\boldsymbol{X}_a = \boldsymbol{X}_b = \mathbf{1}_m$.

objective prior distributions. For the conventional prior, we get

$$\pi_2(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b) = \frac{det(\boldsymbol{\Sigma}_c)^{-1/2}\Gamma(\frac{k}{2})}{2\pi^{k/2}} \Big((\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)^t \frac{\boldsymbol{\Sigma}_c^{-1}}{n} (\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) \Big)^{-\frac{k}{2}}. \tag{16}$$

and for the intrinsic:

$$\pi_2^I(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b) = \frac{det(\boldsymbol{\Sigma}_I)^{-1/2}\Gamma(\frac{k+1}{2})}{2^{3/2}\pi^{(k-2)/2}} \Big((\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)^t \boldsymbol{\Sigma}_I^{-1} (\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) \Big)^{-\frac{k+1}{2}}. \tag{17}$$

The simplicity of the expressions, (16) and (17), allow for easy interpretation of the differences between the intrinsic and conventional priors. Note that both are negative powers of a similar quadratic function with the exponent in intrinsic prior, $-\frac{k+1}{2}$, being slightly smaller than the one in the conventional prior $-\frac{k}{2}$, which results in heavier tails for the conventional π_2. Also, π_2^I will be more peaked around $\boldsymbol{\beta}_a - \boldsymbol{\beta}_b = \mathbf{0}$ than π_2. Nevertheless, both proposals are very similar

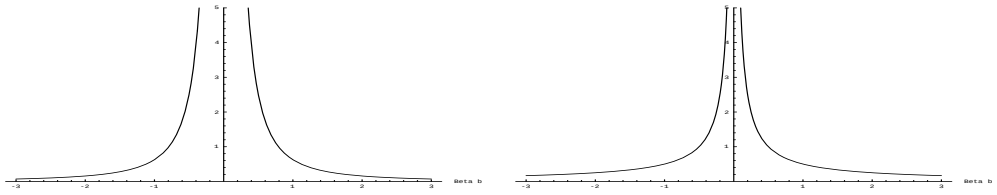

Figure 3: Intrinsic $\pi_2(\beta_b \mid \beta_a = 0)$ (left) and conventional $\pi_2(\beta_b \mid \beta_a = 0)$ (right) for $\boldsymbol{X}_a = \boldsymbol{X}_b = \mathbf{1}_m$ and $\pi_2(\beta_a) = 1$.

and it is expected that they provide similar results in a given situation. The derivation of the conventional priors is, however, somewhat simpler.

For the simple case $\boldsymbol{X}_a = \boldsymbol{X}_b = \mathbf{1}_m$, we have

$$\pi_2(\beta_a, \beta_b) = \frac{1}{2|\beta_a - \beta_b|} \quad \text{and} \quad \pi_2^I(\beta_a, \beta_b) = \frac{\pi^{1/2}}{2^{3/2}(\beta_a - \beta_b)^2}.$$

In Figure 3 we show the (improper) conditional distributions $\pi_2(\beta_b \mid \beta_a = 0)$ and $\pi_2^I(\beta_b \mid \beta_a = 0)$ (assuming $\pi_2(\beta_a) = 1$ in both cases). Again, the conventional prior can be seen to be more spiked around $\boldsymbol{\beta}_a$ and having heavier tails.

5.2 Equality of treatment effects

We next turn to a most traditional ANOVA problem, namely that of testing the equality of treatment effects. Let

$$Y_{ij} = \tilde{\mu} + \tilde{\tau}_i + \epsilon_{ij} \qquad i = 1, \ldots, a, \quad j = 1, \ldots, n_i,$$

where the ϵ_{ij} are i.i.d., $\epsilon_{ij} \sim N(0, \sigma^2)$. We are interested in testing the equality of the a treatment effects, that is, in testing:

$$H_1 : \tilde{\tau}_1 = \tilde{\tau}_2 = \cdots = \tilde{\tau}_a, \quad \text{vs} \quad H_2 : \tilde{\tau}_i \neq \tilde{\tau}_j, \text{ for at least one pair } (i, j).$$

Calling $n = \sum_{i=1}^a n_i$, $\boldsymbol{y}^t = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{a1}, \ldots, y_{an_a})$, and $\tilde{\boldsymbol{\tau}}^t = (\tilde{\mu}, \tilde{\tau}_1, \tilde{\tau}_2, \cdots, \tilde{\tau}_a)$ this problem can be written as the model selection problem:

$$M_1 : \{N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\tau}}, \sigma^2 \boldsymbol{I}_n) : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\tau}} = \mathbf{0}\} \quad \text{vs} \quad M_2 : N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\tau}}, \sigma^2 \boldsymbol{I}_n), \tag{18}$$

where $\tilde{\boldsymbol{X}} = (\mathbf{1}_n \ , \ \oplus_{i=1}^{a} \mathbf{1}_{n_i})$, $\tilde{\boldsymbol{C}}^t = (\mathbf{0}_{a-1} \ , \ \mathbf{1}_{a-1} \ , \ -\boldsymbol{I}_{a-1})$. Note that $\tilde{\boldsymbol{X}} : n \times (a+1)$ has rank a, and hence these models are not of full rank. However, it is well known that $\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\tau}} = \mathbf{0}$ is testable.

