Bayesian Analysis (2018) **13**, Number 2, pp. 627–679

Prior Distributions for Objective Bayesian Analysis

Guido Consonni*, Dimitris Fouskakis†, Brunero Liseo‡, and Ioannis Ntzoufras§

Abstract. We provide a review of prior distributions for objective Bayesian analysis. We start by examining some foundational issues and then organize our exposition into priors for: i) estimation or prediction; ii) model selection; iii) high-dimensional models. With regard to i), we present some basic notions, and then move to more recent contributions on discrete parameter space, hierarchical models, nonparametric models, and penalizing complexity priors. Point ii) is the focus of this paper: it discusses principles for objective Bayesian model comparison, and singles out some major concepts for building priors, which are subsequently illustrated in some detail for the classic problem of variable selection in normal linear models. We also present some recent contributions in the area of objective priors on model space. With regard to point iii) we only provide a short summary of some default priors for high-dimensional models, a rapidly growing area of research.

Keywords: objective Bayes, model comparison, criteria for model choice, noninformative prior, reference prior, variable selection, high-dimensional model.

MSC 2010 subject classifications: Primary 62F15, 62-02; secondary 62J05, 62A01.

1 Objective Bayes methods

In many situations a researcher is not able to express his/her prior opinion into a prior distribution. This may happen, for example, in complex applications, where the parameter space has large dimension and a genuine elicitation of the prior dependence structure among the parameters can be out of reach. In other cases, a very limited knowledge of the problem at hand is available, and one would like to encapsulate prior ignorance into a probability distribution. In both cases, it would be helpful to use a *noninformative* prior in order to make Bayes' theorem work, without introducing subjective inputs into the analysis. This has been, in the last decades, like a search of the "philosopher's stone" for the Bayesian community. However, using Savage's words, as reported in Kass and Wasserman (1995), ... *it has proved impossible to give a precise definition of the tempting expression "know nothing."* The focus subsequently moved to the search of priors with a minimal impact on the corresponding posterior analysis,

*Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Milan, Italy, guido.consonni@unicatt.it

†Department of Mathematics, National Technical University of Athens, Athens, Greece, fouskakis@math.ntua.gr

‡Dipartimento di Metodi e Modelli per il Territorio, l'Economia e la Finanza, Sapienza Università di Roma, Rome, Italy, brunero.liseo@uniroma1.it

§Computational and Bayesian Statistics Lab, Department of Statistics, Athens University of Economics and Business, Athens, Greece, ntzoufras@aueb.gr

 https://doi.org/10.1214/18-BA1103

an important motivation for scientific communication. These priors has been named in many different, sometimes misleading, ways, from vague to objective, from default to noninformative or reference. Each of these terms describes a different aspect of the same problem, and Objective Bayes (OB, hereafter) has emerged as a broad term which tries to include all these strands. It is therefore not surprising that Berger (2006) warns his readers upfront that "there is no unanimity as to the definition of OB analysis, not even on its goals". We believe that after more than ten years this conclusion is still fair.

If we disregard goals, and rather focus on implementation issues, a commonly held view is that an OB method should only use the information contained in the statistical model, and no other external information (Bayarri and García-Donato, 2007); see, however, Leisen et al. (2017) for a radically different view. The above view of "objectivity" presupposes that a model has a different theoretical status relative to the prior: it is the latter which encapsulates the subjective uncertainty of the researcher, while the model is less debatable, possibly because it can usually be tested through data. Another justification is offered by the subjective-predictive approach to inference, as explicated in de Finetti's theory; see Bernardo and Smith (1994, Ch. 4) for an accessible introduction. At first sight this might look surprising, because in the celebrated representation theorem for exchangeable random variables both the model and the prior originate from a unique (subjective) predictive distribution, so that they seem to stand on an equal footing. Dawid (1982) however, in an insightful paper, clarifies how a philosophical distinction between model and prior can be drawn, even within the subjective paradigm, with the former representing a common "intersubjective" component, and the latter being specific to each individual. As an illustration, consider a sequence of 0/1 random variables. While each subject may have a distinct predictive opinion on sequences of such random variables, the very fact that each predictive distribution satisfies the condition of exchangeability implies that all subjects will share the same statistical model (product of i.i.d. Bernoulli laws in this case), while their disagreement will be confined to the distribution of the random probability of success, indexing the statistical model. Representation theorems for exchangeable processes beyond the 0/1 case are of course available, with a similar pattern emerging, although some further structural assumptions are needed to nail down a common intersubjective statistical model among different subjects; see again Bernardo and Smith (1994, Ch. 4).

