Statistical Science
2008, Vol. 23, No. 3, 332–353
DOI: 10.1214/08-STS258

Compatibility of Prior Specifications Across Linear Models

Guido Consonni and Piero Veronese

Abstract. Bayesian model comparison requires the specification of a prior distribution on the parameter space of each candidate model. In this connection two concerns arise: on the one hand the elicitation task rapidly becomes prohibitive as the number of models increases; on the other hand numerous prior specifications can only exacerbate the well-known sensitivity to prior assignments, thus producing less dependable conclusions. Within the subjective framework, both difficulties can be counteracted by linking priors across models in order to achieve simplification and compatibility; we discuss links with related objective approaches. Given an encompassing, or full, model together with a prior on its parameter space, we review and summarize a few procedures for deriving priors under a submodel, namely marginalization, conditioning, and Kullback–Leibler projection. These techniques are illustrated and discussed with reference to variable selection in linear models adopting a conventional g-prior; comparisons with existing standard approaches are provided. Finally, the relative merits of each procedure are evaluated through simulated and real data sets.

Key words and phrases: Bayes factor, compatible prior, conjugate prior, g-prior, hypothesis testing, Kullback–Leibler projection, nested model, variable selection.

1. INTRODUCTION

Model comparison is an important and active area of research especially from the Bayesian viewpoint; see, for example, George (1999) and Robert (2001, Chapter 7). In particular, the problem of variable selection in linear models has received considerable attention; see the review paper of George (2000) and a few survey chapters in the book edited by Dey and Rao (2005). Two critical issues emerge from the very beginning: the elicitation of prior probabilities for the various models under consideration and the assignment of prior distributions on the parameter space of each model, which we simply call *priors*. In this paper we focus on the latter.

Guido Consonni is Professor, Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia, Via S. Felice 7, 27100 Pavia, Italy (e-mail: guido.consonni@unipv.it). Piero Veronese is Professor, Department of Decision Sciences, L. Bocconi University, Via Roentgen, 20136 Milano, Italy (e-mail: piero.veronese@unibocconi.it).

Occasionally, when the model space is not large and detailed prior information is available, subjective prior elicitation on each model can be carried out; see Garthwaite and Dickey (1996). More often, however, because of the potentially very high number of models under investigation, prior elicitation can represent a formidable task, and hence practically implementable procedures have been actively looked for. In the objective framework (see Berger and Pericchi, 1996b), a convenient approach is to start with a default, typically improper, prior under each model, and then to circumvent the indeterminacy of the normalizing constant through an intrinsic prior procedure (see also Casella and Moreno, 2006, for an application to variable selection in linear models). A more general approach, namely expected posterior prior, is described in Pérez and Berger (2002).

Outside the purely objective view, pragmatic simplification of the elicitation task in the variable selection problem has been achieved through hierarchical mixture priors as in George and McCulloch (1997), or using an empirical Bayes approach, as in George

and Forster (2000), and more recently in Yuan and Lin (2005), or employing a blend of noninformative and conjugate procedures, as exemplified in Fernández, Ley and Steel (2001). Recently Liang et al. (2008) have proposed mixtures of g-priors as an efficient tool for Bayesian variable selection.

Within the subjective framework, which uses proper priors, the idea of relating priors across models does not seem to be pervasive. Notable exceptions are Dickey (1971) and Poirier (1985), in the context of linear models; see also the discussion in O'Hagan and Forster (2004, Sections 11.29–11.31). Neal (2001) introduces the idea of transferring prior information from a "donor model" to a "recipient model." His motivation is primarily pragmatic: priors for complex models are harder to elicit than those for simple models; accordingly one can try to carefully elicit a prior under a simple "donor" model and then transfer this information to a complex "recipient" model. Technically Neal's method is similar to, although more general than, the expected posterior prior of Perez and Berger (2002). The paper by Dawid and Lauritzen (2001) stands out as an attempt to discuss, in a general setting, methods to construct "compatible priors" for nested models using a variety of strategies. Their motivation is mixed: on the one hand they state that conceptually there is no compelling reason to relate priors across models (since they express subjective opinions conditionally on a different state of information); on the other hand such relationships may be highly desirable on pragmatic grounds (the effort spent in eliciting a prior under a model should somehow be transferred to other models) and also to achieve some sort of compatibility in order to lessen the sensitivity of the Bayes factor to prior specifications.

Following up this comment, we believe that priors for model comparison deserve to be carefully investigated by the Bayesian community. Traditional priors, which individually perform quite effectively within a single model, need not work satisfactorily when collectively employed for comparing models of varying dimensions. This fact has been informally recognized at least since Jeffreys, who refrained from using conventional priors for comparing two nested hypotheses; see also Zellner and Siow (1980) in the framework of linear models.

In the context of comparing a sharp null hypothesis H_0 versus a composite alternative H, Morris (1987) argued forcibly for the prior under H to be "centered around H_0"; otherwise the prior under H would be "wasting away" prior probability mass in regions that are often too unlikely to be supported by the data, thus unduly favoring H_0, as lucidly spelled out in Casella and Moreno (2007); see also Consonni and La Rocca (2008). Carefully extending this argument to several models would surely be of great value and interest in order to enhance our understanding of the issue of compatibility of priors for model comparison. While this paper falls short of providing a comprehensive treatment of this point, it nevertheless tries to offer some guidance for further reflection and research. Specifically, we try to elucidate the meaning of the term "submodel," or nested model, in order to highlight differences between a couple of approaches which are implicit in the literature and better understand specific strategies to relate priors across models. Although the scope of our considerations is general, we will illustrate the main ideas with reference to the problem of variable selection in linear models.

The structure of the paper is as follows. Section 2 deals with two notions of nested models and discusses the corresponding parametrization, distinguishing between nuisance and common parameters. Section 3 deals with strategies to assign priors on parameters of submodels starting from a prior on the (full) model; we discuss conditioning and projection (including marginalization) and propose, in Sections 3.1 and 3.2, two criteria to evaluate such strategies, which we name nuisance- and nested-coherence. Section 4 deals with priors for linear models. Starting with a g-prior under the full model, a variety of prior specifications on submodels is obtained through the procedures described in Section 3; in particular Section 4.2 contains a discussion of the so-called "information paradox." Section 5 presents three examples to evaluate the performance of the various priors under consideration in terms of model comparison, with special references to sensitivity issues. Finally, Section 6 provides a few points for discussion. To ease the flow of ideas, technical aspects have been relegated to the Appendix.

2. SUBMODELS

2.1 A Preliminary Example

We start by discussing a very simple example with the aim of presenting the main issues at stake. Consider the following model:

$$\mathcal{M}: \; y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$
$$(\alpha, \beta, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$$

where, conditionally on σ^2, $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. An obvious submodel, say $\mathcal{M}^*$, removes the predictor, thus changing the mean structure. However, several instances of $\mathcal{M}^*$ are available, namely:

$$\mathcal{M}_A^*: \; y_i = \alpha + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2);$$

$$\mathcal{M}_B^*: \; y_i = \alpha + \varepsilon_i^*, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^{*2});$$

$$\mathcal{M}_C^*: \; y_i = \alpha^* + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2);$$

$$\mathcal{M}_D^*: \; y_i = \alpha^* + \varepsilon_i^*, \quad \varepsilon_i^* \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^{*2}).$$

Model $\mathcal{M}_A^*$ originates in the setting of hypothesis testing postulating that $\beta = 0$ under $\mathcal{M}$; in other words, $\mathcal{M}_A^*$ is equivalent to the hypothesis $H^*: y \sim \mathcal{M}$ and $\beta = 0$. As a consequence the parameters α and σ^2 are "common" to both models, although one might further distinguish between them, since σ^2 pertains to the error structure (which is not affected explicitly by the submodel specification), and thus can be regarded as a "nuisance" parameter. Model $\mathcal{M}_B^*$ originates from the consideration that the error component in the submodel might, and perhaps should, be allowed to be different from that under $\mathcal{M}$. In particular, since one can anticipate a worse fit under $\mathcal{M}_B^*$ than under $\mathcal{M}$, one should have $E(\sigma^{*2}) \geq E(\sigma^2)$ or even $\sigma^{*2} \geq \sigma^2$ (with probability 1). Model $\mathcal{M}_C^*$ originates from the consideration that the meaning of the intercept is actually quite different under the two models, and so should be distinct from that under $\mathcal{M}$. On the other hand σ^2 remains the same, since it is regarded as a "nuisance" parameter. Finally model $\mathcal{M}_D^*$ combines the specific features of $\mathcal{M}_B^*$ and $\mathcal{M}_C^*$, and has no direct link, unlike the previous versions, to $\mathcal{M}$. For a related discussion on alternative interpretations of submodels, see Berger and Pericchi (2001, Section 1.5, "Difficulty 4").

In an abstract sense, all instances of $\mathcal{M}^*$ above represent the same submodel, since they share the same family of distributions. However, the distinctive features that we have tried to underline should make it clear that they are different objects, or perhaps different ways of looking at the same object. For a given prior π on $(\alpha, \beta, \sigma^2)$ under $\mathcal{M}$, we require a prior, π^* say, under $\mathcal{M}^*$. We claim that each instance of $\mathcal{M}^*$ naturally suggests a different procedure to obtain π^* from π.

Consider first model $\mathcal{M}_A^*$. There are two natural candidates for π^*, namely $\pi(\alpha, \sigma^2)$ and $\pi(\alpha, \sigma^2 | \beta = 0)$, that is, the marginal and the conditional (on $\beta = 0$) distribution derived from $\pi(\alpha, \beta, \sigma^2)$. The latter might appear more natural, if the hypothesis-testing interpretation of $\mathcal{M}_A^*$ is strictly adhered to. Note that the two procedures lead to the same priors if (α, σ^2) is independent of β, as it occurs using default priors. For model $\mathcal{M}_B^*$, instead, no obvious indications are provided for the specification of $\pi^*(\sigma^{*2})$; on the other hand, since α is "common" to both models, a natural suggestion would be to take $\pi^*(\alpha) = \pi(\alpha)$. Of course the problem of combining the two marginal distributions into a joint one remains open. Under model $\mathcal{M}_C^*$ a situation somewhat similar to that under $\mathcal{M}_B^*$ obtains, if we interchange the role of the intercept and the variance. Finally, neither marginalization nor conditioning appears as obvious recommendations under $\mathcal{M}_D^*$, because no effective link with $\mathcal{M}$ is specified. The next sections explore these issues in greater generality.

2.2 Nested Models

It could be argued that each of the models $\mathcal{M}^*$ described in Section 2 is *nested* in $\mathcal{M}$. However, we feel some other clarification is needed.

Consider a model $\mathcal{M} = \{f(\cdot|\theta), \theta \in \Theta\}$. There seem to be two interpretations of a nested model $\mathcal{M}^*$ in the literature, often not clearly distinguished. Both start from the assumption that it is possible to write $\theta = (\lambda, \phi)$, where $\lambda \in \Lambda$ and $\phi \in \Phi$, with λ and ϕ being variation-independent, so that $\Theta = \Lambda \times \Phi$ and model $\mathcal{M}^*$ is identified through the constraint $\phi = \phi_0$, with ϕ_0 a fixed value. As suggested by a referee, this setting covers only the case in which the parameter space Θ^* associated with $\mathcal{M}^*$ has dimension strictly smaller than that of $\mathcal{M}$, and thus it does not account for other interesting nesting situations in which $\dim(\Theta^*) = \dim(\Theta)$ (e.g., when Θ^* is a restriction of Θ). However, the above (λ, ϕ)-representation is especially useful from the perspective of "prior assignment" under submodels, which is the primary focus of this paper. We describe these two interpretations below.

S-N (*Strongly nested interpretation*): The sampling distribution of y under $\mathcal{M}^*$ is given by $f^*(\cdot|\lambda), \lambda \in \Lambda$, where $f^*(\cdot|\lambda) = f(\cdot|\lambda, \phi = \phi_0)$. This interpretation can be clarified in terms of the underlying generating process of y: "If Nature chooses $\lambda \in \Lambda$ and $\phi = \phi_0$, then the distribution of the observables under $\mathcal{M}$ and $\mathcal{M}^*$ is the same."

W-N (*Weakly nested interpretation*): The sampling distribution of the observations y under $\mathcal{M}^*$ can be written as $f^*(\cdot|\gamma), \quad \gamma \in \Lambda$, with $f^*(\cdot|\gamma) = f(\cdot|\lambda = \gamma, \phi = \phi_0)$. In this way γ, although structurally equivalent to λ, is distinct from it. Clearly, each distribution

in $\mathcal{M}^*$ also belongs to $\mathcal{M}$.

Interpretation S-N is rooted in a hypothesis-testing context, that is, $H^*: \phi = \phi_0$, where the actual objective of the analysis is verifying whether $\phi = \phi_0$, *other things being held equal*. On the other hand, W-N is better suited when the objective is model simplification, and each model competes against the other ones according to whatever criterion is deemed to be appropriate (e.g., a combination of fit and parsimony, or on predictive grounds; see, e.g., Gelfand and Ghosh, 1998 and Marriott, Spencer and Pettitt, 2001). With regard to the example in Section 2.1, $\mathcal{M}_A^*$ is the only instance of $\mathcal{M}^*$ that falls under interpretation S-N. The S-N view is probably the most pervasive and is regarded as a natural framework by, for example, Poirier (1985), O'Hagan and Forster (2004, Section 7.15) and Davison (2003, page 127). It seems implicit in George and Forster (2000) and other workers mostly interested in computational aspects, for example, Smith and Kohn (1996), Nott and Green (2004) and Cripps, Carter and Kohn (2005). On the other hand, authors like Berger and Pericchi (1996a) and also Robert (2001, Section 7.2) seem to prefer interpretation W-N.

Within the interpretation S-N, consider a collection of submodels $\mathcal{M}_k$ and suppose that, for each $\mathcal{M}_k$, there exists a reparametrization of $\mathcal{M}$ as $(\delta, \eta_k, \omega_k)$, so that $\mathcal{M}_k$ is identified by $\eta_k = \eta_{k0}$. Since δ is never involved in any submodel specification we can regard it as a *nuisance* parameter; on the other hand we call ω_k the parameter *common* to the pair $(\mathcal{M}, \mathcal{M}_k)$. In the setting of variable selection for linear models, the nuisance parameter is clearly represented by the error variance σ^2, while common parameters are the regression coefficients that are not set to zero in the submodel specification.

We close this section with a caveat that hopefully will not disconcert the reader. Despite our insistence on model interpretation and parametric description, we emphasize that what matters in a Bayesian analysis is the *prior distribution* attached to the parameters of the various models regardless of their formal representation. The latter, however, may become relevant when structuring prior specification across models. This is the topic of the next section.