The Conventional Bayes factor, as given in (8), is

$$B_{21} = \int \left(1 + tn \, \frac{SSE_f}{SSE_r}\right)^{-(n-1)/2} (1+tn)^{(n-a)/2} IGa(t \mid \frac{1}{2}, \frac{1}{2}) \, dt, \tag{19}$$

where here

$$SSE_f = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a} n_i \bar{y}_i^2, \quad SSE_r = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2,$$

with $\bar{y} = \sum_i \sum_j y_{ij}/n$ and $\bar{y}_i = \sum_j y_{ij}/n_i$.

Another objective Bayes factor for this problem depending on the data only through the ratio of the residual sums of squares is the one derived in Spiegelhalter and Smith (1982). Their Bayes factor is computed with the usual non-informative priors and then the arbitrary constant is fixed by considering imaginary training samples. The resulting Bayes factor is:

$$B_{21}^{SM} = \left(\frac{(a+1)\prod_{i=1}^{a} n_i}{2n}\right)^{-1/2} \left(\frac{SSE_f}{SSE_r}\right)^{-n/2}.$$

As it is shown in Westfall and Gönen (1996), B_{21}^{SM} can be obtained as the limit of actual Bayes factors computed from proper priors. The Bayesian information criterion (BIC), for this problem, gives rise (see for instance Berger and Pericchi 2001) to the following approximation to an objective Bayes factor

$$B_{21}^{BIC} = n^{(1-a)/2} \left(\frac{SSE_f}{SSE_r}\right)^{-n/2}.$$

Kass and Wasserman (1995) justify use of B_{21}^{BIC} by showing that it approximately corresponds to the one obtained with the *unit information priors*. Note that both B_{21}^{BIC} and B_{21}^{SM} are only approximations to actual Bayes factors, while the conventional Bayes factor (19) is an actual Bayes factor, derived from prior distributions whose explicit forms can be easily derived. We next turn to the derivation of the CPD.

For the full rank factorization $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{E}$ of Proposition 5, we can take $\boldsymbol{E} = (\mathbf{1}_a \ , \ \boldsymbol{I}_a)$, $\boldsymbol{X} = \oplus_{i=1}^{a} \mathbf{1}_{n_i}$ and $\boldsymbol{C}^t = (\mathbf{1}_{a-1} \ , \ -\boldsymbol{I}_{a-1})$. Then $\boldsymbol{\tau} = \boldsymbol{E}\tilde{\boldsymbol{\tau}}$, with $\boldsymbol{\tau}^t = (\tau_1, \ldots, \tau_a)$ is the usual reparameterization $\tau_i = \tilde{\mu} + \tilde{\tau}_i$. Now, by Proposition 6, the problem can be expressed as the full rank model selection problem:

$$M_1^* : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\tau}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\tau} = \mathbf{0}\} \quad \text{vs} \quad M_2^* : N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\tau}, \sigma^2 \boldsymbol{I}_n).$$

For this full rank parameterization, the CPD's were derived in Proposition 7, and are given

by:

$$\begin{aligned}
\pi_1^*(\boldsymbol{\tau},\sigma) &= \sigma^{-1}\, 1_{a-1}(\boldsymbol{C}^t\boldsymbol{\tau}=\mathbf{0}),\\
\pi_2^*(\boldsymbol{\tau},\sigma) &= \sigma^{-1} PIC_a(\boldsymbol{\tau} \mid \frac{0\oplus(\boldsymbol{V}^t\boldsymbol{V})}{n\sigma^2}, \boldsymbol{R}) =\\
&= \sigma^{-1} Ca_{a-1}\Big((\tau_2,\ldots,\tau_a)^t \mid \tau_1\mathbf{1}_{a-1}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}\Big), \qquad (20)
\end{aligned}$$

where $\boldsymbol{R}^t = (\boldsymbol{A}\ ,\ \boldsymbol{C})$, and $\boldsymbol{A}$ is any matrix for which $|det\boldsymbol{R}| = 1$. More insights can be gained by considering the balanced case, $n_1 = n_2 = \cdots = n_a = m$, for which, taking $\boldsymbol{A}^t = (1\ ,\ \mathbf{0}_{a-1}^t)$, it can be shown that $(\boldsymbol{V}^t\boldsymbol{V})^{-1}$ is the intra class correlation matrix:

$$(\boldsymbol{V}^t\boldsymbol{V})^{-1} = \frac{1}{m}(\boldsymbol{I}_{a-1} + \mathbf{1}_{a-1}\mathbf{1}_{a-1}^t),$$

and a simpler, more intuitive expression for π_2^* can be derived, as given in the next Proposition:

Proposition 9. *For $i = 1,\ldots,a-1$, let*

$$\bar{\tau}_{i+1} = \frac{\sum_{l=i+1}^{a}\tau_l}{a-i}, \quad S_{i+1}^2 = \frac{\sum_{l=i+1}^{a}(\tau_l - \bar{\tau}_{i+1})^2}{a-i},$$

and

$$\Sigma_{i+1} = \frac{a(a-i+1)}{(a-i)^2}\,\sigma^2 + \frac{a-i+1}{a-i}\,S_{i+1}^2.$$

Then, the CPD (20), can be expressed as

$$\pi_2^*(\boldsymbol{\tau},\sigma) = \frac{1}{\sigma}\,\pi_{2.a}^*(\tau_a \mid \sigma)\,\prod_{i=1}^{a-1}\,\pi_{2.i}^*(\tau_i \mid \tau_{i+1},\ldots,\tau_a,\sigma)\,,$$

where $\pi_{2.a}^*(\tau_a \mid \sigma) = 1$, and

$$\pi_{2.i}^*(\tau_i \mid \tau_{i+1},\ldots,\tau_a,\sigma) = St_1(\tau_i \mid \bar{\tau}_{i+1}, \Sigma_{i+1}, a-i),$$

that is, a univariate Student with location parameter $\bar{\tau}_{i+1}$, scaled by Σ_{i+1} and with $a-i$ degrees of freedom (df).