Even if we take for granted a given statistical model, the actual implementation of any OB principle is likely to incorporate, besides the statistical model, some additional context information. This happens for instance in the construction of reference priors (Bernardo, 1979) for a parameter-vector, where the notion of inferential importance of the component parameters is crucial for a correct application of the methodology; see also Section 2. Another notable case is represented by the inferential "goal" of the analysis where the OB prior will be employed. We will argue below that a useful distinction is between priors for estimation (including prediction) and for model selection; again context matters.

In the end, our view of what constitutes an OB analysis is unavoidably pragmatic. First of all, we firmly believe that OB and subjective Bayesian analysis should complement each other, the former being helpful in particular scenarios (prior elicitation is too

hard, or time consuming, or for reference analysis in scientific reporting). Subjective analysis is still a great resource, especially in applications where information about context is available and can be meaningfully incorporated. Secondly, the quality of an OB method should be judged both in terms of its theoretical foundations, and on the correspondence it exhibits to actual Bayesian procedures; see Berger and Pericchi (2001).

A communication problem with the OB approach is that the word "objective" is loaded with many interpretations and expectations. This has led Gelman and Hennig (2017) to propose a radically different approach to the subjective *versus* objective debate in Statistics, which actually transcends the Bayesian approach. They argue that "the words 'objectivity' and 'subjectivity' in statistical discourses are used in a mostly unhelpful way, and [...] propose to replace each of them with broader collections of attributes, with objectivity replaced by transparency, consensus, impartiality, and correspondence to observable reality, and subjectivity replaced by awareness of multiple perspectives and context dependence". The advantage of their reformulation is that the replacement terms do not oppose each other, but rather complement each other, not just from a practical viewpoint, but also from a conceptual one.

We will distinguish between priors for estimation (and prediction) purposes within a given model, and priors for model selection (or comparison), where a collection of models is entertained. This distinction however is currently challenged in the analysis of high-dimensional problems characterized by a huge number of parameters and models, where sparsity inducing priors are devised for the dual purpose of selection and estimation. In this review we will mostly focus on priors for model selection, and especially priors on the parameter space of each entertained model. One reason for this choice is that research on objective priors for estimation/prediction has a long tradition and, accordingly, it has received considerable attention over the past years; see in particular the excellent reviews by Kass and Wasserman (1995) and Ghosh (2011). On the other hand, the OB methodology for priors tailored to model selection started more recently, and its development and applications to various models have increased over the last few years, so that they could not be included in previous reviews such as Berger and Pericchi (1996), Berger and Pericchi (2001), and Pericchi (2005).

2 Prior distributions for estimation and prediction

"Noninformative prior" has been, for many years, the most common name for indicating any kind of prior which was proposed in an attempt to prepare the Bayesian omelette without breaking the Bayesian eggs (Savage, 1954); that is, to obtain probabilistic likelihood-based inferences without relying on informative prior distributions. For the sake of brevity, here we cannot review the long history of the selection of objective priors in Bayesian inference. The interested reader can refer to Kass and Wasserman (1996) and Ghosh (2011). Here we limit ourselves to list the most well-known existing methods and to discuss the most recent advances.

i. *Uniform prior.* Based on a somehow misinterpreted principle of indifference, one can use a prior for a scalar (continuous) parameter which assigns equal probabilities

to intervals having the same length. However a uniform prior is not invariant under re-parametrization and in many real cases there is no natural parametrization for a given model (Jaynes, 2003). In addition, a uniform prior on an unbounded parameter space is improper (i.e. its total mass is not finite). Then, there is no guarantee that the posterior will be proper and a case by case check must be considered.

ii. *Invariant prior*. The lack of invariance of the uniform prior has led many researchers to look for objective priors which are invariant under a certain class of transformations.

Let $(\mathcal{P}, \Theta)$ be a statistical model for the observation X, where $\mathcal{P}$ is the distribution model (a family of distributions), and Θ is the parameter associated to it. Let $Y = s(X)$ be a transformation, and suppose that the distribution model for Y is still $\mathcal{P}$, and denote with Λ the parameter. Notice that $\mathcal{P}$ is unchanged, and therefore we say that the model is invariant to the transformation $s(\cdot)$. If only $\mathcal{P}$ is allowed to inform our choice of the prior, then one should require that the prior for θ, π_θ, and that for λ, π_λ be such that $P^{\pi_\theta}\{\theta \in A\} = P^{\pi_\lambda}\{\lambda \in A\}$, for all sets A. This is named *context invariance* in Dawid (2006), and represents a very strong requirement because it means that it is only the structure of $\mathcal{P}$ that matters, irrespective of the context in which it is applied.