3. STRATEGIES TO ASSIGN PRIORS ON PARAMETERS OF SUBMODELS

Within the objective Bayesian framework, the expected posterior prior (EPP) methodology of Pérez and Berger (2002) is a method to construct prior distributions for model comparison; see also Neal (2001) for related concepts. The idea is to start with a prior distribution under each model, compute its posterior under "imaginary" observations, and formally average the posterior through a marginal data distribution that is common to all models. The method is quite general, but is especially effective if one starts with a default, possibly improper, prior under each model. In this way the EPP method allows to use improper priors for model comparison through Bayes factors, or posterior model probabilities, since the indeterminate normalizing constants cancel out. More generally, EPP is a method to make priors "compatible" across models, through their dependence on a common marginal data distribution; thus this methodology can be applied also with subjectively specified (proper) prior distributions.

Although appealing and flexible, implementing the EPP methodology may be problematic. First of all the choice of the common distribution is not unique. For instance, there exist at least two competing choices, namely that corresponding to the "simplest" model, if it exists, and that corresponding to the empirical distribution, which requires the identification of a minimal training sample; see Berger and Pericchi (2004) for a discussion of potential difficulties associated to this concept. More importantly, to judge the relative merits of the above two choices is not straightforward. A second concern refers to the actual implementation of the EPP, which may require careful computational strategies.

A more specific approach is the intrinsic prior methodology, which has received a great deal of attention both for hypothesis testing and for model selection. Again the primary motivation is the use of default noninformative priors under each model; see Pericchi (2005) for a review. When several models are entertained the intrinsic method requires a nesting strategy. One approach, labeled "encompassing from above," chooses as benchmark a full model wherein all other models are nested. In this way, however, the prior under the full model changes in each pairwise comparison, thus producing an overall incoherent probabilistic answer. Yet posterior probabilities can still be formally defined on the basis of the collection of Bayes factors of each model relative to the full one; see Casella and Moreno (2006) for an application to variable selection in linear models. On the other hand, if the simplest model (i.e., one being nested within any other model) is available, an alternative "encompassing from below" intrinsic prior procedure can be followed, which is

probabilistically correct; for an application to variable selection see Moreno and Giron (2007). Notice that the two alternative encompassing procedures will typically lead to distinct answers. As with the EPP methodology, analytic evaluation of intrinsic priors is typically very hard and actual implementation of the procedure requires a good deal of computational ingenuity; see Casella and Moreno (2005) in the context of contingency tables.

Although the EPP and intrinsic prior methodologies produce priors that are "related" through a common underlying marginal data distribution, they do not explicitly address the issue of prior compatibility across models. The latter issue is lucidly tackled in Dawid and Lauritzen (2001), who present several strategies for the derivation of compatible priors; see also Roverato and Consonni (2004) in the context of directed graphical models and Consonni, Gutiérrez-Peña and Veronese (2007) for general exponential families with a detailed application to testing the Hardy–Weinberg model in studies of population genetics.

Starting with a model $\mathcal{M} = \{f(y|\lambda, \phi)\}$ and a joint distribution $\pi(\lambda, \phi)$, we briefly review below four main strategies for prior specification under a nested model $\mathcal{M}^*$ identified through $\phi = \phi_0$.

Marginalization (M). This approach is most natural under interpretation S-N where $\mathcal{M}^* = \{f^*(y|\lambda), \lambda \in \Lambda\}$, so that $\mathcal{M}$ and $\mathcal{M}^*$ share the same parameter λ, and states that $\pi^{\mathrm{M}}(\lambda) = \pi(\lambda)$, where $\pi(\lambda)$ is the marginal of λ under $\pi(\lambda, \phi)$. Two critical aspects should be taken into consideration: (i) marginalization does not explicitly take into consideration the constraint $\phi = \phi_0$; in fact it disregards this information by averaging with respect to the distribution of ϕ; (ii) on a more technical side, this procedure is not invariant to reparametrization. Consider, for instance, model $\mathcal{M}$ of Section 2.1, and suppose to recenter the data as $x_i \to x_i - \bar{x}$, with $\bar{x}$ the mean of the x_i. The model $\mathcal{M}$ becomes $(\alpha - \beta\bar{x}) + \beta x_i$ suggesting the following reparametrization: $(\alpha, \beta) \mapsto (\gamma, \delta)$, where $\gamma = \alpha - \beta\bar{x}$, and $\delta = \beta$. Notice that α and γ are the same quantities under $\mathcal{M}^*$ and so should share the same prior under the latter model. On the other hand, α and γ are distinct under $\mathcal{M}$ and will have typically different priors, a feature which will be inherited under $\mathcal{M}^*$ through the procedure M, thus establishing its lack of invariance.

Usual conditioning (UC). As with M, this procedure applies more naturally under interpretation S-N, and states that $\pi^{\mathrm{UC}}(\lambda) = \pi(\lambda|\phi = \phi_0)$, where the right-hand side is the conditional distribution of λ given $\phi = \phi_0$ under $\pi(\lambda, \phi)$. A clear advantage of UC is that it incorporates explicitly the information available in the specification of model $\mathcal{M}^*$, through the constraint $\phi = \phi_0$. The major drawback of UC is that it is not invariant to the choice of the conditioning function (typically an event having zero probability) which identifies the submodel. For instance, assume that $\mathcal{M}$ is as in Section 2.1, and that (α, β) are jointly normal with zero mean, variances σ_α^2, σ_β^2 and correlation coefficient ρ. Then the distribution of α given $\beta = 0$ is normal with zero mean and variance $\sigma_\alpha^2(1 - \rho^2)$. On the other hand, model $\mathcal{M}^*$ could also be identified through the constraint $\xi = 0$, where $\xi = \beta/\alpha$. It can be checked that the conditional distribution of α given $\xi = 0$ is no longer normal. This represents an instance of the Borel–Kolmogoroff paradox.

Jeffreys conditioning (JC). This procedure is a variation of UC and hence is most appropriate again under interpretation S-N. It was proposed by Dawid and Lauritzen (2001) to overcome the lack of invariance of UC. First recall that the density obtained through UC can be expressed as $\pi^{\mathrm{UC}}(\theta) \propto \pi(\theta), \quad \theta \in \tilde{\Theta}^*$, where $\tilde{\Theta}^* = \{(\lambda, \phi), \lambda \in \Lambda, \phi = \phi_0\}$. Now let $H(\theta)$ denote the Fisher information matrix for θ under $\mathcal{M}$, and similarly for $H^*(\theta)$ under $\mathcal{M}^*$. Set $j(\theta) \propto |H(\theta)|^{1/2}$, where $|H|$ is the determinant of H, so that $j(\theta)$ is the Jeffreys prior for θ under $\mathcal{M}$, and define analogously $j^*(\theta)$ under model $\mathcal{M}^*$. The JC density is defined as

$$\pi^{\mathrm{JC}}(\theta) \propto \pi(\theta)\frac{j^*(\theta)}{j(\theta)}, \quad \theta \in \tilde{\Theta}^*. \tag{1}$$

Typically, one would re-express the JC density as a function of λ only, and write $\pi^{\mathrm{JC}}(\lambda)$ accordingly; we shall follow this style in the next section. A useful feature of Jeffreys conditioning is invariance to model reparametrization, because of the multiplicative term given by the ratio of the Jeffreys densities. A potential difficulty with Jeffreys conditioning is that the resulting prior $\pi^{\mathrm{JC}}(\lambda)$ may be improper even though $\pi(\theta)$ is proper, because of its nonprobabilistic nature.

Kullback–Leibler (KL) projection. This procedure is part of a more general approach to the construction of priors on related models based on *projection maps*, and is especially appropriate under interpretation W-N. Consider a model $\mathcal{M}$ and a submodel $\mathcal{M}^*$, parametrized by $\theta^* \in \Theta^*$ for the same observable, and suppose that each distribution in $\mathcal{M}$ has an image in $\mathcal{M}^*$ through the (projection) map $\tau : \Theta \mapsto \Theta^*$. Given a prior $\pi(\theta)$ on Θ, the prior induced on $\tau(\theta)$ is called the τ-projection prior.

For reasons to be specified shortly below, we shall take $\tau(\theta)$ as the Kullback–Leibler (KL)-projection of

θ onto Θ^*, that is,

$$\tau_\theta^{\mathrm{KL}}(\theta) = \arg\min_{\theta^* \in \Theta^*} KL(f(\cdot|\theta),\, f^*(\cdot|\theta^*)),$$

where

$$KL(p, q) = E^p\left(\log \frac{p(X)}{q(X)}\right)$$

denotes the KL-divergence between the density p and q relative to a common dominating measure. In this case we call the resulting prior KL-projection prior, or KL-prior for short, and denote it with $\pi^{\mathrm{KL}}(\theta^*)$, that is, $\pi^{\mathrm{KL}}(\theta^*) = \pi^{\theta}_{\theta^\perp}(\theta^*)$, where $\pi^{\theta}_{\theta^\perp}$ is the prior on $\theta^\perp = \tau_\theta^{\mathrm{KL}}(\theta)$ induced from the prior $\pi(\theta)$. KL-priors were originally presented in McCulloch and Rossi (1992) to compute Bayes factors; they are applied in Viele and Srinivasan (2000) to ANOVA models, and in Consonni, Gutiérrez-Peña and Veronese (2007) to a particular multinomial model. Goutis and Robert (1998) and Dupuis and Robert (2003) use KL-projection for comparing models, but do not rely on the idea of KL-priors.

Notice that $KL(p, q)$ is not symmetric. The intrinsic discrepancy between p and q, $\delta(p, q) = \min\{KL(p, q), KL(q, p)\}$ (see Bernardo and Rueda, 2002), overcomes this difficulty. However, we will still use $KL(p, q)$ because (i) we take p as the encompassing model, whose validity is not questioned within our approach, while q is a simplified version of p; from this point of view taking expectations with respect to p, as in $KL(p, q)$, appears a sensible procedure; (ii) for regular nested models (wherein the support is independent of the parameter), p and q have the same support so that $KL(p, q)$ is well defined; (iii) the use of $\delta(p, q)$, instead of $KL(p, q)$, adds complexity from an analytical viewpoint (for a detailed discussion on these points see Consonni, Gutiérrez-Peña and Veronese, 2007).

From our perspective, a very important feature of the KL-projection is its invariance to reparametrization. Thus if $\eta = g(\theta)$ is a reparametrization under $\mathcal{M}$, then $\tau_\eta^{\mathrm{KL}}(\eta) = \tau_\theta^{\mathrm{KL}}(g^{-1}(\eta))$. Accordingly, prior assignments based on KL-projection do not depend on the specific parametrization that is chosen. To illustrate the KL-procedure, consider the simple linear model $\mathcal{M}$ of Section 2.1 with the submodel specified by $\mathcal{M}_D^*$. It can be checked that the KL-projection of $(\alpha, \beta, \sigma^2)$ onto the space $\{(\alpha^*, \sigma^{*2}) \in \mathbb{R} \times \mathbb{R}^+\}$ is given by

$$\begin{aligned}(\alpha, \beta, \sigma^2)^\perp &= \left(\alpha + \beta\bar{x}, \sigma^2 + \beta^2 \frac{1}{n}\sum (x_i - \bar{x})^2\right) \\ &= (\alpha^\perp, \sigma^{2\perp}),\end{aligned}$$

with some abuse of notation for the latter equality. It is interesting to remark that the projection corresponding to the variance is given by σ^2 plus a quadratic term: as a consequence σ^{*2} is stochastically larger, under the KL-prior, than σ^2, whatever the prior on σ^2 under $\mathcal{M}$. This seems to be consistent with the views of those authors who state that σ^{*2} should perhaps be larger than σ^2, to account for an anticipated worse fit of the submodel; see Berger and Pericchi (2001, Section 1.5) and Robert (2001, page 349). A similar, although less stringent, view is held by George and McCulloch (1997) according to whom the *expectation* of σ^2 under the smaller model should be larger. The exact form of the joint KL-prior for (α^*, σ^{*2}) is typically unavailable because of the complicated structure of $\sigma^{2\perp}$; however, we will provide an analytical approximation in the next section. Alternatively, one could resort to stochastic simulation since a draw from $\pi^{\mathrm{KL}}(\cdot)$ can be easily obtained by first generating $\tilde{\theta}$ from $\pi(\cdot)$ and then calculating $\tau_\theta^{\mathrm{KL}}(\tilde{\theta})$, possibly through numerical methods.

3.1 Coherence of Procedures With Respect to Nuisance Parameters

In this section we plan to evaluate the procedures to construct priors under submodels from the point of view of coherence with respect to the nuisance parameter as defined in Section 2.2.

If δ is a nuisance parameter, then it could be integrated out from the very beginning (see O'Hagan and Forster, 2004, Sections 3.13–3.14), using a prior under $\mathcal{M}$. A new *integrated* model $\mathbb{1}\mathcal{M}$ would then be obtained, which in turn generates an integrated submodel $\mathbb{1}\mathcal{M}^*$. Let y be a future observation to be forecast. We say that a procedure is *nuisance-coherent* if the marginal distributions of y under submodel $\mathcal{M}^*$ and the corresponding integrated submodel $\mathbb{1}\mathcal{M}^*$ are the same, that is,

$$f^*_{\mathcal{M}^*}(y) = f^*_{\mathbb{1}\mathcal{M}^*}(y). \tag{2}$$

In other words, integrating out the nuisance parameter "at the beginning" (using π) or "at the end" (using the procedure-induced prior) does not make any difference. If (2) holds, then the predictive distributions under the two models are equivalent; moreover, the Bayes factor for the pair $(\mathcal{M}, \mathcal{M}^*)$ coincides with that for $(\mathbb{1}\mathcal{M}, \mathbb{1}\mathcal{M}^*)$, since $f_{\mathcal{M}}(y) = f_{\mathbb{1}\mathcal{M}}(y)$ by definition of integrated model.

The following proposition establishes results on nuisance-coherence for the procedures M, UC and JC.

PROPOSITION 1. *Consider a model $\mathcal{M}$ parametrized by (λ, δ, ϕ) with δ a nuisance parameter, and prior $\pi(\lambda, \delta, \phi)$. Let $\mathcal{M}^*$ be a submodel identified through $\phi = \phi_0$. Then*:

(i) *the UC procedure is nuisance-coherent*;

(ii) *the M procedure is nuisance-coherent if δ is conditionally independent of ϕ given λ under $\pi(\lambda, \delta, \phi)$*;

(iii) *the JC procedure is nuisance-coherent if the ratio of the Jeffreys priors relative to the pair $(\mathcal{M}, \mathcal{M}^*)$ is proportional to that for the pair $(\mathcal{I}\mathcal{M}, \mathcal{I}\mathcal{M}^*)$, provided the resulting priors are proper.*

PROOF. See the Appendix. □

In general nuisance-coherence does not hold for the KL-procedure; see Section 4.1.3.