Proof. See the Appendix. □

The marginal distribution for σ in this prior is the usual objective invariant prior; the conditional distribution $\pi_2^*(\boldsymbol{\tau} \mid \sigma)$ when expressed as the product of the following conditional distributions:

$$\pi_{2.a}^*(\tau_a \mid \sigma),\ \pi_{2.a-1}^*(\tau_{a-1} \mid \tau_a,\sigma),\ \ldots,\pi_{2.1}^*(\tau_1 \mid \tau_2,\ldots,\tau_a,\sigma),$$

each of which can be seen to be more informative than the previous one. Indeed, the first one,

$\pi_{2.a}(\tau_a \mid \sigma)$ is an improper (constant) distribution, the rest of them are Student distributions, with degrees of freedom increasing from one df in $\pi_{2.a-1}(\tau_{a-1} \mid \tau_a, \sigma)$ (a Cauchy distribution) to the $(a-1)$ df of $\pi_{2.1}(\tau_1 \mid \tau_2, \ldots, \tau_a, \sigma)$, which will be close to a normal distribution if the number of groups a is moderate or large. Each Student conditional distribution $\pi_{2.i}(\tau_i \mid \tau_{i+1}, \ldots, \tau_a, \sigma)$ has location equal to the mean of $\tau_{i+1}, \ldots, \tau_a$ and scale Σ_{i+1}, a linear combination of σ^2 and S^2_{i+1} (the variance of $\tau_{i+1}, \ldots, \tau_a$). This prior thus nicely demonstrates the "partially informative" nature of the CPD's. (Note that this is only a conveniently intuitive expression of the joint distribution of the τ's, which is, of course, independent of any ordering of the components.)

If wished, the CPD's for the original problem (18), where $\tilde{\boldsymbol{\tau}}^t = (\tilde{\mu}, \tilde{\tau}_1, \tilde{\tau}_2, \cdots, \tilde{\tau}_a)$ can also be derived. Applying Proposition 8 with $\boldsymbol{Q}_2^t = (1, 0, \overset{a}{\ldots}, 0)$, $(|det\boldsymbol{Q}| = 1)$ we get the CPD's:

$$\begin{aligned}
\pi_1(\tilde{\boldsymbol{\tau}}, \sigma) &= \sigma^{-1}\, h_1^1(\tilde{\mu})\; 1_{a-1}(\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\tau}} = \boldsymbol{0}), \\
\pi_2(\tilde{\boldsymbol{\tau}}, \sigma) &= \sigma^{-1}\, h_1^2(\tilde{\mu}) PIC_a(\boldsymbol{E}\tilde{\boldsymbol{\tau}} \mid \frac{0 \oplus (\boldsymbol{V}^t\boldsymbol{V})}{n\sigma^2}, \boldsymbol{R}) \\
&= \sigma^{-1}\, h_1^2(\tilde{\mu}) Ca_{a-1}\Big((\tilde{\tau_2}, \ldots, \tilde{\tau_a})^t \mid \tilde{\tau_1}\boldsymbol{1}_{a-1}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}\Big),
\end{aligned}$$

where h_1^j are arbitrary (proper) probability densities in $\mathcal{R}$. That is, π_2 has the usual invariant non-informative prior for $(\sigma, \tilde{\tau_1})$, $\tilde{\mu}$ gets an arbitrary proper prior, and the conditional of $(\tilde{\tau_2}, \ldots, \tilde{\tau_a})$ given $(\sigma, \tilde{\tau_1}, \tilde{\mu})$ is a Cauchy centered at the null and with the usual type of scale; again, an intuitively appealing prior for this problem.

6. MULTIPLE AND NON NESTED LINEAR MODEL SELECTION

Our proposal can easily be applied to problems of selecting among more than two linear models. In a similar way to generalizations of JZS priors to multiple models scenarios, we simply compare each model to a 'base' reference model to derive the relevant pairwise Bayes factors. In Liang et al. (2005) performance of JZS Bayes factors for two distinct base models (the 'null' or encompassed model, and the full or encompassing model) is investigated. We prefer to choose the null as base model for pairwise comparisons. This is also used, for example, in Pérez (1998).

Suppose we want to select among the following d competing models:

$$M_i : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) :\, \boldsymbol{C}_i^t\boldsymbol{\beta} = \boldsymbol{0}\}, \quad i = 1, 2, \ldots, d,$$

where $\boldsymbol{C}_i^t : s_i \times k$ has full row rank for all i. (This formulation includes comparison between non nested models .)