To exemplify, consider a model whose density is

$$f(x;\sigma) = \frac{1}{\sigma} g(x/\sigma), \quad \sigma > 0,$$

where $g(\cdot)$ is a density on $\mathbb{R}$. The distribution model is scale invariant because $X \sim f(x;\sigma)$ implies that $Y = cX \sim f(y;c\sigma)$, for all $c > 0$. We can imagine X being the price of a commodity measured in \$, and Y the corresponding price in Japanese yen. The scale invariance requirement for a prior π on σ leads to

$$\int_A \pi(\sigma)d\sigma = \int_{c^{-1}A} \pi(\sigma)d\sigma = \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma, \quad \text{for all measurable sets } A,$$

whence $\pi(\sigma) = \pi(c^{-1}\sigma)c^{-1}$ for all σ. Setting $\sigma = c$, and noting that the equality must hold for all $c > 0$, one concludes that the only measure which satisfies the requirement is $\pi(\sigma) \propto \sigma^{-1}$ which is improper, although not uniform. It is important to note that this is the right Haar invariant measure on the group of scale transformations. A complete description of the uses of invariance in Bayesian analysis can be found in Berger (1985), Dawid (2006) and Robert (2007).

iii. *Matching prior*. The rationale behind this approach is that a noninformative prior should provide inferences which are similar to those obtained from a frequentist perspective, for example in terms of credible versus confidence intervals. In this perspective, a probability matching prior is a prior distribution under which the posterior probabilities of certain regions coincide with their frequentist coverage probabilities, either exactly or approximately; see Datta and Mukerjee (2004) for details.

iv. *Maximum entropy prior* (Jaynes, 2003). This approach selects the prior which maximizes the entropy over a class of priors satisfying some basic restrictions. In the continuous case, the entropy of a distribution $\pi(\boldsymbol{\theta})$ is given by

$$\text{Ent}(\pi) = -\int_{\Theta} \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

and can be considered a measure of *un-informativeness* of $\pi(\cdot)$ for $\boldsymbol{\theta}$.

The maximum entropy prior approach is based on the following two steps. First, one chooses a large class Γ of potential prior distributions, characterized by a set of k constraints, usually in the form of quantiles or moments; the generic set of constraints can then be written as

$$\mathbb{E}\big(g_j(\boldsymbol{\theta})\big) = \mu_j, \qquad j = 1, \cdots k,$$

for suitable functions $g_j(\cdot)$. Next the maximum entropy prior is selected as any element in Γ maximizing $\text{Ent}(\pi)$.

v. *Jeffreys and reference prior*

In practical applications, however, at least before the advent of Markov Chains Monte Carlo (MCMC) methods, the vast majority of researchers used Jeffreys' prior (Jeffreys, 1961)

$$\pi^J(\boldsymbol{\theta}) \propto \det\left(\mathcal{I}(\boldsymbol{\theta})\right)^{1/2},$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix, whose generic element $\mathcal{I}_{ij}(\boldsymbol{\theta})$ – under very general conditions – and assuming a continuous parameter space, is given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(Y|\boldsymbol{\theta})\right),$$

where $\mathbb{E}_{\boldsymbol{\theta}}$ denotes the expected value over the sampling space for a given value of the parameter $\boldsymbol{\theta}$, and Y is an observable random variable.

Besides being parametrization invariant, Jeffreys' prior enjoys many optimality properties in the absence of nuisance parameters. It maximizes the asymptotic divergence between the prior and the posterior for $\boldsymbol{\theta}$, under several different metrics. It is also a second order matching prior (Datta and Mukerjee, 2004) when $\boldsymbol{\theta}$ is a scalar.

Although the Jeffreys' prior is probably still the most popular objective prior method among practitioners, it has some potential drawbacks which is important to discuss. The Jeffreys' prior may be improper and there is no guarantee that the resulting posterior distribution will be proper for all possible data sets: interesting counterexamples may be found in Ye and Berger (1991) and Berger et al. (2001). Jeffreys himself, in his original proposal, developed the method for the case of a scalar parameter. In the multidimensional case, the use of $\pi^J(\boldsymbol{\theta})$ may lead to incoherence and paradoxes (Dawid et al., 1973).