3.2 Coherence of Procedures Across Nested Models

We now address the issue of coherence across a collection of submodels. It is actually enough to consider only three models. For simplicity of exposition we shall formulate the problem within interpretation S-N (see Section 2.2). Specifically, consider the following models:

$$\mathcal{M}: \quad f(y|\lambda, \phi_1, \phi_2), \tag{3}$$

$$\mathcal{M}^*: \quad f^*(y|\lambda, \phi_2) = f(y|\lambda, \phi_1 = \phi_1^0, \phi_2), \tag{4}$$

$$\begin{aligned} \mathcal{M}^{**}: \ f^{**}(y|\lambda) &= f(y|\lambda, \phi_1 = \phi_1^0, \phi_2 = \phi_2^0) \\ &= f^*(y|\lambda, \phi_2 = \phi_2^0), \end{aligned} \tag{5}$$

so that $\mathcal{M}^*$ is a submodel of $\mathcal{M}$ and $\mathcal{M}^{**}$ is a submodel of $\mathcal{M}^*$ (and so also of $\mathcal{M}$). Let $\pi(\lambda, \phi_1, \phi_2)$ be the prior under $\mathcal{M}$, $\pi^*(\lambda, \phi_2)$ that under $\mathcal{M}^*$ and finally $\pi^{**}(\lambda)$ that under $\mathcal{M}^{**}$. For each given procedure to construct priors on submodels, the prior $\pi^{**}(\lambda)$ can be obtained either with respect to the pair $(\mathcal{M}, \mathcal{M}^{**})$, which we label $\pi^{**}_{\mathcal{M}}(\lambda)$, or with respect to the pair $(\mathcal{M}^*, \mathcal{M}^{**})$, which we label $\pi^{**}_{\mathcal{M}^*}(\lambda)$.

We say that a procedure is *nested-coherent* if $\pi^{**}_{\mathcal{M}}(\lambda) = \pi^{**}_{\mathcal{M}^*}(\lambda)$.

PROPOSITION 2. *Consider the three models described in* (3)–(5). *The M, UC and JC procedures are nested-coherent.*

PROOF. See the Appendix. □

We remark that nested-coherence fails in general for the KL-procedure as we report in Section 4.1.3 with reference to linear models.

4. LINEAR MODELS

Consider the general linear model $\mathcal{M}$

$$y = X\beta + \varepsilon, \tag{6}$$

where y is an n-dimensional vector of observations on the dependent variable, X an $(n \times p)$ matrix of predictors having rank p, β a p-dimensional vector of regression coefficients and ε an n-dimensional vector of error terms with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$, conditionally on σ^2. We assume that the constant term is always included in the model, so that the first column of X is the unit vector. It is useful to think of (6) as the *full* model.

If subjective information is limited, we can easily resort to conventional proper priors such as the conjugate normal inverted gamma (NIGa) family; see, for example, O'Hagan and Forster (2004, Section 11.4). Specifically, under a $\mathrm{NIGa}(b, V, d, a)$ prior, the conditional distribution of β given σ^2 is $\mathrm{N}(b, \sigma^2 V)$ while the marginal distribution of σ^2 is $\mathrm{IGa}(d/2, a/2)$. Here, $\mathrm{N}(b, \Sigma)$ denotes a normal distribution with expectation b and variance matrix Σ, while $\mathrm{IGa}(d/2, a/2)$ stands for an inverted gamma distribution having expectation $a/(d-2), d > 2$. In many applications, and especially in econometric analysis, a simplified version of the NIGa prior is usually considered. The suggestion of Zellner (1986), called *g-prior*, is to set $V = g(X^T X)^{-1}$, with $g > 0$. The choice of g has been extensively analyzed in several papers, for example, George and Foster (2000), Clyde and George (2004) and Fernández, Ley and Steel (2001).

Some authors have raised criticism against the use of g-priors for model selection (see for a clear exposition Berger and Pericchi, 2001), and have suggested alternative conventional priors, such as the Cauchy prior by Zellner and Siow (1980), recently discussed in Bayarri and Garcia-Donato (2007). Liang et al. (2008) propose to use a prior on the parameter g leading to a mixture of g-priors, which includes as a special case that by Zellner and Siow. This prior does not suffer from the "information paradox" which represents a major drawback of g-priors; see Section 4.2. However, we still employ a g-prior on the full model because of its simplicity and analytical tractability. At any rate the compatible priors that we derive under the various submodels differ from the g-priors traditionally employed.

We take as prior for (β, σ^2) under $\mathcal{M}$

$$\pi(\beta, \sigma^2) = \mathrm{NIGa}(\beta, \sigma^2; b, g(X^T X)^{-1}, d, a), \tag{7}$$

hierarchically specified through

$$\begin{aligned} \pi(\beta|\sigma^2) &= \mathrm{N}(\beta; b, g\sigma^2(X^T X)^{-1}); \\ \pi(\sigma^2) &= \mathrm{IGa}(\sigma^2; d/2, a/2), \end{aligned} \tag{8}$$

and refer informally to (7) as the gNIGa prior.

Concerning the choice of $E(\beta) = b$, three default options are

$$b_0^T = (0, \ldots, 0), \quad \bar{b}^T = (\bar{y}, 0, \ldots, 0), \quad \hat{b} = \hat{\beta}, \tag{9}$$

where $\hat{\beta}$ represents the OLS estimate of β under the full model. In this way the elicitation of the gNIGa prior reduces simply to choosing the three hyperparameters d, a and g. Possible choices for g are extensively discussed in Fernández, Ley and Steel (2001). In particular, based on simulation results, they recommend using $g = \max\{n, p^2\}$, so that typically $g = n$, because n ordinarily exceeds p^2.

4.1 Priors for Submodels

We now review some techniques for prior specification under a generic linear submodel. Let $\mathcal{M}_k$ represent a submodel that uses p_k predictors with $p_k < p$. Write $X = (X_k \vdots X_{\setminus k})$, where X_k is an $(n \times p_k)$ matrix. We assume that each submodel includes the intercept term, so that the first column of X_k is the unit vector; for this reason there exist 2^{p-1} possible models. Let $\beta^T = (\beta_k^T, \beta_{\setminus k}^T)$ be the partition corresponding to that of X.

If we adopt interpretation S-N of nested models, we can write $\mathcal{M}_k$ as $y = X_k\beta_k + \varepsilon$, which is equivalent to the hypothesis $H_k : \beta_{\setminus k} = 0$. On the other hand if one follows interpretation W-N, $\mathcal{M}_k$ can be expressed as

$$y = X_k\beta_k^* + \varepsilon_k, \tag{10}$$

with $\varepsilon_k \sim \mathrm{N}(0, \sigma_k^2 I)$, and β_k^* a p_k-dimensional vector. Notice that in this setting each submodel presents a specific parametric representation, with a distinct β_k^* and σ_k^2. To simplify the exposition, in the following we will make use exclusively of representation (10) which reduces to the S-N case by setting $\beta_k^* = \beta_k$ and $\sigma_k^2 = \sigma^2$.

It is common practice to "replicate" the gNIGa prior described in (7), under each $\mathcal{M}_k$, in particular using the same values of g, d and a. We will show that the UC and JC procedures, as well as KL based on a conjugate approximation, lead instead to

$$\begin{aligned}&\pi_k(\beta_k^*, \sigma_k^2)\\ &\quad = \mathrm{NIGa}(\beta_k^*, \sigma_k^2; b_k^*, g_k(X_k^T X_k)^{-1}, d_k, a_k),\end{aligned} \tag{11}$$

with model-specific hyperparameters. As a consequence, the marginal distribution of y is an n-dimensional Student t-distribution and the Bayes factor for model $\mathcal{M}_k$ versus model $\mathcal{M}_s$ can be written as

$$\begin{aligned}B_{ks} = C_{ks}\Bigg\{&a_s + \frac{g_s}{1+g_s} y^T M_s y\\ &+ \frac{1}{1+g_s}(y - X_s b_s^*)^T\\ &\cdot (y - X_s b_s^*)\Bigg\}^{(d_s+n)/2}\\ \cdot \Bigg[\Bigg\{&a_k + \frac{g_k}{1+g_k} y^T M_k y\\ &+ \frac{1}{1+g_k}(y - X_k b_k^*)^T\\ &\cdot (y - X_k b_k^*)\Bigg\}^{(d_k+n)/2}\Bigg]^{-1};\end{aligned} \tag{12}$$

where

$$C_{ks} = \frac{\Gamma(d_s/2)(a_k)^{d_k/2}\Gamma((d_k+n)/2)(1+g_s)^{p_s/2}}{\Gamma(d_k/2)(a_s)^{d_s/2}\Gamma((d_s+n)/2)(1+g_k)^{p_k/2}},$$

with $M_k = I - X_k(X_k^T X_k)^{-1}X_k^T = I - P_k$, where P_k is the projection matrix onto the column space of X_k. Accordingly $y^T M_k y$ represents the residual sum of squares SSR_k of model $\mathcal{M}_k$ and similarly for M_s.

Notice that the marginalization procedure does not lead to the gNIGa prior (11). Indeed, conditionally on σ_k^2, the variance matrix of β_k^* is given by $g\sigma_k^2[(X^T \cdot X)^{-1}]_{kk}$, where $[(X^T X)^{-1}]_{kk}$ is the submatrix of $(X^T X)^{-1}$ containing the first k rows and k columns, which is not equal to $(X_k^T X_k)^{-1}$. This reason, together with the lack of invariance and of nuisance-coherence of the marginalization procedure in this case, suggest to disregard it in our future investigations.

4.1.1 *Standard Approach*. The conventional prior that is used in most Bayesian analyses of linear models *assumes* that, under $\mathcal{M}_k$, (β_k^*, σ_k^2) follows a gNIGa distribution, with hyperparameters $(b_k^{\mathrm{S}}, g, d, a)$, where the superscript S stands for "standard." Often the prior on σ_k^2 is taken to be improper ($d \to 0$ and $a \to 0$) and the resulting prior will be denoted with $\pi^{\mathrm{I}}(\beta_k^*, \sigma_k^2)$, where I stands for "improper." Standard choices for b_k^{S} reproduce the default options (9) and can be formally recovered as $b_k^{\mathrm{S}} = (X_k^T X_k)^{-1}X_k^T Xb$. Using results in Rao and Toutemburg (1999, pages 41–42), it can be checked that when $b = \hat{b}$ the corresponding b_k^{S} will coincide with the OLS estimate of β_k under $\mathcal{M}_k$.

We conclude this section remarking that the standard approach does not satisfy nuisance-coherence (it is enough to check that the marginal variance of y under $\mathcal{M}_k$ differs from that under $\mathcal{M}_k$); on the other hand nested-coherence trivially holds.

4.1.2 *Usual Conditioning*. The prior for (β_k^*, σ_k^2) in this case is given by

$$\begin{aligned}\pi_k^{\mathrm{UC}}(\beta_k^*, \sigma_k^2) &= \pi(\beta_k^*, \sigma_k^2 | \beta_{\setminus k} = 0) \\ &= \pi(\beta_k^* | \beta_{\setminus k} = 0, \sigma_k^2)\pi(\sigma_k^2 | \beta_{\setminus k} = 0) \\ &= \pi_k^{\mathrm{UC}}(\beta_k^* | \sigma_k^2)\pi_k^{\mathrm{UC}}(\sigma_k^2). \end{aligned} \tag{13}$$

It can be checked that the UC prior is gNIGa, that is,

$$\begin{aligned}&\pi_k^{\mathrm{UC}}(\beta_k^*, \sigma_k^2) \\ &\quad = \mathrm{NIGa}(\beta_k^*, \sigma_k^2; b_k^{\mathrm{UC}}, \\ &\qquad g_k^{\mathrm{UC}}(X_k^T X_k)^{-1}, d_k^{\mathrm{UC}}, a_k^{\mathrm{UC}})\end{aligned} \tag{14}$$

with

$$\begin{aligned} b_k^{\mathrm{UC}} &= b_k + (X_k^T X_k)^{-1}(X_k^T X_{\setminus k}) b_{\setminus k}, \\ g_k^{\mathrm{UC}} &= g, \quad d_k^{\mathrm{UC}} = d + (p - p_k), \end{aligned} \tag{15}$$

$$a_k^{\mathrm{UC}} = a + b_{\setminus k}^T X_{\setminus k}^T M_k X_{\setminus k} b_{\setminus k}. \tag{16}$$

Analogous results were derived in Poirier (1985). Notice that under UC the hyperparameters change across models. In particular d_k^{UC} increases as p_k decreases (the model becomes smaller). George and McCulloch (1997) also allow different priors for the variance under the various models, although their choice is not based on formal probabilistic derivations. In their case, the larger the model, the smaller the expected variance, which is not necessarily the case under UC. Notice that if $b_{\setminus k} = 0$, one obtains $E(\sigma_k^2) = a/(d_k^{\mathrm{UC}} - 2)$, which decreases as p_k decreases. While this feature may appear somewhat counterintuitive, it will turn out to have useful implications as detailed in Section 4.2.

4.1.3 *Kullback–Leibler Projection*. The following lemma is instrumental in deriving KL-projections.

LEMMA 3. *Consider the linear model $\mathcal{M}$ defined in* (6), *and the submodel $\mathcal{M}_k$ defined in* (10). *Then*

(i) *the KL-divergence between $\mathcal{M}$ and $\mathcal{M}_k$ is given by*

$$\begin{aligned} KL(\mathcal{M}, \mathcal{M}_k) = \frac{1}{2\sigma_k^2}(X\beta - X_k\beta_k^*)^T(X\beta - X_k\beta_k^*) \\ + \frac{n}{2}\left[\frac{\sigma^2}{\sigma_k^2} - \log\left(\frac{\sigma^2}{\sigma_k^2}\right) - 1\right]; \end{aligned}$$

(ii)

$$\arg\min_{\beta_k^*} KL(\mathcal{M}, \mathcal{M}_k) = \beta_k^{\perp} = (X_k^T X_k)^{-1} X_k^T X\beta; \tag{17}$$

(iii) $\arg\min_{\beta_k^*, \sigma_k^2} KL(\mathcal{M}, \mathcal{M}_k) = (\beta_k^{\perp}, \sigma_k^{2\perp})$,

where $\beta_k^{\perp}$ is defined in (17) *and*

$$\sigma_k^{2\perp} = \sigma^2 + Q_k(\beta), \tag{18}$$

with

$$\begin{aligned} Q_k(\beta) &= \frac{1}{n}\beta^T X^T M_k X\beta \\ &= \frac{1}{n}\beta_{\setminus k}^T X_{\setminus k}^T M_k X_{\setminus k}\beta_{\setminus k}. \end{aligned} \tag{19}$$

PROOF. Point (i) follows specializing to our case the KL-divergence between two multivariate normal distributions, given for example in Whittaker (1990, page 387). Points (ii) and (iii) are obtained by a direct calculation. □

We now distinguish two cases, namely projection with respect to β_k^* for given σ_k^2, and projection with respect to both β_k^* and σ_k^2. Consider the former case. This is appropriate, for instance, when we want to take the same prior on σ_k^2 for all models; in this case we need only minimize $KL(\mathcal{M}, \mathcal{M}_k)$ with respect to β_k^* and thus $\beta_k^{\perp}$ is given by (17) (for interesting related results, obtained using a predictive point of view, see Ibrahim, 1997, and Celeux, Marin and Robert, 2006).