Let us define the 'null' model (and add it to the list of models in case it is not there)

$$M_0 : \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) :\, \boldsymbol{C}_i^t\boldsymbol{\beta} = \boldsymbol{0},\ i = 1, 2, \ldots, d\}.$$

Note that M_0 is nested in M_i for all i. The conventional priors corresponding to each pairwise

comparison M_i vs M_0, say π_i and π_{0i}, can be easily constructed. Which is appealing in this approach (when compared with others like comparison with the 'full model') is that π_{0i} does not change throughout the comparisons (i.e. $\pi_{0i} = \pi_{0j} = \pi_0$, say, $\forall i, j$) giving rise to a coherent Bayesian procedure in the sense that there is only one prior for each model.

Once the Bayes factors B_{k0}, $k = 1, \ldots, d$ have been obtained, the Bayes factor for comparing any two models M_i and M_j, B_{ij} can be simply computed as $B_{ij} = B_{i0}(B_{j0})^{-1}$.

We limit ourselves to deriving the conventional priors in the full rank scenario ($\boldsymbol{X} : n \times k$ has full column rank). The conventional priors for the ANOVA case are similar. In order to avoid unnecessary notational complexity we assume that $\boldsymbol{C} = (\boldsymbol{C}_1 \ , \ \boldsymbol{C}_2 \ , \cdots \ , \boldsymbol{C}_d)$ has rank $r(\boldsymbol{C}) = \sum_{j=1}^{d} s_i$. (If this does not hold, for example if $\boldsymbol{C}_2 = (\boldsymbol{C}_1 \ , \ \boldsymbol{C}^*)$, then $\boldsymbol{C}$ would be constructed with only the non redundant restrictions and some aesthetic adjustments would be needed in the expressions to follow.)

Let $\boldsymbol{A} : k \times k_0$ ($k_0 = k - \sum_j s_j$) be any matrix so that $\boldsymbol{R}^t = (\boldsymbol{A} \ , \ \boldsymbol{C})$ with $|det \boldsymbol{R}| = 1$ has inverse $\boldsymbol{R}^{-1} = (\boldsymbol{S} \ , \ \boldsymbol{T}_1 \cdots \boldsymbol{T}_d)$. Define the following matrices:

$$
\begin{array}{ll}
\boldsymbol{C}_{-i} = (\boldsymbol{C}_1 \cdots \boldsymbol{C}_{i-1} \ , \ \boldsymbol{C}_{i+1} \cdots \boldsymbol{C}_d), & \boldsymbol{X}_e^{-i} = \boldsymbol{X}(\boldsymbol{T}_1 \cdots \boldsymbol{T}_{i-1} \ , \ \boldsymbol{T}_{i+1} \cdots \boldsymbol{T}_d), \\
\boldsymbol{X}_0 = \boldsymbol{X}\boldsymbol{S}, & \boldsymbol{P}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^t \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^t, \\
\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{X}_e^{-i}, & \boldsymbol{H}_i = \boldsymbol{0}_{k_0 \times k_0} \oplus \boldsymbol{V}_i^t \boldsymbol{V}_i.
\end{array}
$$

By similar arguments to those used in the pairwise comparisons, it can be shown that the conventional priors (under the encompassed null model approach) are

$$
\pi_i(\boldsymbol{\beta}, \sigma) = \sigma^{-1} \, PIC_{k-s_i}\Big(\begin{pmatrix} \boldsymbol{A}^t \\ \boldsymbol{C}_{-i}^t \end{pmatrix} \boldsymbol{\beta} \mid \frac{\boldsymbol{H}_i}{n\sigma^2}, \boldsymbol{I}_{k-s_i} \Big) \ 1_{s_i}(\boldsymbol{C}_i^t \boldsymbol{\beta} = \boldsymbol{0}), \qquad i = 1, \ldots, d.
$$

7. CONCLUSIONS

The original conventional theory of Jeffreys (1961) and Zellner and Siow (1980, 1984) was formulated for the problem of covariate selection in (full rank) regression models. We have extended this theory by considering that the full model is not necessarily of full rank, and that the simpler model is defined more generally by a linear (testable) combination of the parameters.

From a practical point of view, we derive a unique expression for the conventional Bayes factor for all these testing problems. Moreover, the Bayes factor is defined in terms of standard statistics, widely available, and expressed as a unidimensional integral; hence, it is very easy to compute by either numerical or MC methods.

On a more theoretical side, we derive explicit expressions for the prior distributions (CPD's) producing these conventional Bayes factors, thus given Bayesian validity to the conventional Bayes factor. Moreover, the CPD's can be used to judge the adequacy of the conventional methodology in a specific problem. A generalization of the Partially Informative Distributions

defined by Speckmann and Sun (2003) has proven very useful to express the CPD's in a unified, attractive way.

We apply the results to two standard statistical problems: *change point* and the *equality of treatment effects*. Conventional Bayes factors are very easy to derive (always the case for conventional Bayes factors). Explicit expressions for the CPD's are also given. These conventional prior distributions are found to be of a very reasonable form, very intuitive, and nicely demonstrating the fundamental "partially informative" nature of these priors for objective Bayesian model selection.