Jeffreys also suggested to separately deal with location parameters. If $\boldsymbol{\theta} = (\boldsymbol{\phi}, \lambda)$, where $\boldsymbol{\phi}$ is a vector of location parameters, then the Jeffreys' proposal is to use a prior proportional to $(\det(\lambda))^{1/2}$, keeping $\boldsymbol{\phi}$ fixed. This prior is called "non-location Jeffreys' prior" in Kass and Wasserman (1996). Another popular variant of the Jeffreys' method is the so-called "independent Jeffreys priors", which are made of a product of conditional Jeffreys' priors, i.e., by computing the Jeffreys prior one pa-

rameter at a time with all other parameters considered to be fixed (Robert, 2014). This prior is not invariant with respect to parametrization.

Another serious drawback of the Jeffreys' method for selecting objective priors is that there is no guarantee of a "satisfactory" behavior when the parameter of interest is a low dimensional function $\psi(\boldsymbol{\theta})$ of the entire parameter vector $\boldsymbol{\theta}$. Here "satisfactory" means that, in repeated sampling, the use of the Jeffreys' prior should produce statistical procedures with good a frequentist performance; for an interesting and well-known counterexample, see for example, Robert (2007), pag.133. This point is important because it suggests a deeper conclusion: a "good" objective prior for a vector $\boldsymbol{\theta}$ may have an unsatisfactory performance with regard to a function of the parameter which is of interest. The problem of selecting an objective prior for a specific parameter of interest $\psi(\boldsymbol{\theta})$ in the presence of other nuisance parameters has been one of the main motivations for the development of the so-called *reference prior* method (Bernardo, 1979; Berger and Bernardo, 1992). The goal of the reference prior approach, introduced by Bernardo (1979), is to find a prior distribution which maximizes – over the sample space – a limiting version of the average divergence between the prior and the corresponding posterior for a specific quantity of interest $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$. The method has been refined and improved in a series of papers (Berger et al., 1989; Berger and Bernardo, 1989; Berger et al., 2012, 2015). The reference prior method has introduced two main innovations in OB thinking: *i*) the explicit use of the notion of information contained in a statistical experiment, measured in terms of the Shannon–Lindley relative entropy; *ii*) the necessity of declaring in advance an ordering of *inferential importance* among the parameters of the model. In fact, for a given statistical model, the reference prior for the parameter vector $\boldsymbol{\theta}$ may well depend on that ordering (Berger and Bernardo, 1992). This reinforces the point that OB methods are, in general, context-dependent. Berger et al. (2015) deeply discuss this issue, and argue that there are many situations where having a single, overall objective prior would be desirable. They also propose two methods for achieving this goal.

In the scalar case, under general conditions, the reference prior coincides with Jeffreys' prior, at least when the latter can be calculated. Reference priors show, in general, very good frequentist properties in terms of coverage probability of a Bayesian credible interval. Further details on the methods for constructing priors discussed so far may be found in Kass and Wasserman (1996) or Berger (2006).

The remaining part of this section is devoted to some more recent developments.

Discrete parameter space

When the support of some of the parameters is discrete, traditional OB methods, like Jeffreys' or reference priors, cannot be directly used since they are based on the Fisher information matrix which assumes differentiability with respect to the parameters. It is important to stress that here we are not considering the case when the parameter is a model index, as for instance when it identifies a subset of covariates in a variable selection problem: see Section 3 for more details. We rather consider cases where the parameter is discrete due to the structure of the statistical model. Important examples

include the number of degrees of freedom ν in a Student-t sampling model, the unknown population size N in a capture-recapture model, and change-point problems.

Berger et al. (2012) discuss in detail several methods to tackle the problem. In particular they propose to embed the discrete parameter into a continuous parameter space and then apply the usual reference methodology. However it is not always clear how to practically perform the embedding. Under particular circumstances, one could add a hierarchical level to the model depending on a continuous hyperparameter, say $\boldsymbol{\theta}$, then find a reference prior for $\boldsymbol{\theta}$ and use it to indirectly derive the prior for the discrete parameter.