PROPOSITION 4. *Consider the linear model $\mathcal{M}$ specified in* (6) *with a* $\mathrm{NIGa}(b, g(X^T X)^{-1}, d, a)$ *prior on (β, σ^2) described in* (7)–(8) *and a submodel $\mathcal{M}_k$ specified in* (10). *Conditionally on the assumption that σ_k^2 has the same distribution as σ^2, that is,* $\mathrm{IGa}(d/2, a/2)$, *the KL-prior on (β_k^*, σ_k^2) is given by*

$$\mathrm{NIGa}(\beta_k^*, \sigma_k^2; b_k^{\mathrm{KL}}, g(X_k^T X_k)^{-1}, d, a), \tag{20}$$

with

$$b_k^{\mathrm{KL}} = b_k + (X_k^T X_k)^{-1}(X_k^T X_{\setminus k}) b_{\setminus k}, \tag{21}$$

where $(b_k^T, b_{\setminus k}^T)$ is the decomposition of $b^T = E(\beta)^T$ corresponding to $\mathcal{M}_k$.

PROOF. Recalling that $\beta_k^{\perp}$ is a linear transformation of β, it follows immediately that the distribution of $\beta_k^{\perp}$ given σ_k^2 is normal. Now $E(\beta_k^{\perp} | \sigma_k^2) = (X_k^T X_k)^{-1} X_k^T X E(\beta) = (X_k^T X_k)^{-1} X_k^T X b$, and (21) follows immediately rewriting $X = (X_k \vdots X_{\setminus k})$ and $b^T = (b_k^T, b_{\setminus k}^T)$. Furthermore, $\mathrm{Var}(\beta_k^{\perp} | \sigma_k^2) = g \times \sigma_k^2 (X_k^T X_k)^{-1} W_k$, where $W_k = X_k^T P X_k (X_k^T X_k)^{-1}$ with $P = X(X^T X)^{-1} X^T$. Let now $M_{\setminus k} = (I - P_{\setminus k})$, where $P_{\setminus k}$ denote the projection matrix onto the column space of $X_{\setminus k}$. Using the equality $P = I - M_{\setminus k} + M_{\setminus k} X_k (X_k^T M_{\setminus k} X_k)^{-1} X_k^T M_{\setminus k}$ provided in Searle (1982, exercise 8, page 269), it follows that $W_k = I$, which gives the result. □

Consider now the projection with respect to (β_k^*, σ_k^2) whose corresponding expressions are provided in point (iii) of Lemma 3. Notice that $\beta_k^\perp$ is unchanged relative to the previous case; on the other hand $\sigma_k^{2\perp} \geq \sigma^2$ since $Q_k(\beta) \geq 0$. [This follows because $Q_k(\beta)$ can be written as $w^T w$ with $w = M_k X\beta$, using the fact that M_k is a projection matrix.] As a consequence the KL-projection variance under $\mathcal{M}_k$ will always exceed σ^2, justifying the intuition that the variance under $\mathcal{M}_k$ should be larger to account for a greater lack of fit. This case generalizes the simple linear regression example introduced shortly before Section 3.1.

The KL-prior of (β_k^*, σ_k^2), that is, that induced from (7) on $(\beta_k^\perp, \sigma_k^{2\perp})$, is unfortunately not analytically available, because of the awkward dependence of $\sigma_k^{2\perp}$ on (β, σ^2). Of course one can easily simulate from the KL-prior on (β_k^*, σ_k^2) using draws from the gNIGa prior on (β, σ^2) and mapping them into draws from π_k^{KL} through $(\beta_k^\perp, \sigma_k^{2\perp})$. However, we will not follow this course of action and derive an analytical approximation along the lines described in Consonni, Gutiérrez-Peña and Veronese (2007). Essentially, we employ a conjugate prior that minimizes the KL-divergence relative to the true π_k^{KL}. We call the resulting prior the KL-conjugate approximation, but for simplicity, we still identify it as π_k^{KL}. Specifically, we approximate the true KL-prior within the conjugate gNIGa family, whose hyperparameters $b_k^{\mathrm{KL}}, g_k^{\mathrm{KL}}, a_k^{\mathrm{KL}}, d_k^{\mathrm{KL}}$ are given in the following proposition.

PROPOSITION 5. *Consider the linear model $\mathcal{M}$ specified in (6) with a* NIGa$(b, g(X^T X)^{-1}, d, a)$ *prior on (β, σ^2) described in (7)–(8) and a submodel $\mathcal{M}_k$ specified in (10). Then the KL-conjugate approximation prior on (β_k^*, σ_k^2) is the* NIGa$(b_k^{\mathrm{KL}}, g_k^{\mathrm{KL}}(X_k^T X_k)^{-1}, d_k^{\mathrm{KL}}, a_k^{\mathrm{KL}})$ *where the hyperparameters can be identified in the following way*:

- *If $b_{\setminus k} = 0$, they are the solutions of the following system of equations*:

$$b_k^{\mathrm{KL}} = b_k, \tag{22}$$

$$g_k^{\mathrm{KL}} = g\,E(R_k(\beta, \sigma^2)), \tag{23}$$

$$a_k^{\mathrm{KL}} = d_k^{\mathrm{KL}} \frac{a}{d} \frac{1}{E[R_k(\beta, \sigma^2)]}, \tag{24}$$

$$\begin{aligned} \psi(d_k^{\mathrm{KL}}/2) - \log(d_k^{\mathrm{KL}}/2) &= \psi(d/2) - \log(d/2) \\ &\quad + E\{\log[R_k(\beta, \sigma^2)]\} \\ &\quad - \log\{E[R_k(\beta, \sigma^2)]\}, \end{aligned} \tag{25}$$

  where $R_k(\beta, \sigma^2) = (1 + Q_k(\beta)/\sigma^2)^{-1}$, and $\psi(\alpha) = \frac{\partial}{\partial \alpha} \log(\Gamma(\alpha))$ is the digamma function.
- *If $b_{\setminus k} \neq 0$, they are* approximately *the solutions of the following system of equations*:

$$b_k^{\mathrm{KL}} = b_k + (X_k^T X_k)^{-1}(X_k^T X_{\setminus k}) b_{\setminus k}, \tag{26}$$

$$g_k^{\mathrm{KL}} = \frac{g}{E[R_k(\beta, \sigma^2)^{-1}]}, \tag{27}$$

$$a_k^{\mathrm{KL}} = d_k^{\mathrm{KL}} \frac{a}{d} E[R_k(\beta, \sigma^2)^{-1}] \tag{28}$$

  and

$$\begin{aligned} &\psi(d_k^{\mathrm{KL}}/2) - \log(d_k^{\mathrm{KL}}/2) \\ &\quad = \psi(d/2) - \log(d/2) \\ &\qquad + \frac{1}{2} \frac{\mathrm{Var}[R_k(\beta, \sigma^2)^{-1}]}{E[R_k(\beta, \sigma^2)^{-1}]^2}. \end{aligned} \tag{29}$$

The analytical expressions for $E[R_k(\beta, \sigma^2)]$, $E[R_k(\beta, \sigma^2)^{-1}]$ and $\mathrm{Var}[R_k(\beta, \sigma^2)^{-1}]$ are given in Lemma A.1 in the Appendix.

PROOF. See the Appendix. □

Notice that both the expressions of b_k^{KL} in Propositions 4 and 5 coincide with that of b_k^{UC}. Furthermore, (22)–(25), as well as (26)–(29), do not admit a closed-form solution. Yet, a few results can be established which we report without proof: $d_k^{\mathrm{KL}} < d$; $d_k^{\mathrm{KL}}/d \to 0$ for $d \to \infty$; $a_k^{\mathrm{KL}} \to 0$ for $d \to \infty$, whence $a_k^{\mathrm{KL}} < a$ for large d; $E(\sigma_k^{-2}) = d_k^{\mathrm{KL}}/a_k^{\mathrm{KL}} < d/a = E(\sigma^{-2})$, as expected. Finally nested-coherence is satisfied on the space or regression parameter, while it fails on the variance space. Moreover it can be established empirically that nuisance-coherence fails.

4.2 Information Paradox

A major objection to the use of g-priors falls under the heading of Information Paradox; see Liang et al. (2008) for a recent discussion. Suppose that the regression model $\mathcal{M}_k$ is compared with the "Null" model $\mathcal{M}_0$ having no predictors. Assume the data overwhelmingly support $\mathcal{M}_k$, that is, $\|\beta_k\|^2 = \beta_k^T \beta_k \to \infty$, so that the coefficient R^2 under $\mathcal{M}_k$ tends to 1 and $SSR_k = y^T M_k y \to 0$. Using a g-prior under both models with zero expectation for the regression parameters and $d_k = d$, $a_k = a$ and $g_k = g$, the Bayes factor B_{k0} of $\mathcal{M}_k$ against $\mathcal{M}_0$ remains bounded whereas one would expect it to diverge. However, the paradox does not necessarily arise if we assume *different* g-priors under the two models as implied by the UC and KL-procedures, as we now show.

First notice that $\beta_k^T \beta_k \to \infty$ implies also $y^T y \to \infty$. If $b = E(\beta)$ is independent of the data, for example; b_0 in (9), it can be easily checked using (12) that B_{k0} is asymptotic to

$$\frac{(y^T A y)^{(d_0+n)/2}}{(1/(1+g_k) y^T y)^{(d_k+n)/2}} \quad \text{with } A = \left(I - \frac{g_0}{n(1+g_0)} J\right),$$

where I is the identity matrix and J is the matrix with all elements equal to 1. Since $\lambda_{\min} \le y^T A y / y^T y \le \lambda_{\max}$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of A, it follows that $y^T A y = O(y^T y)$ since $\lambda_{\min} > 0$. As a consequence the limiting behavior of B_{k0} depends on the hyperparameters d_0 and d_k deduced from the specific compatible procedure. In the case of UC, we have $g_k^{\mathrm{UC}} = g$ and $d_k^{\mathrm{UC}} = d + (p - p_k) < d + (p-1) = d_0^{\mathrm{UC}}$ and thus $B_{k0} \to \infty$ so that the paradox does not arise. However, this result does not hold for the KL-procedure, since $d_k^{\mathrm{KL}} > d_0^{\mathrm{KL}}$. The same conclusions can be obtained, using similar arguments, if we assume $E(\beta) = \bar{b} = (\bar{y}, 0, \ldots, 0)$.

Suppose now $E(\beta) = \hat{b}$, that is, the expectation of β is fully data-dependent. In this case both b_k^{UC} and b_k^{KL} reduce to the OLS estimate of β_k under $\mathcal{M}_k$, that is, $b_k^{\mathrm{UC}} = b_k^{\mathrm{KL}} = (X_k^T X_k)^{-1} X_k^T y$, while $b_0^{\mathrm{UC}} = b_0^{\mathrm{KL}} = \bar{y}$. Thus, from (12), B_{k0} is asymptotic to

$$C_{k0} \frac{(a_0 + y^T M_0 y)^{(d_0+n)/2}}{a_k^{(d_k+n)/2}} \quad \text{with } M_0 = \left(I - \frac{1}{n} J\right). \tag{30}$$

Under the UC procedure, only the hyperparameter a_k^{UC} can depend on the data y through $\hat{b}$ [see (15) and (16)], and we have

$$\begin{aligned} a_k^{\mathrm{UC}} &= a + y^T M_k X_{\setminus k}^T (X_{\setminus k}^T M_k X_{\setminus k})^{-1} X_{\setminus k}^T M_k y \\ &= a + y^T (P - P_k) y \\ &= a + y^T (M_k - M) y \to a, \end{aligned} \tag{31}$$

recalling that $b_{\setminus k}^{\mathrm{UC}} = \hat{\beta}_{\setminus k} = (X_{\setminus k}^T M_k X_{\setminus k})^{-1} X_{\setminus k}^T M_k y$, and using formula 3.98 on page 42 and Theorem A.45 on page 367 in Rao and Toutemburg (1999). The result follows noting that $y^T (M_k - M) y \to 0$ because the *SSR* of $\mathcal{M}$ must be less than that of $\mathcal{M}_k$ which tends to zero by hypothesis. Thus B_{k0} in (30) trivially goes to infinity, since $C_{k0} \to$ *constant* and $y^T M_0 y \to \infty$, and there is no paradox.

Under the KL-procedure instead, from (25), (58) and (59), it appears that the dependence of the hyperparameters on the data happens only through $Q_k(\hat{\beta})$. Now

$$\begin{aligned} Q_k(\hat{\beta}) &= \frac{1}{n} \hat{\beta}^T X^T M_k X \hat{\beta} = \frac{1}{n} y^T P M_k P y \\ &= \frac{1}{n} y^T (P - P_k) y = \frac{1}{n} y^T (M_k - M) y \end{aligned}$$

which tends to zero as in (31). Accordingly the hyperparameters behave as constants in the limit, and thus also in this case the information paradox does not arise.

5. EXAMPLES

In this section we present three examples in order to evaluate the performance of the various priors discussed in Section 4.1. The first one considers the very simple situation of testing a normal model with a submodel $\mathcal{M}^*$ having mean zero: in this way different priors of σ^{*2} can be more easily compared. Features of the priors, and their consequences on variable selection, are then assessed in a more complex simulation study, and in a real data set (Hald data), frequently analyzed in the literature.