APPENDIX: PROOFS

Proof of Proposition 2:
We show that f_i^* is a reparameterization of f_i, for $i = 1, 2$:

For M_1, let $(\boldsymbol{\beta}_1, \sigma) = g_1(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \sigma)$, then

$$\begin{aligned} f_1^*(\boldsymbol{y} \mid g_1(\boldsymbol{\beta}, \sigma)) &= f_1^*(\boldsymbol{y} \mid \boldsymbol{A}^t\boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{A}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \\ &= \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{S}\boldsymbol{A}^t\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{T}\boldsymbol{C}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} \\ &= \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\} = f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma). \end{aligned}$$

For M_2, let $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = g_2(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \boldsymbol{C}^t\boldsymbol{\beta}, \sigma)$, then

$$\begin{aligned} f_2^*(\boldsymbol{y} \mid g_2(\boldsymbol{\beta}, \sigma)) &= f_2^*(\boldsymbol{y} \mid \boldsymbol{A}^t\boldsymbol{\beta}, \boldsymbol{C}^t\boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{A}^t\boldsymbol{\beta} + \boldsymbol{X}_e\boldsymbol{C}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \\ &= N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{S}\boldsymbol{A}^t\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{T}\boldsymbol{C}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \\ &= N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) = f_2(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma). \end{aligned}$$

Proof of Proposition 3:
We have to show that $SSE_1^* = SSE_r$ and $SSE_2^* = SSE_f$.

Since the design matrix in M_2^* is $\boldsymbol{X}\boldsymbol{R}^{-1}$, we have

$$SSE_2^* = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{R}^{-1}\big((\boldsymbol{X}\boldsymbol{R}^{-1})^t(\boldsymbol{X}\boldsymbol{R}^{-1})\big)^{-1}\boldsymbol{X}\boldsymbol{R}^{-1})\boldsymbol{y} = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t)\boldsymbol{y} = SSE_f.$$

Now with $\boldsymbol{X}_2 = \boldsymbol{X}\boldsymbol{R}^{-1}$, and using a standard result (Searle 1997) we have

$$SSE_1^* = SSE_2^* + \boldsymbol{y}^t\boldsymbol{X}_2(\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1}\boldsymbol{C}_a(\boldsymbol{C}_a^t(\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1}\boldsymbol{C}_a)^{-1}\boldsymbol{C}_a^t(\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^t\boldsymbol{y},$$

with $\boldsymbol{C}_a^t = (\boldsymbol{0}_{k_e \times k_1}\ ,\ \boldsymbol{I}_{k_e}) = (\boldsymbol{R}^{-1})^t\boldsymbol{C}$. Then, since $SSE_2^* = SSE_f$:

$$SSE_1^* = SSE_f + \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C}(\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} = SSE_r.$$

Proof of Proposition 4:
Let $\tilde{\boldsymbol{X}}_e : n \times k_e$. We will show that

$$r(\tilde{\boldsymbol{X}}_e^t(\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e) < k_e.$$

Since $r(\tilde{\boldsymbol{X}}) = r < k$ and $\tilde{\boldsymbol{X}}_1$ has full column rank,there exists a vector $\boldsymbol{v}$ for which $\tilde{\boldsymbol{X}}_e = (\tilde{\boldsymbol{X}}_e^1\ ,\ \tilde{\boldsymbol{X}}_1\boldsymbol{v})$, where $\tilde{\boldsymbol{X}}_e^1 : n \times (k_e - 1)$. Now

$$(\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e = (\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)(\tilde{\boldsymbol{X}}_e^1\ ,\ \tilde{\boldsymbol{X}}_1\boldsymbol{v}) = \big((\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e^1\ ,\ \boldsymbol{0}_{n\times 1}\big),$$

and hence:

$$\begin{aligned} r(\tilde{\boldsymbol{X}}_e^t(\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e) = r((\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e) = r((\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e^1\ ,\ \boldsymbol{0}_{n\times 1}) = r((\boldsymbol{I} - \tilde{\boldsymbol{P}}_1)\tilde{\boldsymbol{X}}_e^1) \\ \leq r(\tilde{\boldsymbol{X}}_e^1) \leq k_e - 1 < k_e. \end{aligned}$$

Proof of Proposition 5:
Assume first that $\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is testable. Let

$$\boldsymbol{Q}^t\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0}_{r\times(k-r)} \\ \boldsymbol{0}_{(k-r)\times r} & \boldsymbol{0}_{(k-r)\times(k-r)} \end{pmatrix} \tag{A1}$$

be the spectral decomposition of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$, with $\boldsymbol{D}$ the diagonal matrix containing the r non null eigenvalues of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$. Consider the partition of $\boldsymbol{Q} = (\boldsymbol{E}^t\ ,\ \boldsymbol{Q}_2)$, where $\boldsymbol{E}^t : k \times r$ has rank r. Define $\boldsymbol{X} = \tilde{\boldsymbol{X}}\boldsymbol{E}^t$ and $\boldsymbol{C}^t = \tilde{\boldsymbol{C}}^t\boldsymbol{E}^t$. Note that $\boldsymbol{X} : n \times r$ has also rank r:

$$r(\boldsymbol{X}) = r(\tilde{\boldsymbol{X}}\boldsymbol{E}^t) = r(\tilde{\boldsymbol{X}}\boldsymbol{E}^t\ ,\ \boldsymbol{0}_{n\times(k-r)}) = r(\tilde{\boldsymbol{X}}\boldsymbol{Q}) = r(\tilde{\boldsymbol{X}}) = r.$$