Example. *The Hypergeometric model* (Berger et al., 2012). Write the sampling distribution for the observation R as

$$P\left(R = r|n, N, M\right) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}, \quad r = 0, 1, \ldots, M,$$

where $M \in \{0, 1, \ldots, N\}$ is the unknown parameter. If we assume that, given p, $M \sim \text{Bin}(N, p)$, it is easy to see that the marginal model is given by

$$Pr\left(R = r|n, N, p\right) = \binom{n}{r} p^r (1-p)^{n-r}.$$

The natural objective prior for p would be the Jeffreys prior, that is a Beta$(0.5, 0.5)$; the prior for M would then be given by

$$\begin{aligned}\pi(M|N) &= \int_0^1 \binom{N}{M} p^M (1-p)^{N-M} \frac{1}{\pi} p^{-0.5} (1-p)^{-0.5} dp \\ &= \frac{1}{\pi} \frac{\Gamma(M+0.5)}{\Gamma(M+1)} \frac{\Gamma(N-M+0.5)}{\Gamma(N-M+1)}.\end{aligned}$$

However, the above situation is not so common and other approaches are discussed in Berger et al. (2012), mainly based on asymptotic arguments.

A radically different approach is discussed in a series of papers by Villa and Walker (2014b, 2015b, 2015a), where the authors propose a general method for producing objective priors starting from the so called "self-information" loss combined with the notion of the Kullback–Leibler divergence between models. A measure of the information loss associated to an event E having probability $\pi(E)$ is called self-information loss. The most natural one is given by: $I(E) = \log(1/\pi(E)) = -\log \pi(E)$. Then, they state a version of Bayes' theorem in terms of losses. In this framework, the formal derivation of the prior distribution for $\boldsymbol{\theta}$ can be expressed as follows. Consider a discrete collection of models indexed by $\{\boldsymbol{\theta}, f(\cdot|\boldsymbol{\theta})\}$. The *worth* associated to a particular value of $\boldsymbol{\theta}$ is represented by the Kullback–Leibler divergence between the model indexed by $\boldsymbol{\theta}$ and its nearest neighbor. That is,

$$u(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}^* \neq \boldsymbol{\theta}} D_{KL}\big(f(\cdot|\boldsymbol{\theta}) \big|\big| f(\cdot|\boldsymbol{\theta}^*)\big),$$

where $D_{KL}(f_j||f_k) = \int f_j(\boldsymbol{y}) \log(f_j(\boldsymbol{y})/f_k(\boldsymbol{y}))d\boldsymbol{y}$. Then, the above quantity represents the negative of the information loss in keeping the value $\boldsymbol{\theta}$ in the parameter space. At $\boldsymbol{\theta}$, the information loss can be also measured in terms of self-information loss. By equating the two expressions, one can derive the objective prior for $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}) \propto \exp\left\{ \min_{\boldsymbol{\theta}^* \neq \boldsymbol{\theta} \in \Theta} D_{KL}\big(f(\cdot|\boldsymbol{\theta}) \big|\big| f(\cdot|\boldsymbol{\theta}^*)\big) \right\} - 1.$$

More specialized topics related to estimation in discrete parameter spaces are: change-point problems (Girón et al., 2007), exponential families restricted to a lattice (Choirat and Seri, 2012), the degrees of freedom ν of a Student t distribution (Villa and Walker, 2014b) where the new prior is compared with two versions of the Jeffreys' prior proposed in Fonseca et al. (2008), the estimation of the number of trials in binomial and capture-recapture experiments (Villa and Walker, 2014a), and for assessing objective prior probabilities in a model selection scenario (Villa and Walker, 2015b).

Hierarchical Normal Model

The hierarchical normal model is still a very useful and routinely applied model because of its flexibility and modularity. However the formal derivation of objective priors has proven to be highly challenging. The most basic situation, which we now discuss, has been considered by Berger and several co-authors in a series of papers (Berger and Strawderman, 1996; Berger et al., 2005; Sun et al., 2001).