5.1 A Simple Illustration

Consider the two models

$$\begin{aligned} \mathcal{M}&: \; y_i = \mu + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\ \mathcal{M}^*&: \; y_i = \varepsilon_i^*, \quad \varepsilon_i^* \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^{*2}), \end{aligned}$$

with $i = 1, \ldots, n$, and assume as a prior for (μ, σ^2) under $\mathcal{M}$ the following gNIGa: $\pi(\mu|\sigma^2) = \mathrm{N}(\mu; 0, g\sigma^2/n)$; $\pi(\sigma^2) = \mathrm{IGa}(\sigma^2; d/2, a/2)$. If $St_n(\cdot; \eta, \Lambda, \nu)$ denotes an n-dimensional Student t-distribution with expectation η, degrees of freedom ν and variance matrix $\nu\Lambda^{-1}/(\nu-2)$, $\nu > 2$, the marginal density of y is

$$f(y) = St_n\left(y; 0, \frac{d}{a}\left(I - \frac{g/n}{1+g} J\right), d\right).$$

The submodel $\mathcal{M}^*$ only requires a prior on σ^{*2}. The Standard, UC and KL-procedures lead to priors π^{S}, π^{UC} and π^{KL} for σ^{*2} which are all of type $\mathrm{IGa}(d^*/2, a^*/2)$. Specifically, one obtains

$$\begin{aligned} &(d^{\mathrm{S}} = d, \quad a^{\mathrm{S}} = a); \\ &(d^{\mathrm{UC}} = d + 1, \quad a^{\mathrm{UC}} = a). \end{aligned} \tag{32}$$

We consider also the typical improper prior on σ^2 given by $\pi^{\mathrm{I}}(\sigma^2) \propto \sigma^{-2}$ which can be formally obtained from π^{S} setting $d = 0$, $a = 0$. Consider now the

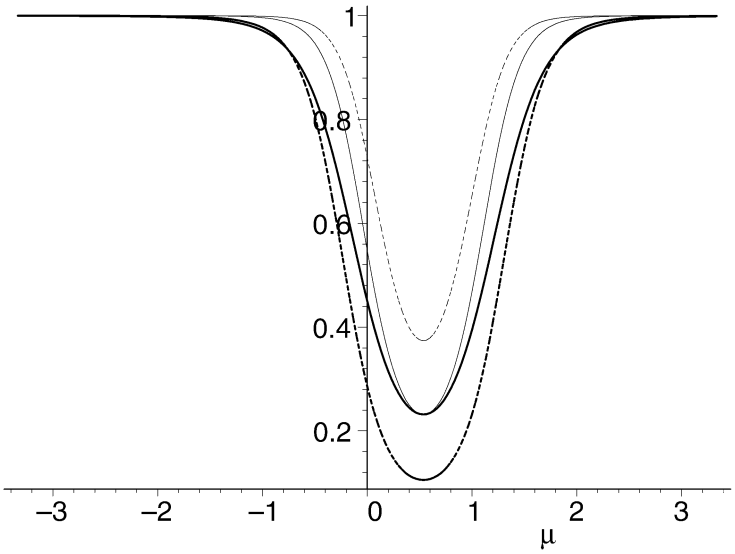

FIG. 1. *Posterior probability of $\mathcal{M}$ for hyperparameters $d = 5$, $a = 1$: $p^{\mathrm{KL}}(\mathcal{M}|y)$ dash thick, $p^{\mathrm{S}}(\mathcal{M}|y)$ solid thin, $p^{\mathrm{UC}}(\mathcal{M}|y)$ dash thin, $p^{\mathrm{I}}(\mathcal{M}|y)$ solid thick.*

KL-prior. A direct computation yields $\sigma^{2\perp} = \sigma^2 + \mu^2$, which can also be deduced from (18) by setting P_k equal to the zero matrix since in this case X_k is void, so that $Q_k(\mu) = \mu^2$. The values of d^{KL} and a^{KL} can be recovered from (24) and (25). For illustration, in the following we use three different values of (d, a), namely $(d = 1, a = 1)$, $(d = 5, a = 1)$, $(d = 3, a = 25)$ leading respectively to $(d^{\mathrm{KL}} = 0.93, a^{\mathrm{KL}} = 1.42)$, $(d^{\mathrm{KL}} = 3.38, a^{\mathrm{KL}} = 1.03)$, $(d^{\mathrm{KL}} = 2.36, a^{\mathrm{KL}} = 29.98)$.

In order to appreciate the effect of the different priors, we compute the posterior probability of the two models $\mathcal{M}$ and $\mathcal{M}^*$. In particular assuming prior odds 1, we have $\Pr(\mathcal{M}|y) = 1/(1 + B^*)$, where $B^* = f^*(y)/f(y)$ is the Bayes factor of $\mathcal{M}^*$ versus $\mathcal{M}$. Notice that $f^*(y) = St_n(y; 0, (d^*/a^*)I, d^*)$ with d^* and a^* depending on the specific procedure. We fix $n = g = 25$ and perform a simulation study, generating a vector ε from a multivariate standard normal distribution, and set $y = \mu\iota_n + \varepsilon$, where ι_n is the n-dimensional unit vector. In Figures 1 and 2 the posterior probability of $\mathcal{M}$ is plotted as a function of μ. Notice that the minimum of the curves does not occur at $\mu = 0$, because

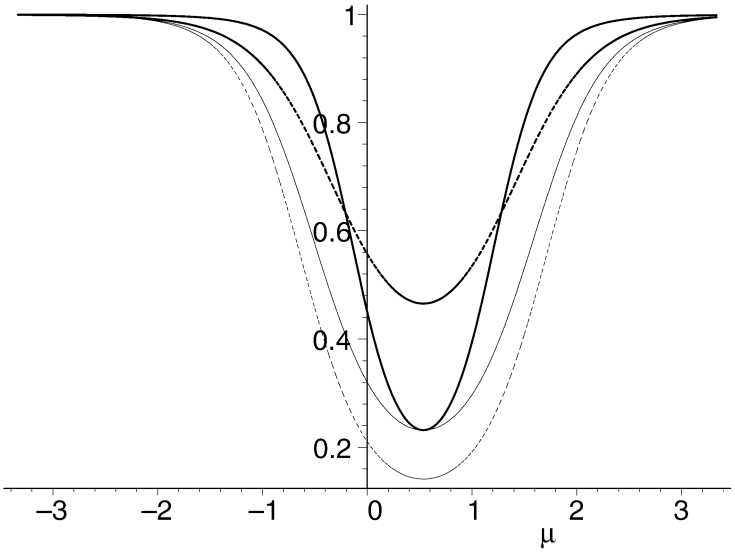

FIG. 2. *Posterior probability of $\mathcal{M}$ for hyperparameters $d = 3$, $a = 25$: $p^{\mathrm{KL}}(\mathcal{M}|y)$ dash thick, $p^{\mathrm{S}}(\mathcal{M}|y)$ solid thin, $p^{\mathrm{UC}}(\mathcal{M}|y)$ dash thin, $p^{\mathrm{I}}(\mathcal{M}|y)$ solid thick.*

the generated errors in the simulation had a negative mean of about -0.5. Ideally the posterior probability curve should reach a minimum close to zero for $\mu \approx 0$ and then increase rapidly as μ moves away from zero. When $d = a$ all curves overlap to a large extent. Differences emerge for unequal a and d with the curves corresponding to π^{I} and π^{S} occupying intermediate positions, while those associated to π^{KL} and π^{UC} represent "extreme" curves. A strong sensitivity of π^{UC} and π^{KL} is apparent and in particular when a is greater than d, π^{UC} favors $\mathcal{M}^*$ most strongly, while π^{KL} favors $\mathcal{M}$ (and conversely when d is greater than a). For $a > d$, the curve corresponding to π^{S} is somewhat flatter than that under π^{I}.

We now consider the problem of model comparison from a predictive viewpoint as described in Gelfand and Ghosh (1998); see also Marriot, Spencer and Pettitt (2001). In the simple case corresponding to squared error loss, each model $\mathcal{M}_k$ is assigned a score $D^{(k)}$ made up of two parts: an error sum of squares component $G^{(k)}$ and a predictive variance component $P^{(k)}$,

$$D^{(k)} = \frac{c}{c+1} G^{(k)} + P^{(k)}, \quad c > 0, \tag{33}$$

where

$$G^{(k)} = \sum_{i=1}^{n} \bigl(\mu_i^{(k)} - y_i\bigr)^2,$$

$$P^{(k)} = \sum_{i=1}^{n} \sigma_i^{2\,(k)};$$

$$\mu_i^{(k)} = E^{(k)}(y_{i,\mathrm{rep}}|y),$$

$$\sigma_i^{2\,(k)} = \mathrm{Var}^{(k)}(y_{i,\mathrm{rep}}|y).$$

In the above setting $y^T = (y_1, \ldots, y_n)$ are the data, while $y_{i,\mathrm{rep}}$ represents a future replicate observation (the number of replicates being equal to that of the data). Model selection is achieved through a minimization of $D^{(k)}$ for a given choice of c. The term $P^{(k)}$ represents a penalty which aims at discouraging models that either strongly underfit or overfit the data, because in both cases predictive variances will tend to be inflated. Since our objective is to compare the performances of the various priors under model $\mathcal{M}^*$ we simply need to evaluate D^* for each distinct prior.

Consider first μ_i^*. This is

$$\mu_i^* = E^*(y_{i,\mathrm{rep}}|y) = E^*[E^*(y_{i,\mathrm{rep}}|y, \sigma^{2*})|y]$$

$$= E^*[E^*(y_{i,\mathrm{rep}}|\sigma^{2*})|y] = 0,$$

since under $\mathcal{M}^*$ each observation has expectation zero, conditionally on σ^{2*}. As a consequence $D^* = P^* + \sum_{i=1}^n y_i^2$, and thus only the term P^* matters for comparison purposes. Now

$$\sigma_i^{2*} = \mathrm{Var}^*(y_{i,\mathrm{rep}}|y) = E^*[\mathrm{Var}^*(y_{i,\mathrm{rep}}|y, \sigma^{2*})|y]$$

$$= E^*(\sigma^{2*}|y) = \frac{a_n^*}{d_n^* - 2}, \quad d_n^* - 2 > 0,$$

since under each prior the posterior distribution of σ_i^{*2} is $\mathrm{IGa}(d_n^*/2, a_n^*/2)$, with $d_n^* = d^* + n$, and $a_n^* = a^* + \sum_{i=1}^n y_i^2$. In conclusion the predictive criterion of Gelfand and Ghosh (1998) suggests to base model comparison on $P^* = n a_n^*/(d_n^* - 2)$.

From (32), it is immediate to conclude that $P^{\mathrm{UC}} < P^{\mathrm{S}}$ so that π^{UC} supports $\mathcal{M}^*$ more than π^{S}. On the other hand, since $d^{\mathrm{KL}} < d$ it follows that $P^{\mathrm{KL}} > P^{\mathrm{S}}$ whenever $a^{\mathrm{KL}} > a^{\mathrm{S}}$ (calculations show that this occurs for moderate values of d, specifically $d < 5.45$); in other words the KL-prior would tend to favor $\mathcal{M}^*$ less than π^{S}. These conclusions are broadly in accord with the curves describing $P(\mathcal{M}|y)$ depicted in Figures 1 and 2.

5.2 Simulation Study

As a second example, we consider a simulation study along the lines presented in George and McCulloch (1993), Raftery, Madigan and Hoeting (1997) and Fernández, Ley and Steel (2001). We consider $p = 6$ predictors, the constant plus $(X_1, \ldots, X_5)$ and $n = 30$ observations. Let Z_j, $j = 1, \ldots, 5$ be independent n-dimensional vectors, whose components are independent standard normal variables, and set

$$X_1 = Z_1, \quad X_2 = Z_2, \quad X_3 = Z_3,$$

$$(X_4, X_5) = (X_1, X_2)(0.3\ 0.7)^T (1\ 1) + (Z_4, Z_5).$$

In this way there is a correlation between the first two predictors and the last two. We generate the response y according to three different models:

$$\mathcal{M}_1 : y = C + 2.5\varepsilon, \tag{34}$$

$$\mathcal{M}_2 : y = C + 2X_1 - X_3 + 1.5X_5 + 2.5\varepsilon, \tag{35}$$

$$\mathcal{M}_3 : y = C + 2X_1 - X_3 + X_4 + 1.5X_5 + 2.5\varepsilon, \tag{36}$$

where C is a fixed constant and the n elements of ε are independent standard normal variables. In particular, the case in which the data were generated from $\mathcal{M}_1$ was analyzed in a frequentist way by Freedman (1983). He showed that, under this "null model," standard variable selection procedures, such as stepwise regression, may lead to misleading results, for example, retaining a subset of predictors with a highly significant F-statistic and reasonably high R^2.

TABLE 1
Frequency of correct identification of the true model $\mathcal{M}_i$ $(i = 1, 2, 3)$ with $g = n = 30$ for various compatible priors and different choices of (d, a) and $E(\beta)$

		$\pi^{\mathbf{KL}}$			$\pi^{\mathbf{S}}$			$\pi^{\mathbf{UC}}$			$\pi^{\mathbf{I}}$		
d	a	b_0	$\bar{b}$	$\hat{b}$	b_0	$\bar{b}$	$\hat{b}$	b_0	$\bar{b}$	$\hat{b}$	b_0	$\bar{b}$	$\hat{b}$
$\mathcal{M}_1$ true model													
0	0										0.60	0.56	0.54
1	1	0.24	0.40	0.24	0.56	0.54	0.52	0	0	0.76			
1	10	0.08	0.24	0.48	0.64	0.56	0.56	0.32	0.26	0.64			
5	5	0.26	0.44	0.24	0.50	0.50	0.48	0.06	0.06	0.86			
10	1	0.34	0.48	0.30	0.40	0.36	0.36	0	0	0.96			
10	50	0.04	0.06	0	0.56	0.54	0.52	0.46	0.42	0.60			
$\mathcal{M}_2$ true model													
0	0										0.46	0.56	0.60
1	1	0.70	0.60	0.66	0.48	0.58	0.60	0.68	0.68	0.32			
1	10	0.70	0.66	0.64	0.42	0.54	0.58	0.58	0.62	0.56			
5	5	0.68	0.60	0.68	0.58	0.60	0.62	0.64	0.64	0.18			
10	1	0.66	0.52	0.60	0.62	0.64	0.66	0	0	0.02			
10	50	0.66	0.68	0.68	0.50	0.58	0.60	0.58	0.62	0.58			
$\mathcal{M}_3$ true model													
0	0										0.26	0.38	0.54
1	1	0.64	0.44	0.66	0.26	0.42	0.54	0.68	0.68	0.26			
1	10	0.74	0.54	0.52	0.24	0.36	0.05	0.30	0.42	0.54			
5	5	0.54	0.32	0.50	0.40	0.54	0.56	0.64	0.64	0.04			
10	1	0.22	0.16	0.66	0.56	0.56	0.60	0	0	0			
10	50	0.74	0.56	0.60	0.34	0.48	0.54	0.50	0.54	0.52			

In order to compare the different priors, we consider the Bayes factor for each submodel versus the full model with six predictors (including the constant) for 50 simulated data sets and report the frequency of times in which the highest Bayes factor is associated to the correct model (i.e., the model which has generated the data). We fix $g = n$ and for each choice of $E(\beta)$, namely $b_0, \bar{b}, \hat{b}$ [see (9)] check the robustness of the various priors to the choice of the hyperparameters (d, a) of the inverse-gamma distribution on σ^2 (each time leaving unchanged the values of the predictors).