We now show that $\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{C}}^t = \boldsymbol{C}^t\boldsymbol{E}$. First it is immediate to see that:

$$\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}\boldsymbol{E}^t\boldsymbol{E} = \tilde{\boldsymbol{X}}\boldsymbol{E}^t\boldsymbol{E} + \tilde{\boldsymbol{X}}\boldsymbol{Q}_2\boldsymbol{Q}_2^t = \tilde{\boldsymbol{X}}\boldsymbol{Q}\boldsymbol{Q}^t = \tilde{\boldsymbol{X}}\boldsymbol{I}_k = \tilde{\boldsymbol{X}}.$$

Next, since $\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is testable, then (Ravishanker and Dey 2002) $\tilde{\boldsymbol{C}}^t\boldsymbol{G}(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{C}}^t$ where $\boldsymbol{G}$ is a generalized inverse of $(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}})$, so that

$$\tilde{\boldsymbol{C}}^t = \tilde{\boldsymbol{C}}^t\boldsymbol{G}(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{C}}^t\boldsymbol{G}\boldsymbol{E}^t\boldsymbol{D}\boldsymbol{E}. \tag{A2}$$

It is straightforward to show that the matrix

$$\boldsymbol{E}^t(\boldsymbol{E}\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}\boldsymbol{E}^t)^{-1}\boldsymbol{E} \tag{A3}$$

is a generalized inverse of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$. Substituting (A3) for $\boldsymbol{G}$ in (A2), it follows that $\tilde{\boldsymbol{C}}^t = \boldsymbol{C}^t\boldsymbol{E}$.

Assume now that there exists a full rank factorization of $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{E}$ such that $\tilde{\boldsymbol{C}}^t = \boldsymbol{C}^t\boldsymbol{E}$, for some matrices $\boldsymbol{X}, \boldsymbol{E}$ and $\boldsymbol{C}$. Then

$$\tilde{\boldsymbol{C}}^t = \boldsymbol{C}^t\boldsymbol{E} = \boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{E} = \boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\tilde{\boldsymbol{X}} = \boldsymbol{T}^t\tilde{\boldsymbol{X}},$$

for $\boldsymbol{T}^t = \boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$. But $\tilde{\boldsymbol{C}}^t = \boldsymbol{T}^t\tilde{\boldsymbol{X}}$ is a sufficient condition for $\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ to be testable (Ravishanker and Dey 2002), and the result follows.

Proof of Proposition 6:
We show that f_i^* is a reparameterization of f_i, for $i = 1, 2$:

For M_1, let $(\boldsymbol{\beta}, \sigma) = g_1(\tilde{\boldsymbol{\beta}}, \sigma) = (\boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma)$, then

$$f_1^*(\boldsymbol{y} \mid g_1(\tilde{\boldsymbol{\beta}}, \sigma)) = f_1^*(\boldsymbol{y} \mid \boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{E}\tilde{\boldsymbol{\beta}} = \boldsymbol{0}\} = f_1(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma),$$

since $\boldsymbol{C}^t\boldsymbol{E} = \tilde{\boldsymbol{C}}^t$ and $\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}$.

For M_2, let $(\boldsymbol{\beta}, \sigma) = g_2(\tilde{\boldsymbol{\beta}}, \sigma) = (\boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma)$, then $f_2^*(\boldsymbol{y} \mid g_2(\tilde{\boldsymbol{\beta}}, \sigma)) = f_2(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma)$, since $\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}$.

Proof of Theorem 1:
We show that $SSE_f = SSE_f^*$ and $SSE_r = SSE_r^*$, using the identities $SSE_f = \boldsymbol{y}^t(\boldsymbol{I}_n - \tilde{\boldsymbol{X}}\boldsymbol{G}\tilde{\boldsymbol{X}}^t)\boldsymbol{y}$ and

$$SSE_r = SSE_f + \boldsymbol{y}^t\tilde{\boldsymbol{X}}\boldsymbol{G}\tilde{\boldsymbol{C}}(\tilde{\boldsymbol{C}}^t\boldsymbol{G}\tilde{\boldsymbol{C}})^{-1}\tilde{\boldsymbol{C}}^t\boldsymbol{G}\tilde{\boldsymbol{X}}^t\boldsymbol{y},$$

where $\boldsymbol{G}$ is any generalized inverse of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$.

Since $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{E}$ is a full rank factorization, $(\boldsymbol{X}\boldsymbol{E})^t\boldsymbol{X} \cdot \boldsymbol{E}$ is a full rank factorization of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$, and hence the matrix

$$\boldsymbol{E}^t(\boldsymbol{E}\boldsymbol{E}^t)^{-1}(\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{E}\boldsymbol{E}^t\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{E},$$

is a generalized inverse of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$. Substitute for $\boldsymbol{G}$ in the identities above and the result follows.

Proof of Proposition 7:
According to Proposition 2, the problem (3) with competing models f_1 and f_2 is equivalent to the problem (4) with competing models f_1^* and f_2^*. It was shown in Example 1 that the CPD's in the parameterization f_i^* are:

$$\pi_1^*(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1} \quad \text{and} \quad \pi_2^*(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1}PIC_k(\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_e \end{pmatrix} \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{I}_k).$$

We show that (11) holds for $i = 1, 2$.