Consider

$$\boldsymbol{Y}_i = \boldsymbol{B}_i\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, m, \tag{1}$$

with the $\boldsymbol{\varepsilon}_i$'s mutually independent with a $N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ distribution, with $\boldsymbol{\Sigma}_i$ known; for simplicity assume $\boldsymbol{B}_i = \boldsymbol{I}_k$, for all i's. Also, assume that

$$\boldsymbol{\theta}_i = \boldsymbol{z}_i\boldsymbol{\beta} + \boldsymbol{\tau}_i, \qquad i = 1, \ldots, m, \tag{2}$$

with $\boldsymbol{\tau}_i \sim N_k(\boldsymbol{0}, \boldsymbol{V})$. Here the issue is to find objective priors for $(\boldsymbol{\beta}, \boldsymbol{V})$ *with reasonably good properties*. This common situation is hardly manageable both from a classical and Empirical Bayes perspectives: even when $k = 1$ the marginal likelihood may provide estimates of V equal to zero! On the other hand, the usual Jeffreys' prior $\pi(V) \propto V^{-1}$ would give an improper posterior, and the problem is only hidden, not solved, if one uses a vague proper inverse gamma prior on V with very small values of the shape and the scale parameters. This issue is discussed in detail in Berger and Strawderman (1996). In general, when an improper prior produces an improper posterior, the use of a vague proper prior does not solve the problem and the posterior distribution will pile up at the boundary of the parameter space, with a dramatic dependence on the values of the hyperparameters.

The problem of finding *robust* objective priors for this model has been tackled from a different perspective. Given that a formal reference prior cannot be derived, the idea is to leverage the notion of admissibility. Proper priors always provide admissible estimators for $\boldsymbol{\beta}$; also, improper priors may be seen as the limit of appropriate sequences of proper priors. As a consequence, they are at the 'boundary of admissibility'. So, if

a given improper prior results in an admissible estimator, it can be considered a valid candidate prior for an objective analysis. For the above situation, Berger et al. (2005) have proposed the following prior with independent components

$$\pi(\boldsymbol{\beta}, \boldsymbol{V}) \propto \frac{1}{(1+||\boldsymbol{\beta}||^2)^{(d-1)/2}} |\boldsymbol{V}|^{1/(2k)-1} \left(\prod_{i<j} (\lambda_i - \lambda_j) \right)^{-1}, \tag{3}$$

where the λ's are the eigenvalues of $\boldsymbol{V}$, and d is the dimension of $\boldsymbol{\beta}$. The admissibility of this prior has been recently proved by Berger et al. (2005). The above result, although very important, is not easy to extend outside the Gaussian set,-up, where a useful characterization of admissibility actually exists (Brown, 1971). An important exception can be found in Spitzner (2005). For the broad class of generalized linear models, two new classes of priors are proposed from an Empirical Bayes perspective. These classes of priors 'correct' the Jeffreys' prior, produce a shrinkage effect on the maximum likelihood estimator and achieve a risk reduction.

Nonparameteric models

While this review is focused on objective Bayesian methods for parametric models, it has theoretically some relevance also for Bayesian nonparametric (BNP) methods, because BNP could be more fittingly defined as "massively parametric Bayes" (Müller and Mitra, 2013). In practice however objective BNP methods are far less developed, and one can find a few reasons for this. In principle, one could argue that BNP methods are intrinsically objective in the sense that they use models with very large, if not full, support. In this context, trying to be "objective" also in the choice of the hyperparameters would seem like a daring enterprise. In the BNP literature, the Dirichlet process and its generalizations represents the staple approach to inference. Along this line of research Bush et al. (2010) and Lee et al. (2014) have proposed a minimally informative version of the mixture of Dirichlet process model, in which the *size* M and the base measure F_0 are selected using the concept of local mass. In a broader perspective, one can interpret the extensions of the Dirichlet process, such us the normalized generalized gamma process (De Blasi et al., 2015), as an impulse towards objectivity, or at least towards the construction of more flexible and robust priors, which allow different tail behaviors for some specific functionals of interest. Another link between objective inference and BNP can be found in the search of those prior processes which attain a minimax (adaptive) posterior concentration rates (Rivoirard and Rousseau, 2012; Hoffmann et al., 2015).

High-dimensional models

As already hinted in Section 1, current research is progressively developing objective methods which produce proper priors that can be used both in estimation and testing scenarios. One reason is the sheer complexity and dimensionality of the problems involved that make the derivation of a formal objective prior too hard or even impossible.

A second motivation is that objective improper priors for estimation may not guarantee proper posterior when the number of parameters exceeds the sample size. Actually the difficulty is more acute because even proper objective priors may lead to posterior

distributions which are not satisfactory from several perspectives. To illustrate this point, let us consider the following example.