We can summarize our results, which are in part reported in Table 1, as follows:

(i) π^{UC} appears to be the least robust prior relative to the various choices of $E(\beta)$ and (d, a); this is consistent with the fact that the marginal of the data under π^{UC} is more peaked on its expectation; see the discussion in Section 5.1. Its frequency of correct model identification can reach very low values especially when d exceeds a, in accord with the fact that as d increases relative to a larger models receive greater support under π^{UC}; see Figure 1. To provide an explanation of this phenomenon, consider the Bayes factor B_k of the submodel $\mathcal{M}_k$ versus the full model $\mathcal{M}$. If the prior under $\mathcal{M}_k$ is obtained through UC, then calculations show that

$$B_k = \frac{\pi(\beta_{\setminus k} = 0|y)}{\pi(\beta_{\setminus k} = 0)}, \tag{37}$$

where $\pi(\beta_{\setminus k} = 0|y)$ and $\pi(\beta_{\setminus k} = 0)$ are respectively the marginal posterior and prior density of $\beta_{\setminus k}$, evaluated at the value 0. The expression (37) for B_k is known as "Savage's density ratio"; see, for example, O'Hagan and Forster (2004, Section 7.16). Now if the data are at least moderately more informative than the prior, the numerator will be essentially dominated by the likelihood, and thus will be fairly robust to prior specifications, while this does not clearly occur for the denominator. In particular, if d increases relative to a, the distribution of σ^2 tends to concentrate on smaller values, so that the marginal of $\beta_{\setminus k}$ becomes more peaked around the mode (which coincides with 0 under b_0 or $\bar{b}$), thus lowering B_k, and supporting $\mathcal{M}$ more than $\mathcal{M}_k$.

(ii) π^{KL} is reasonably robust and shows good performance, save when the generating model corresponds to the "null model" $\mathcal{M}_1$ and a is large (this

is in accord with the fact exhibited in Figure 2 that for large a bigger models are preferred under π^{KL}).

(iii) π^{S} and π^{I} exhibit a relatively similar behavior, as already remarked in the previous section, and have a better performance than the other priors at identifying the "null model."

Overall, the frequency of correct model identification is comparable, or even superior, to similar investigations carried out in a Bayesian framework, although using different model choice criteria and different priors; see Marriot, Spencer and Pettitt (2001).

5.3 Hald Data

Our third example involves the Hald data, often analyzed in the literature, in order to evaluate model selection procedures; see, for instance, Draper and Smith (1981). It consists of 13 observations on one response variable with four predictors. A specific feature of this data set is represented by the strong correlation between X_1 and X_3 and between X_2 and X_4. We consider all the possible 16 models in which the constant term is always included.

A detailed subjective Bayesian analysis of this data set has been performed in Laud and Ibrahim (1995, 1996) and Ibrahim (1997), especially in terms of prior specification. We follow Laud and Ibrahim (1995) and fix a prior on (β, σ^2) under the full model which is a $\mathrm{NIGa}(\tilde{b}, g(X^TX)^{-1}, 25, 125)$ with $E(\beta) = \tilde{b} = (X^TX)^{-1}X^T\eta$, where η is a subjective prediction for y given by $\eta = (79, 77, 104, 90, 99, 108, 105, 73, 93, 111, 88, 115, 113)$. We also report the value $\gamma = 1/(g+1)$, which represents a weight on the prior guess η. Notice that the choice of $d = 25$ and $a = 125$ implies $E(\sigma^{-2}) = 0.2$ and $\Pr(\sigma^{-2} < 0.5) \approx 0.95$.

Table 2 summarizes the results of a Bayesian analysis using the conventional value $g = n = 13$, as well as $g = 9$ (Ibrahim's choice) which correspond to weights $\gamma = 0.07$, respectively 0.10, representing weak prior information. Moreover we consider two choices for $E(\beta)$, namely $\bar{b}$ and $\tilde{b}$. We do not report explicitly results for $E(\beta) = b_0$ because posterior model probabilities are relatively more diffuse and no subset of models emerges as a clear winner. The column π^{Ibr} reports the results obtained in Ibrahim (1997) which assumes a fixed $\sigma^{-2} = 0.2$. The highest probability is given to model $\{1, 2\}$ under all priors, save for π^{KL} that indicates a slight preference for more complex models, for example, $\{1, 2, 4\}$ for $g = 13$. Overall there is broad agreement with standard frequentist model selection procedures as reported in Laud and Ibrahim (1995, Table 1).

We also performed a sensitivity analysis (not reported here) with respect to γ $(0.01 \le \gamma \le 0.95)$ for the two choices $E(\beta) = b_0$, respectively $\tilde{b}$, in order to make a comparison with the results of Tables 2 and 3 of Ibrahim (1997). The results are appreciably sensitive to the choice of b_0 or $\tilde{b}$, although this fact is definitely

TABLE 2

Posterior probability of top four models with $g = n = 13$ ($\gamma = 0.07$) and $g = 9$ ($\gamma = 0.1$) for various compatible priors and different choices of $E(\beta)$; in first column is Ibrahim's results

		π^{KL}		π^{S}		π^{UC}		π^{I}	
Model	π^{Ibr}	$\bar{b}$	$\tilde{b}$	$\bar{b}$	$\tilde{b}$	$\bar{b}$	$\tilde{b}$	$\bar{b}$	$\tilde{b}$
$g = 13$									
$\{1, 2\}$		0.175	0.203	0.340	0.290	0.276	0.293	0.329	0.271
$\{1, 4\}$								0.221	
$\{1, 2, 3\}$		0.181	0.227	0.145	0.207	0.167	0.211	0.112	0.213
$\{1, 2, 4\}$		0.184	0.234	0.151	0.220	0.174	0.223	0.114	0.229
$\{1, 3, 4\}$		0.169	0.174	0.127	0.155	0.147	0.146		0.153
Total		0.709	0.838	0.763	0.872	0.764	0.873	0.776	0.866
$g = 9$									
$\{1, 2\}$	0.272	0.217	0.210	0.310	0.262	0.238	0.268	0.294	0.248
$\{1, 4\}$		0.171		0.165				0.219	
$\{1, 2, 3\}$	0.215	0.157	0.230	0.143	0.215	0.171	0.222	0.111	0.219
$\{1, 2, 4\}$	0.214	0.156	0.216	0.143	0.209	0.171	0.217	0.111	0.213
$\{1, 3, 4\}$	0.164		0.173		0.163	0.153	0.157		0.159
Total	0.865	0.701	0.829	0.761	0.852	0.733	0.864	0.735	0.839

less manifest for the prior π^{Ibr} (under which, however, σ^2 is assumed fixed). Overall it is confirmed that the choice of b_0 is the least satisfactory, as it tends to shift posterior model probability toward "extreme" models, such as the null or full model, when γ approaches either boundary. On the other hand, under $\tilde{b}$ the results are fairly insensitive to the choice of γ as far as the identification of the top model is concerned, which is usually $\{1, 2\}$, and either $\{1, 2, 3\}$ or $\{1, 2, 4\}$. In particular π^{KL} exhibits a high stability, with respect to γ, of the posterior probability mass on the top model which always contains three predictors.

The Hald data have been also analyzed in a Bayesian objective framework, in particular by Berger and Pericchi (1996b) using intrinsic Bayes factor, and by Casella and Moreno (2006) and Moreno and Giron (2007) using intrinsic priors. The models they identify are essentially those exhibited as most probable in Table 2. However, under their approach, model $\{1, 2\}$ receives a posterior probability in excess of 50%. Based on an objective predictive approach, Barbieri and Berger (2004) develop a theory for model choice. They show that the optimal model is not necessarily the highest posterior probability model, but rather the "median probability model." For the Hald data the latter is represented by $\{1, 2, 4\}$ which, curiously, is also the model with the highest posterior probability under the KL-prior with $g = n$; see Table 2.

6. DISCUSSION

For a given proper prior on the parameter space of a full model, we reviewed and analyzed procedures for the specification of prior distributions on the parameter space of a collection of submodels. We presented two interpretations of nested models, in order to explicate more naturally the rationale of each procedure. In particular, we investigated four methods for the specification of a compatible prior under a submodel, namely marginalization, usual and Jeffreys conditioning and Kullback–Leibler projection. Next, each procedure was evaluated from two perspectives, nuisance- and nested-coherence. Given a full linear model with a normal inverted gamma g-prior on the parameters, we considered the problem of variable selection, and applied the above procedures for the construction of priors under each submodel $\mathcal{M}_k$. For completeness we also considered, for each $\mathcal{M}_k$, a g-prior on the regression parameters combined with an inverted gamma (d, a) distribution on σ_k^2, labeled π^{S}, as well as a conventional improper prior on σ_k^2, identified with π^{I}.

Three examples were used to illustrate the behavior of the various procedures for prior specification, leading to the conclusions that results are quite sensitive to the choice of the hyperparameters. Overall the improper prior π^{I} performs comparably to the standard prior π^{S}, when d and a are similar. The usual conditioning prior π^{UC}, despite its theoretically attractive coherence properties exhibited in Propositions 1 and 2, shows remarkable sensitivity to the choice of the hyperparameters, oscillating between highly simple and complicated models. The Kullback–Leibler projection prior exhibits a performance which is comparable or superior to that of π^{S} when using the OLS estimate as prior expectation on β, provided that the true model is not very close to the "null" model with no predictors. This is consistent with the general attitude of the KL-prior to favor more complex models.

When the goal of model choice is prediction, one might consider orthogonalizing the matrix of predictors, as in Clyde, DeSimone and Parmigiani (1996). In this case a g-prior on the regression coefficient under the full model admits a diagonal variance matrix. As a consequence the M, UC and KL-procedures would generate the same prior under each submodel $\mathcal{M}_k$ conditionally on σ_k^2; yet they would imply distinct priors for the variance. We remark, however, that this approach cannot be implemented in a variable selection problem, where the focus is on the original predictors.

Consistency of the posterior distribution on model space under different choices of the hyperparameter g_k^* in the gNIGa prior (11), with $d_k = d$ and $a_k = a$, has been recently discussed in Fernández, Ley and Steel (2001). They prove, under mild conditions, that consistency obtains under both the standard and improper priors π^{S} and π^{I}. Using similar arguments one can prove that the same result holds for the UC procedure under b_0 and $\bar{b}$, defined in (9). As far as π^{KL} is concerned the limiting probability of model $\mathcal{M}_k$ is zero provided the true model is not nested within $\mathcal{M}_k$; on the other hand when $\mathcal{M}_k$ is moderately larger than the true model this result may fail, and π^{KL} may lead to choose slightly overparametrized models.

It is well known that a standard use of g-priors for variable selection cannot be recommended because it suffers from the information paradox. However, our analysis shows that, when g-priors under submodels are *derived* using compatibility criteria, the paradox either does not arise (UC procedure), or can be avoided (KL-procedure) through a suitable choice of the initial hyperparameters.

Recent contributions in the area of linear models (see Liang et al., 2008 and Bayarri and Garcia-Donato, 2007), suggest to use a noninformative improper prior on the nuisance parameter and a proper mixture of g-priors on the regression coefficients. It would be interesting to apply the methods discussed in this paper to the latter distribution of the regression coefficients in order to derive a compatible mixture of g-priors under the various submodels.

APPENDIX

PROOF OF PROPOSITION 1. Assume that the sampling distribution under model $\mathcal{M}$ is $\{f(y|\lambda,\delta,\phi)\}$, where δ is the nuisance parameter. Then, for a given prior $\pi(\lambda,\delta,\phi)$, the integrated model $\mathcal{IM}$ has sampling distribution $f(y|\lambda,\phi) = \int f(y|\lambda,\delta,\phi)\pi(\delta|\lambda,\phi)\,d\delta$, while the corresponding integrated submodel $\mathcal{IM}^*$ has density $f^*(y|\lambda) = f(y|\lambda,\phi=\phi_0)$. Let the prior under $\mathcal{IM}$ be $\pi_{\mathcal{IM}}(\lambda,\phi) = \pi(\lambda,\phi)$, that is, the marginal distribution of (λ,ϕ) under π. Consider now a procedure to construct a prior under a submodel. Let

$$
\begin{aligned}
&f^*_{\mathcal{M}^*}(y)\\
&= \int\int f^*(y|\lambda,\delta)\pi^*_{\mathcal{M}^*}(\lambda,\delta)\,d\lambda\,d\delta \qquad (38)\\
&= \int\int f(y|\lambda,\delta,\phi=\phi_0)\pi^*_{\mathcal{M}^*}(\lambda,\delta)\,d\lambda\,d\delta
\end{aligned}
$$

and

$$
\begin{aligned}
&f^*_{\mathcal{IM}^*}(y)\\
&= \int f^*(y|\lambda)\pi^*_{\mathcal{IM}^*}(\lambda)\,d\lambda\\
&= \int\left\{\int f(y|\lambda,\delta,\phi=\phi_0)\pi(\delta|\lambda,\phi=\phi_0)\,d\delta\right\} \qquad (39)\\
&\qquad\cdot \pi^*_{\mathcal{IM}^*}(\lambda)\,d\lambda,
\end{aligned}
$$

where $\pi^*_{\mathcal{M}^*}(\lambda,\delta)$ is the output of the procedure applied to $(\mathcal{M},\mathcal{M}^*)$ starting from $\pi(\lambda,\delta,\phi)$, while $\pi^*_{\mathcal{IM}^*}(\lambda)$ is the output of the procedure applied to $(\mathcal{IM},\mathcal{IM}^*)$ starting from $\pi(\lambda,\phi)$.

(i) Recall that $\pi^{\mathrm{UC}}_{\mathcal{M}^*}(\lambda,\delta) = \pi(\lambda,\delta|\phi=\phi_0)$ and consider $\pi^{\mathrm{UC}}_{\mathcal{IM}^*}$. We have $\pi^{\mathrm{UC}}_{\mathcal{IM}^*}(\lambda) = \pi_{\mathcal{IM}}(\lambda|\phi=\phi_0) = \pi(\lambda|\phi=\phi_0)$. As a consequence we get from (38)

$$
\begin{aligned}
&f^{\mathrm{UC}}_{\mathcal{M}^*}(y)\\
&= \int\left\{\int f(y|\lambda,\delta,\phi=\phi_0)\pi(\delta|\lambda,\phi=\phi_0)\,d\delta\right\} \qquad (40)\\
&\qquad\cdot \pi(\lambda|\phi=\phi_0)\,d\lambda,
\end{aligned}
$$

while from (39) we get

$$
\begin{aligned}
&f^{\mathrm{UC}}_{\mathcal{IM}^*}(y)\\
&= \int\left\{\int f(y|\lambda,\delta,\phi=\phi_0)\pi(\delta|\lambda,\phi=\phi_0)\,d\delta\right\} \qquad (41)\\
&\qquad\cdot \pi(\lambda|\phi=\phi_0)\,d\lambda,
\end{aligned}
$$

and the two densities clearly coincide.