Under M_1, it was shown in proof of Proposition 2, that $f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{S}\boldsymbol{A}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$,

so that

$$\begin{aligned}
&\int f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma)\pi_1(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta} d\sigma = \\
&= \int N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{S}\boldsymbol{A}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)\sigma^{-1} 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0})\, d\boldsymbol{\beta} d\sigma = \\
&\qquad \text{Change: } \boldsymbol{\beta}_1 = \boldsymbol{A}^t\boldsymbol{\beta}, \boldsymbol{\beta}_e = \boldsymbol{C}^t\boldsymbol{\beta}, \text{with jacobian } |det\boldsymbol{R}|^{-1} = 1 \\
&= \int N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{S}\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)\sigma^{-1} 1_{k_e}(\boldsymbol{\beta}_e = \boldsymbol{0})\, d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_e d\sigma = \\
&= \int N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)\sigma^{-1}\, d\boldsymbol{\beta}_1 d\sigma = \int f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma)\pi_1^*(\boldsymbol{\beta}_1, \sigma)\, d\boldsymbol{\beta}_1 d\sigma.
\end{aligned}$$

Under M_2 the result follows by noting that $\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_e \end{pmatrix} = \boldsymbol{R}\boldsymbol{\beta}$ (with $\boldsymbol{R}$ non singular) and applying Result 2 (part 1).

Proof of Proposition 8:
According to Proposition 6, the problem (6) with competing models f_1 and f_2 is equivalent to the problem (7) with competing models f_1^* and f_2^*. The CPD's for this full rank parameterization f_i^* are (see Proposition 7)

$$\pi_1^*(\boldsymbol{\beta}, \sigma) = \sigma^{-1} 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}),$$

and

$$\pi_2^*(\boldsymbol{\beta}, \sigma) = \sigma^{-1} PIC_r(\boldsymbol{\beta} \mid \frac{\boldsymbol{H}}{n\sigma^2}, \boldsymbol{R}),$$

where $\boldsymbol{H} = \boldsymbol{0}_{k_1 \times k_1} \oplus (\boldsymbol{V}^t\boldsymbol{V})$. We next show that (11) holds for $i = 1, 2$.

For M_1, note that

$$\begin{aligned}
&\int f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma)\pi_1^*(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta} d\sigma = \int f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma)\sigma^{-1} 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0})\, d\boldsymbol{\beta} d\sigma = \\
&= \int f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma)\sigma^{-1} 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0})\, h_{k-r}^1(\boldsymbol{\beta}_0)\, d\boldsymbol{\beta} d\boldsymbol{\beta}_0 d\sigma = \\
&\qquad \text{Change: } (\boldsymbol{\beta}^t, \boldsymbol{\beta}_0^t)^t = (\boldsymbol{E}\ ,\ \boldsymbol{Q}_2^t)\tilde{\boldsymbol{\beta}}, \text{ with jacobian } |det\boldsymbol{Q}| = 1 \\
&= \int f_1^*(\boldsymbol{y} \mid \boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma)\sigma^{-1} 1_{k_e}(\boldsymbol{C}^t\boldsymbol{E}\tilde{\boldsymbol{\beta}} = \boldsymbol{0}) h_{k-r}^1(\boldsymbol{Q}_2^t\tilde{\boldsymbol{\beta}})|det\boldsymbol{Q}|\, d\tilde{\boldsymbol{\beta}} d\sigma = \\
&\qquad \text{which, by definition of } \pi_1 \text{ and } g_1, \text{ and } \boldsymbol{C}^t\boldsymbol{E} = \tilde{\boldsymbol{C}}^t \\
&= \int f_1^*(\boldsymbol{y} \mid g_1(\tilde{\boldsymbol{\beta}}, \sigma))\pi_1(\tilde{\boldsymbol{\beta}}, \sigma)\, d\tilde{\boldsymbol{\beta}} d\sigma = \int f_1(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma)\pi_1(\tilde{\boldsymbol{\beta}}, \sigma)\, d\tilde{\boldsymbol{\beta}} d\sigma,
\end{aligned}$$

since f_1^* reparameterizes f_1. A similar proof applies for M_2.

Proof of Proposition 9:

This result is a immediate consequence of the factorization:

$$\pi_2^*(\boldsymbol{\tau},\sigma) = \pi^*(\tau_{j+1},\ldots,\tau_a,\sigma)\prod_{i=1}^{j}\pi_{2.i}^*(\tau_i \mid \tau_{i+1},\ldots,\tau_a,\sigma), \tag{A4}$$

with $\pi_{2.i}^*$ as given in Proposition 9 and

$$\begin{aligned}\pi^*(\tau_{j+1},\ldots,\tau_a,\sigma) = &\frac{\Gamma(\frac{a-j}{2})}{\pi^{(a-j)/2}(a-j)^{1/2}}\big(1+\frac{a-j}{a\sigma^2}S_{j+1}^2\big)^{-(a-j)/2}\\ &\times a^{-(a-j-1)/2}\sigma^{-(a-j)},\end{aligned}$$

for any $1 \le j \le a-1$. Result(A4) can be proven by induction on j, noting that