(ii) Recall that $\pi^{\mathrm{M}}_{\mathcal{M}^*}(\lambda,\delta) = \pi(\lambda,\delta)$. Consider now $\pi^{\mathrm{M}}_{\mathcal{IM}^*}(\lambda)$: this is the marginal of $\pi_{\mathcal{IM}}(\lambda,\delta)$; the latter, however, coincides with the marginal $\pi(\lambda,\delta)$ under the prior $\pi(\lambda,\delta,\phi)$ by definition of integrated model. We therefore obtain $\pi^{\mathrm{M}}_{\mathcal{IM}^*}(\lambda) = \pi_{\mathcal{IM}}(\lambda) = \pi(\lambda)$. From (38) we get

$$
f^{\mathrm{M}}_{\mathcal{M}^*}(y) = \int\int f(y|\lambda,\delta,\phi=\phi_0)\pi(\delta|\lambda)\pi(\lambda)\,d\delta\,d\lambda,
$$

while from (39) we get

$$
\begin{aligned}
&f^{\mathrm{M}}_{\mathcal{IM}^*}(y)\\
&= \int\left\{\int f(y|\lambda,\delta,\phi=\phi_0)\pi(\delta|\lambda,\phi=\phi_0)\,d\delta\right\}\\
&\qquad\cdot \pi(\lambda)\,d\lambda.
\end{aligned}
$$

Inspection of $f^{\mathrm{M}}_{\mathcal{M}^*}(y)$ and $f^{\mathrm{M}}_{\mathcal{IM}^*}(y)$ reveals that if δ is conditionally independent of ϕ given λ, the two densities are equal.

(iii) Recall that

$$
\pi^{\mathrm{JC}}_{\mathcal{M}^*}(\lambda,\delta) \propto \pi(\lambda,\delta|\phi=\phi_0)\frac{j_{\mathcal{M}^*}(\lambda,\delta)}{j_{\mathcal{M}}(\lambda,\delta,\phi_0)},
$$

where the j-functions are the Jeffreys priors. Passing to the integrated model we therefore obtain

$$
\pi^{\mathrm{JC}}_{\mathcal{IM}^*}(\lambda) \propto \pi_{\mathcal{IM}}(\lambda|\phi=\phi_0)\frac{j_{\mathcal{IM}^*}(\lambda)}{j_{\mathcal{IM}}(\lambda,\phi_0)}.
$$

Let

$$
h(\lambda,\delta) = \frac{j_{\mathcal{M}^*}(\lambda,\delta)}{j_{\mathcal{M}}(\lambda,\delta,\phi_0)}, \qquad g(\lambda) = \frac{j_{\mathcal{IM}^*}(\lambda)}{j_{\mathcal{IM}}(\lambda,\phi_0)}.
$$

Clearly, if $h(\lambda,\delta) \propto g(\lambda)$, then $f^{\mathrm{JC}}_{\mathcal{M}^*}(y)$ and $f^{\mathrm{JC}}_{\mathcal{IM}^*}(y)$ have a representation as in (40), respectively (41), with the integrand in each case multiplied by $g(\lambda)$, and therefore they must coincide. □

PROOF OF PROPOSITION 2. Start with the M procedure. Notice that $\pi^{**}_{\mathcal{M}}(\lambda) = \pi(\lambda)$. On the other hand $\pi^{**}_{\mathcal{M}^*}(\lambda) = \pi^*(\lambda)$, where $\pi^*(\lambda)$ is the marginal prior on λ under $\pi^*(\lambda,\phi_2)$; but the latter is under M equal to $\pi(\lambda,\phi_2)$, whence $\pi^{**}_{\mathcal{M}^*}(\lambda) = \pi(\lambda)$, thus establishing the result.

Consider now the UC procedure. We have $\pi^{**}_{\mathcal{M}}(\lambda) = \pi(\lambda|\phi_1 = \phi_1^0, \phi_2 = \phi_2^0)$. On the other hand

$$\begin{aligned}\pi^{**}_{\mathcal{M}^*}(\lambda) &= \pi^*(\lambda|\phi_2 = \phi_2^0) = \frac{\pi^*(\lambda, \phi_2 = \phi_2^0)}{\pi^*(\phi_2 = \phi_2^0)} \\ &= \frac{\pi(\lambda, \phi_2 = \phi_2^0|\phi_1 = \phi_1^0)}{\pi(\phi_2 = \phi_2^0|\phi_1 = \phi_1^0)} \\ &= \pi(\lambda|\phi_1 = \phi_1^0, \phi_2 = \phi_2^0),\end{aligned}$$

which establishes the result.

Finally consider the JC procedure. We have

$$\pi^{**}_{\mathcal{M}}(\lambda) \propto \pi(\lambda|\phi_1 = \phi_1^0, \phi_2 = \phi_2^0)\frac{j_{\mathcal{M}^{**}}(\lambda)}{j(\lambda, \phi_1^0, \phi_2^0)}. \tag{42}$$

On the other hand

$$\pi^{**}_{\mathcal{M}^*}(\lambda) \propto \pi^*_{\mathcal{M}}(\lambda|\phi_2 = \phi_2^0)\frac{j_{\mathcal{M}^{**}}(\lambda)}{j_{\mathcal{M}^*}(\lambda, \phi_2^0)}, \tag{43}$$

where $\pi^*_{\mathcal{M}}(\lambda|\phi_2 = \phi_2^0)$ is proportional to the JC prior under the $\mathcal{M}^*$ model, evaluated at $(\phi_2 = \phi_2^0)$, namely $\pi^*_{\mathcal{M}}(\lambda, \phi_2 = \phi_2^0)$, where

$$\pi^*_{\mathcal{M}}(\lambda, \phi_2) \propto \pi(\lambda, \phi_2|\phi_1 = \phi_1^0)\frac{j_{\mathcal{M}^*}(\lambda, \phi_2)}{j(\lambda, \phi_1^0, \phi_2^0)}.$$

Substituting into (43), one obtains (42). □

LEMMA A.1. *Assume* $(\beta, \sigma^2) \sim \text{NIGa}(b, g(X^T \cdot X)^{-1}, d, a)$ *and set* $R_k(\beta, \sigma^2) = (1 + Q_k(\beta)/\sigma^2)^{-1}$, *with* $Q_k(\beta) = \beta^T X^T (I - P_k) X\beta/n$. *Then, given* σ^2, $Q_k(\beta)/\sigma^2 \sim (g/n)\chi^2_{p-k}(\delta)$, *with* $\delta = nQ_k(b)/(g\sigma^2)$, *where* $\chi^2_{p-k}(\delta)$ *is a chi-squared distribution with* $(p - k)$ *degrees of freedom and noncentrality parameter* δ. *As a consequence*

$$\begin{aligned}&E[R_k(\beta, \sigma^2)^{-1}] \\ &\quad = \left(1 + \frac{g}{n}(p - k) + Q_k(b)\frac{d}{a}\right),\end{aligned} \tag{44}$$

$$\begin{aligned}&\text{Var}[R_k(\beta, \sigma^2)^{-1}] \\ &\quad = \frac{2d}{a}Q_k(b)\left[\frac{Q_k(b)}{a} + \frac{2g}{n}\right] + \frac{2g^2}{n^2}(p - k).\end{aligned} \tag{45}$$

Furthermore

(i) *if* $b_{\backslash k} = 0$, *then* $R_k(\beta, \sigma^2) = [1 + (g/n)W]^{-1}$ *with* W *distributed as a (central)* χ^2_{p-k}, *whence*

$$\begin{aligned}E[R_k(\beta, \sigma^2)] &= \left(\frac{2g}{n}\right)^{-(p-k)/2}\exp\left(\frac{n}{2g}\right) \\ &\quad \cdot \Gamma\left(1 - \frac{p - k}{2}; \frac{n}{2g}\right),\end{aligned} \tag{46}$$

where $\Gamma(\alpha, z) = \int_z^\infty \exp(-t)t^{\alpha-1}\,dt$ *is the incomplete gamma function.*

(ii) *If* $b_{\backslash k} \neq 0$, *then the first-order approximation of* $E[R_k(\beta, \sigma^2)]$ *given by the delta method is*

$$\begin{aligned}E[R_k(\beta, \sigma^2)] &\approx \frac{1}{E[R_k(\beta, \sigma^2)^{-1}]} \\ &= \left[1 + \frac{g}{n}(p - k) + Q_k(b)\frac{d}{a}\right]^{-1}.\end{aligned} \tag{47}$$

PROOF. First of all notice that because $X = [X_k \vdots X_{\backslash k}]$ and $\beta = [\beta_k^T \vdots \beta_{\backslash k}^T]^T$, we have $Q_k(\beta) = \beta^T X^T M_k X\beta/n = \beta_{\backslash k}^T X_{\backslash k}^T M_k X_{\backslash k}\beta_{\backslash k}/n$. Now $\beta_{\backslash k}|\sigma^2$ is distributed according to a $\text{N}(b_{\backslash k}, g\sigma^2\Sigma_{\backslash k})$ with $\Sigma_{\backslash k} = (X_{\backslash k}^T M_k X_{\backslash k})^{-1}$ (see Searle, 1982, Section 10.5), and consequently $(n/g)Q_k(\beta)/\sigma^2$ given σ^2 is distributed according to a $\chi^2_{p-k}(\delta)$ distribution, where $p - k$ are the degrees of freedom and $\delta = (n/g)Q_k(b)/\sigma^2$ is the noncentrality parameter (see Muirhead, 1982, page 26). Now recalling that the expected value and variance of a $\chi^2_{p-k}(\delta)$ distribution are respectively $p - k + \delta$ and $2(p - k) + 4\delta$, (44) follows immediately from $E[R_k(\beta, \sigma^2)^{-1}] = E^{\sigma^2}\{1 + E^{\beta|\sigma^2}[Q_k(\beta)/\sigma^2]\}$, and $E(1/\sigma^2) = d/a$.

Similarly (45) follows from

$$\begin{aligned}\text{Var}[R_k(\beta, \sigma^2)^{-1}] &= \text{Var}[Q_k(\beta)/\sigma^2] \\ &= \text{Var}^{\sigma^2}\left[E^{\beta|\sigma^2}\left(\frac{Q_k(\beta)}{\sigma^2}\right)\right] \\ &\quad + E^{\sigma^2}\left[\text{Var}^{\beta|\sigma^2}\left(\frac{Q_k(\beta)}{\sigma^2}\right)\right] \\ &= \frac{g^2}{n^2}\left\{\text{Var}^{\sigma^2}\left[p - k + \frac{n}{g\sigma^2}Q_k(b)\right]\right. \\ &\quad \left. + E^{\sigma^2}\left[2(p - k) + 4\frac{nQ_k(b)}{g\sigma^2}\right]\right\}\end{aligned}$$

and $\text{Var}(1/\sigma^2) = 2d/a^2$.

(i) If $b_{\backslash k} = 0$, then $\delta = 0$, so that $W = (n/g)Q_k(\beta)/\sigma^2$ is distributed as a (central) χ^2_{p-k}. Thus $E[R_k(\beta, \sigma^2)] = E^W[(1 + (g/n)W)^{-1}]$ whose analytical expression is given in (46).

(ii) If $b_{\backslash k} \neq 0$, writing $E[R_k(\beta, \sigma^2)] = E[1/R_k(\beta, \sigma^2)^{-1}]$ and recalling that the first-order approximation gives $E(1/W) \approx 1/(E(W))$ for an arbitrary random variable W, we obtain (47). □

PROOF OF PROPOSITION 5. The $\text{NIGa}(b_k, g_k(X_k^T X_k)^{-1}, d_k, a_k)$ distribution on (β_k^*, σ_k^2) can be

written as

$$\pi(\beta_k^*, \sigma_k^2) \propto \exp\Big\{-\frac{1}{2\sigma_k^2}a + b^T X_k^T X_k \frac{\beta_k^*}{g\sigma_k^2} - \frac{1}{2g}\frac{\beta_k^{*T} X_k^T X_k \beta_k^*}{\sigma_k^2} + \frac{d+p+2}{2}\log\Big(\frac{1}{\sigma_k^2}\Big)\Big\},$$

thus it belongs to the exponential family with "canonical statistics" given by $1/\sigma_k^2$, β_k^*/σ_k^2, $\beta_k^{*T} X_k^T X_k \beta_k^*/\sigma_k^2$ and $\log(1/\sigma_k^2)$. Applying Theorem 1 of Consonni, Gutiérrez-Peña and Veronese (2007), it follows that the KL-divergence between π^{KL} and a NIGa$(b_k, g_k(X_k^T \cdot X_k)^{-1}, d_k, a_k)$ distribution is minimized for values $b_k^{\mathrm{KL}}, g_k^{\mathrm{KL}}, d_k^{\mathrm{KL}}$ and a_k^{KL} which are a solution of the following system:

$$E^{\mathrm{KL}}(1/\sigma_k^2) = E^{\mathrm{NIGa}}(1/\sigma_k^2), \tag{48}$$

$$E^{\mathrm{KL}}(\beta_k^*/\sigma_k^2) = E^{\mathrm{NIGa}}(\beta_k^*/\sigma_k^2), \tag{49}$$

$$E^{\mathrm{KL}}(\beta_k^{*T} X_k^T X_k \beta_k^*/\sigma_k^2) = E^{\mathrm{NIGa}}(\beta_k^{*T} X_k^T X_k \beta_k^*/\sigma_k^2), \tag{50}$$

$$E^{\mathrm{KL}}(\log(1/\sigma_k^2)) = E^{\mathrm{NIGa}}(\log(1/\sigma_k^2)), \tag{51}$$

where E^{KL} denotes expectation w.r.t. the KL-projection prior induced by the NIGa$(\beta, \sigma^2; b, g(X^T X)^{-1}, d, a)$, while E^{NIGa} denotes expectation w.r.t. the NIGa$(\beta_k^*, \sigma_k^2; b_k^{\mathrm{KL}}, g_k^{\mathrm{KL}}(X_k^T X_k)^{-1}, d_k^{\mathrm{KL}}, a_k^{\mathrm{KL}})$. Recalling (17) and (18), that is, $\beta_k^\perp = (X_k^T X_k)^{-1} X_k^T X\beta$, $\sigma_k^{2\perp} = \sigma^2 + Q_k(\beta)$, we can compute the terms involving E^{KL} in the previous equations substituting (β_k^*, σ_k^2) with the corresponding expression of $(\beta_k^\perp, \sigma_k^{2\perp})$ and using the prior $\pi(\beta, \sigma^2)$.

First of all recall that if Y is a normal vector with variance matrix Σ, then $Y^T AY$ and CY are stochastically independent if and only if $C\Sigma A = 0$; similarly $Y^T AY$ and $Y^T DY$ are stochastically independent if and only if $A\Sigma D = 0$ (with A, C and D being suitable matrices). It follows that under π and given σ^2, $Q_k(\beta)$ and $\beta_k^\perp$ as well as $Q_k(\beta)$ and $\beta_k^{\perp T} X_k^T X_k \beta_k^\perp$ are independent; the latter implies that also $\beta_k^{\perp T} X_k^T X_k \beta_k^\perp$ and $\sigma_k^{2\perp}$ are independent, given σ^2. The proof is a straightforward calculation.

Consider now (49). The left-hand side is equal to

$$E\Big(\frac{\beta_k^\perp}{\sigma_k^{2\perp}}\Big) = E^{\sigma^2}\Big\{E^{\beta|\sigma^2}\Big[\frac{1}{\sigma^2 + Q_k(\beta)}\Big] \cdot E^{\beta|\sigma^2}[(X_k^T X_k)^{-1} X_k^T X\beta]\Big\} = (X_k^T X_k)^{-1} X_k^T Xb E(1/\sigma_k^2),$$

while the right-hand side is equal to $E^{\mathrm{NIGa}}(\beta_k^*/\sigma_k^2) = b_k^{\mathrm{KL}} E^{\mathrm{NIGa}}(1/\sigma_k^2)$. Using (48) it follows that

$$b_k^{\mathrm{KL}} = (X_k^T X_k)^{-1} X_k^T Xb. \tag{52}$$

Consider (50). First of all notice that, using (17), $\beta_k^{\perp T} X^T X \beta_k^\perp = \beta^T X^T P_k X\beta$. Thus the left-hand side can be written, recalling the independence of $\sigma_k^{2\perp}$ and $\beta_k^{\perp T} X^T X\beta_k^\perp$, given σ^2, as

$$\begin{aligned} &E^{\sigma^2}[E^{\beta|\sigma^2}(1/\sigma_k^{2\perp}) E^{\beta|\sigma^2}(\beta^T X^T P_k X\beta)] \\ &\quad = E^{\sigma^2}\big\{E^{\beta|\sigma^2}\big[(1/\sigma_k^{2\perp}) \cdot \big(\mathrm{tr}(\sigma^2 g P_k P) + b^T X^T P_k Xb\big)\big]\big\} \\ &\quad = E^{\sigma^2}[(kg\sigma^2 + b^T X^T P_k Xb) E^{\beta|\sigma^2}(1/\sigma_k^{2\perp})] \\ &\quad = kgE[R_k(\beta, \sigma^2)] + b^T X^T P_k Xb E(1/\sigma_k^{2\perp}), \end{aligned}$$

where $R_k(\beta, \sigma^2) = [1 + Q_k(\beta)/\sigma^2]^{-1}$.

The right-hand side is equal to

$$\begin{aligned} &E^{\sigma_k^2}[(1/\sigma_k^2) E^{\beta_k^*|\sigma_k^2}(\beta_k^{*T} X_k^T X_k \beta_k^*)] \\ &\quad = E^{\sigma_k^2}\{(1/\sigma_k^2)[\mathrm{tr}(\sigma_k^2 g_k^{\mathrm{KL}}(X_k^T X_k)^{-1}(X_k^T X_k)) + b_k^{\mathrm{KL}T}(X_k^T X_k) b_k^{\mathrm{KL}}]\} \\ &\quad = kg_k^{\mathrm{KL}} + b^T X^T P_k Xb E^{\sigma_k^2}(1/\sigma_k^2), \end{aligned}$$

substituting the expression of b_k^{KL} given in (52). Equating the left- and right-hand sides and using (48) we obtain

$$g_k^{\mathrm{KL}} = gE[R_k(\beta, \sigma^2)]. \tag{53}$$

Consider (51). The left-hand side can be written as $E[\log(1/\sigma^2)] + E[\log(R_k(\beta, \sigma^2))]$ with $E[\log(1/\sigma^2)] = \Psi(d/2) - \log(a/2)$, where $\Psi(\alpha) = \frac{\partial}{\partial\alpha} \cdot \log(\Gamma(\alpha))$ is the digamma function. The right-hand side is equal to $\Psi(d_k^{\mathrm{KL}}/2) - \log(a_k^{\mathrm{KL}}/2)$ and thus we obtain

$$\Psi(d/2) - \log(a/2) + E[\log(R_k(\beta, \sigma^2)] = \Psi(d_k^{\mathrm{KL}}/2) - \log(a_k^{\mathrm{KL}}/2). \tag{54}$$

Assume now that $b_{\setminus k} = 0$ and consider last (48). First notice that the left-hand side can be written as $E[R_k(\beta, \sigma^2)/\sigma^2]$, while the right-hand side is equal to $d_k^{\mathrm{KL}}/a_k^{\mathrm{KL}}$. Since, from Lemma A.1, $R_k(\beta, \sigma^2)$ is independent of σ^2 when $b_{\setminus k} = 0$, (48) becomes

$$E(1/\sigma^2)E[R_k(\beta, \sigma^2)] = d_k^{\mathrm{KL}}/a_k^{\mathrm{KL}}, \tag{55}$$

which implies

$$a_k^{\mathrm{KL}} = d_k^{\mathrm{KL}} \frac{a}{d} \frac{1}{E[R_k(\beta, \sigma^2)]}. \tag{56}$$

Substituting (56) into (54) we obtain

$$\begin{aligned} &\Psi(d_k^{\mathrm{KL}}/2) - \log(d_k^{\mathrm{KL}}/2) \\ &= \Psi(d/2) - \log(d/2) + E\{\log[R_k(\beta, \sigma^2)]\} \\ &\quad - \log\{E[R_k(\beta, \sigma^2)]\}. \end{aligned} \tag{57}$$

Consider now the case $b_{\setminus k} \neq 0$. In order to obtain an explicit expression of (53), we use the approximation of $E[R_k(\beta, \sigma^2)]$ given in (47), so that

$$\begin{aligned} g_k^{\mathrm{KL}} &\approx \frac{g}{E[R_k^{-1}(\beta, \sigma^2)]} \\ &= \frac{g}{[1 + g/n(p-k) + Q_k(b)d/a]}. \end{aligned} \tag{58}$$

Furthermore, we can still use (55) as an approximation of (48) to the first order. Thus we have

$$\begin{aligned} a_k^{\mathrm{KL}} &\approx d_k^{\mathrm{KL}} \frac{a}{d} \frac{1}{E[R_k(\beta, \sigma^2)]} \\ &\approx d_k^{\mathrm{KL}} \frac{a}{d} E[R_k^{-1}(\beta, \sigma^2)] \\ &= d_k^{\mathrm{KL}} \frac{a}{d} \left[1 + \frac{g}{n}(p-k) + Q_k(b)\frac{d}{a}\right], \end{aligned} \tag{59}$$

using (47).

Using the first approximation of (59), formula (57) still holds in an approximate way.

Finally (57) reduces to

$$\begin{aligned} &\Psi(d_k^{\mathrm{KL}}/2) - \log(d_k^{\mathrm{KL}}/2) \\ &\quad \approx \Psi(d/2) - \log(d/2) - \frac{1}{2}\frac{\mathrm{Var}([R_k(\beta, \sigma^2)])}{E[R_k(\beta, \sigma^2)]^2}, \end{aligned}$$

using the further second-order approximation $E[\log(U)] \approx \log[E(U)] - (1/2)\,\mathrm{Var}(U)/[E(U)]^2$, for a positive random variable U.

Since $\mathrm{Var}(U) = \mathrm{Var}(1/U^{-1}) \approx [1/E(U^{-1})]^4 \cdot \mathrm{Var}(U^{-1})$ and $E(U) = E(1/U^{-1}) \approx 1/E(U^{-1})$ we conclude

$$\begin{aligned} &\Psi(d_k^{\mathrm{KL}}/2) - \log(d_k^{\mathrm{KL}}/2) \\ &\quad \approx \Psi(d/2) - \log(d/2) - \frac{1}{2}\frac{\mathrm{Var}([R_k^{-1}(\beta, \sigma^2)])}{E[R_k^{-1}(\beta, \sigma^2)]^2} \end{aligned}$$

with $E[R_k^{-1}\beta, \sigma^2)]$ and $\mathrm{Var}[R_k^{-1}(\beta, \sigma^2)]$ given in (44) and (45). □

ACKNOWLEDGMENTS

GC's research was supported in part by MIUR, Rome (PRIN 2003138887 and PRIN 2005132307) and the University of Pavia. PV's research was supported in part by MIUR, Rome (PRIN 2003138887 and PRIN 2005132307) and by L. Bocconi University. Part of this work was written while the authors visited Université Paris-Sud, Orsay, France. They are grateful to Gilles Celeux, Jean Michel Marin and Christian Robert for helpful discussions. Useful comments on some of the ideas presented in this paper were also provided by Eduardo Gutiérrez-Peña and Manuel Mendoza. The support of the Executive Editor and the reviewers' comments are gratefully acknowledged.

REFERENCES

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

BAYARRI, M. J. and GARCÍA-DONATO, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika* **94** 135–152. MR2367828

BERGER, J. O. and PERICCHI, L. R. (1996a). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065

BERGER, J. O. and PERICCHI, L. R. (1996b). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 25–44. Oxford Univ. Press. MR1425398

BERGER, J. O. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. Inst. Math. Statist., Beachwood, OH. MR2000753

BERGER, J. O. and PERICCHI, L. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.* **32** 841–869. MR2065191

BERNARDO, J. M. and RUEDA, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70** 351–372.

CASELLA, G. and MORENO, E. (2005). Intrinsic meta-analysis of contingency tables. *Stat. Med.* **24** 583–604. MR2134527

CASELLA, G. and MORENO, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101** 157–167. MR2268035

CASELLA, G. and MORENO, E. (2007). Assessing robustness of intrinsic tests of independence in two-way contingency tables. Technical report, Dept. Statistics, Univ. Florida.

CELEUX, G., MARIN, J. M. and ROBERT, C. P. (2006). Sélection bayésienne de variables en régression linéaire. *J. Soc. Française de Statistique* **147** 59–79.

CLYDE, M., DESIMONE, H. and PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91** 1197–1208.

CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

CONSONNI, G., GUTIÉRREZ-PEÑA, E. and VERONESE, P. (2007). Compatible priors for Bayesian model comparison with an application to the Hardy–Weinberg equilibrium model. *Test*. To appear. DOI 10.1007/s11749-007-0057-7. MR1667008

CONSONNI, G. and LA ROCCA, L. (2008). Tests based on intrinsic priors for the equality of two correlated proportions. *J. Amer. Statist. Assoc.* To appear.

CRIPPS, E., CARTER, C. and KOHN, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Handbook of Statistics* **25**. *Bayesian Thinking Modeling and Computation* (D. K. Dey and C. R. Rao, eds.) 519–552. North-Holland, New York.

DAVISON, A. C. (2003). *Statistical Models*. Cambridge Univ. Press, Cambridge. MR1998913

DAWID, A. P. and LAURITZEN, S. L. (2001). Compatible prior distributions. In *Bayesian Methods With Applications to Science, Policy and Official Statistics* (E. George, ed.) 109–118. Monographs of Official Statistics, Luxembourg.

DEY, D. K. and RAO, C. R., eds. (2005). *Handbook of Statistics* **25**. *Bayesian Thinking Modeling and Computation*. North-Holland, New York.

DICKEY, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42** 204–223. MR0309225

DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley, New York. MR0610978

DUPUIS, J. A. and ROBERT, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *J. Statist. Plann. Inference* **111** 77–94. MR1955873

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410

FREEDMAN, D. A. (1983). A note on screening regression equations. *Amer. Statist.* **37** 152–155. MR0702208

GARTHWAITE, P. H. and DICKEY, J. M. (1996). Quantifying and using expert opinion for variable-selection problems in regression. *Chemometrics and Intelligent Laboratory Systems* **35** 1–26.

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11. MR1627258

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. I. (1999). Bayesian model selection. In *Encyclopedia of Statistical Sciences, Update* **3** (S. Kotz, C. Read and D. Banks, eds.) 39–46. Wiley, New York.

GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95** 1304–1308. MR1825282

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GOUTIS, C. and ROBERT, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika* **85** 29–37. MR1627250

IBRAHIM, J. G. (1997). On properties of predictive priors in linear models. *Amer. Statist.* **51** 333–337. MR1484784

LAUD, P. W. and IBRAHIM, J. G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57** 247–262. MR1325389

LAUD, P. W. and IBRAHIM, J. G. (1996). Predictive specification of prior model probabilities in variable selection. *Biometrika* **83** 267–274. MR1439783

LIANG, R. F., PAULO, R., MOLINA, G., CLYDE, M. and BERGER, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423.

O'HAGAN, A. and FORSTER, J. (2004). *Kendall's Advanced Theory of Statistics* **2b**. *Bayesian Inference*, 2nd ed. Edward Arnold, London.

MARRIOTT, J. M., SPENCER, N. M. and PETTITT, N. (2001). A Bayesian approach to selecting covariates for prediction. *Scand. J. Statist.* **28** 87–97. MR1844350

MCCULLOCH, R. E. and ROSSI, P. E. (1992). Bayes factor for nonlinear hypotheses and likelihood distributions. *Biometrika* **79** 663–676. MR1209468

MORENO, E. and GIRON, E. J. (2007). Comparison of Bayesian objective procedures for variable selection in linear regression. *Test*. To appear. DOI 10.1007/s11749-006-0039-1.

MORRIS, C. M. (1987). Comment on: "Reconciling Bayesian and frequentist evidence in the one-sided testing problem," by G. Casella and R. L. Berger and "Testing a point null hypothesis: The irreconcilability of P values and evidence," by J. O. Berger and T. Sellke. *J. Amer. Statist. Assoc.* **82** 131–135.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

NEAL, R. (2001). Transferring prior information between models using imaginary data. Technical Report No. 0108, Dept. Statistics, Univ. Toronto.

NOTT, D. J. and GREEN, P. J. (2004). Bayesian variable selection and the Swendsen–Wang algorithm. *J. Comput. Graph. Statist.* **13** 141–157. MR2044875

PÉREZ, J. M. and BERGER, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* **89** 491–511. MR1929158

PERICCHI, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In *Handbook of Statistics* **25**. *Bayesian Thinking Modeling and Computation* (D. K. Dey and C. R. Rao, eds.) 115–149. North-Holland, New York.

POIRIER, D. J. (1985). Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 711–722. North-Holland, Amsterdam. MR0862514

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92** 179–191. MR1436107

RAO, R. and TOUTEMBURG, H. (1999). *Linear Models: Least Squares and Alternatives.* Springer, New York. MR1707290

ROBERT, C. P. (2001). *The Bayesian Choice*, 2nd ed. Springer, New York. MR1835885

ROVERATO, A. and CONSONNI, G. (2004). Compatible prior distributions for directed acyclic graph models. *J. Roy. Statist. Soc. Ser. B* **66** 47–61. MR2035758

SEARLE, S. R. (1982). *Matrix Algebra Useful for Statistics.* Wiley, Chichester. MR0670947

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

VIELE, K. and SRINIVASAN, C. (2000). Parsimonious estimation of multiplicative interaction in analysis of variance using Kullback–Leibler information. *J. Statist. Plann. Inference* **84** 201–219. MR1748194

YUAN, M. and LIN, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100** 1215–1225. MR2236436

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester. MR1112133

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* (P. K. Goel and A. Zellner, eds.) 233–243. North-Holland, Amsterdam. MR0881437

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratio for selected regression hypotheses. In *Bayesian Statistics* **1** (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. Valencia Univ. Press, Spain. MR0638871