$$\begin{aligned}\Big(1+\boldsymbol{\tau}^t\boldsymbol{C}\frac{\boldsymbol{V}^t\boldsymbol{V}}{n\sigma^2}\boldsymbol{C}^t\boldsymbol{\tau}\Big)^{-a/2} &= St_1(\tau_1 \mid \bar{\tau}_2,\Sigma_2,a-1)\\ &\times(1+\frac{a-1}{a\sigma^2}S_2^2)^{-a/2}\Big(\frac{\Gamma(\frac{a-1}{2})\,((a-1)\pi)^{1/2}}{\Gamma(\frac{a}{2})}\Sigma_2^{1/2}\Big),\end{aligned}$$

and

$$\begin{aligned}1+\frac{a-i}{a\sigma^2}S_{i+1}^2 &= \big(1+\frac{a-i-1}{a\sigma^2}S_{i+2}^2\big)\\ &\times\Big[1+(\tau_{i+1}-\bar{\tau}_{i+2})^2\frac{a-i-1}{(a-i)(a\sigma^2+(a-i-1)S_{i+2}^2)}\Big],\end{aligned}$$

for $i = 1,2,\ldots,a-1$.

References

[1] Berger, J. O., and Bernardo, J. M. (1992), "On the Development of the Reference prior Method," *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford: University Press, pp. 35-60.

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

[3] Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association,* 91, 109-122.

[4] Berger, J. O., and Pericchi, R. L. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion)," *Model Selection*, ed. P. Lahiri, Institute of Mathematical Statistics Lecture Notes- Monograph Series, volume 38, pp. 135-207.

[5] Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhya: The Indian Journal of Statistics. Series A*, 60, 307-321.

[6] Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B*, 57, 473-484.

[7] Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing,* 12, 27-36.

[8] García-Donato, G. (2003), "Factores Bayes y Factores Bayes convencionales: Algunos aspectos relevantes," unpublished Ph.D. dissertation, University of Valencia, Dept. of Statistics.

[9] Han, C., and Carlin, B. P. (2001), "Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review," *Journal of the American Statistical Association,* 96, 1122-1132.

[10] Jeffreys, H. (1961), *Theory of Probability (3rd ed.),* London: Oxford University Press.

[11] Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association,* 90, 773-795.

[12] Kass, R. E., and Vaidyanathan, S. (1992), "Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions," *Journal of the Royal Statistical Society, Series B*, 54, 1, 129-144.

[13] Kass, R. E., and Wasserman, L. (1995), "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *Journal of the American Statistical Association,*, 90, 928-934.

[14] Ibrahim, J., and Laud, P. (1994), "A predictive approach to the analysis of designed experiments," *Journal of the American Statistical Association*, 89, 309-319.

[15] Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. O. (2005), "Mixtures of g -priors for Bayesian Variable Selection," ISDS working paper.

[16] Moreno, E., Bertolino, F., and Racugno, W. (1998), "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing," *Journal of the American Statistical Association*, 93, 1451-1460.

[17] Moreno, E., Torres, F., and Casella, G. (2005), "Testing equality of regression coefficients in heteroscedastic normal regression models," *Journal of Statistical Planning and Inference*, 131, 117-134

[18] O'Hagan, A. (1994), *Kendall's advanced theory of statistics. Volume 2B: Bayesian Inference,* London: Edward Arnold.

[19] Pérez, J. M. (1998), *Development of Expected Posterior Prior Distributions for Model Comparisons,* Unpublished Ph.D. dissertation, Purdue University.

[20] Pérez, J. M., and Berger, J. O. (2002), "Expected posterior prior distributions for model selection," *Biometrika,* 89, 491-512.

[21] Rencher, A. C. (2000), *Linear Models in Statistics,* John Wiley and Sons, Inc.

[22] Raftery, A. E. (1996), "Hypothesis Testing and Model Selection," *Markov Chain Monte Carlo in Practice,* eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, Chapman and Hall/CRC, pp. 163-187.

[23] Ravishanker, N., and Dey, D. K. (2002), *A first course in linear model theory,* Chapman & Hall/CRC.

[24] Searle, S. R. (1997), *Linear models,* New York : John Wiley & Sons (Wiley Classics library).

[25] Speckman, P. L., and Sun, D. (2003), "Fully Bayesian spline smoothing and intrinsic autoregressive priors," *Biometrika*, 90, 289-302.

[26] Spiegelhalter, D., and Smith, A. (1982), "Bayes factors for linear and Log-linear models with vague prior information," *Journal of the Royal Statistical Society, Series B*, 44, 377-387.

[27] Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999), "Posterior distribution of hierarchical models using CAR(1) distributions," *Biometrika*, 86, 341-350.

[28] Westfall, P., and Gönen, M. (1996), "Asymptotic properties of ANOVA Bayes factors," *Communications in Statistics: Theory and Methods*, 25, 3101-3123.

[29] Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti.* North-Holland/Elsevier (Amsterdam; New York).

[30] Zellner, A., and Siow, A. (1980), "Posterior Odds Ratio for Selected Regression Hypotheses," *Bayesian Statistics 1*, eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Valencia: University Press.

[31] Zellner, A., and Siow, A. (1984), *Basic Issues in Econometrics,* Chicago: University of Chicago Press.