Example. *Sparse Multinomial Model* (Berger et al., 2015). Assume a multinomial experiment with many, say $m = 1000$ cells. In the absence of specific quantities of interest, the Jeffreys' and reference priors are both the proper Dirichlet$(1/2, \ldots, 1/2)$ prior. However, this prior is not recommended in the presence of sparsity and small sample size n. For example, with $n = 3$, assume we observe $x_{111} = 2$, $x_{976} = 1$ and all the other $x_j = 0$. The posterior means would be $\mathbb{E}(\theta_i|\boldsymbol{x}) = (x_i + 0.5)/(n + 0.5\ m)$ so that $\mathbb{E}(\theta_{111}|\boldsymbol{x}) = 2.5/503$, $\mathbb{E}(\theta_{976}|\boldsymbol{x}) = 1.5/503$ and all other parameters have a posterior mean equal to $0.5/503$. Then, cells 111 and 976 only have total posterior probability of 0.008 even though all 3 observations are in these cells. Here the problem is that the prior mass, far from being noninformative, overwhelms the role of the data. We discuss in more detail these issues in Section 4.

Further contributions

A recent and promising approach has been developed in Simpson et al. (2017) where the main goal is not to derive formal objective priors for a specific model. Rather the authors aim at identifying those parts of a complex model which require a (hopefully minimal) subjective input to be elicited in a principled way. Suppose one has a base model M_0, characterized by some parameter value ξ_0, say $f_0(\cdot|\xi_0)$. Then, a richer and more flexible model can be denoted by $f(\cdot|\xi)$. In order to characterize the complexity of f compared to f_0, one can build a so called *penalizing complexity* prior on ξ, which depends on a function of the Kullback–Leibler divergence between the base model and the alternative models indexed by ξ, $d(\xi)$. The authors propose to derive the prior based on a principle of constant rate penalization which automatically implies an exponential prior on $d(\xi)$. Details and discussion about advantages, disadvantages, and its debatable status of objectivity can be found in Simpson et al. (2017).

The derivation of an objective prior, whatever method is considered, is strictly dependent on the statistical model under investigation. A list, inevitably incomplete, of priors tailored to specific models that have been proposed in the recent past includes: bivariate copula models (Guillotte and Perron, 2012), skew-symmetric models (Branco et al., 2013; Rubio and Liseo, 2014; Dette et al., 2017), small area models (Datta and Rao, 2010; Arima et al., 2012), capture-recapture models (Xu et al., 2014), autoregressive time series (Liseo and Macaro, 2013; Sorbye and Rue, 2017), survival models (Vallejos and Steel, 2015), Dallal model for bilateral data (M'lan and Chen, 2015), and generalized marginal mixed models (Bodnar et al., 2016).

3 Objective Bayes model comparison

3.1 Some general issues

It is common practice to regard a *statistical model* as a family of distributions for the observable random variables, and we follow suit. Model selection involves the computation of the posterior distribution on a collection of statistical models; we may then

summarize the latter distribution in order to single out a unique representative, which is the typical goal of model selection.

To fix notation for the rest of the paper let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ denote the available observations and suppose we wish to compare the following two models:

$$\begin{aligned} &\text{model } M_0 : f(\boldsymbol{y}|\boldsymbol{\theta}_0, M_0), \quad \boldsymbol{\theta}_0 \in \Theta_0, \\ &\text{model } M_\ell : f(\boldsymbol{y}|\boldsymbol{\theta}_\ell, M_\ell), \quad \boldsymbol{\theta}_\ell \in \Theta_\ell, \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_\ell$ are unknown, model specific, parameters of size d_0 and d_ℓ respectively. If M_0 is nested in M_ℓ, so that $d_0 < d_\ell$, we will henceforth assume that $\boldsymbol{\theta}_\ell = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_{\ell\backslash 0}^T)^T$, so that $\boldsymbol{\theta}_0$ is a parameter 'common' between the two models, whereas $\boldsymbol{\theta}_{\ell\backslash 0}$ is model specific. The use of a 'common' parameter $\boldsymbol{\theta}_0$ in nested model comparison is often made to justify the employment of the same, potentially improper, prior on $\boldsymbol{\theta}_0$ across models. This usage is becoming standard, but is not always appropriate, in particular when the intrinsic prior methodology is adopted; see e.g. Casella and Moreno (2006). We will return briefly to this issue below. Let $\pi(\boldsymbol{\theta}_0|M_0)$ be the prior under the null model M_0, and without loss of generality let the prior under model M_ℓ, have the following hierarchical form:

$$\pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_{\ell\backslash 0}|M_\ell) = \pi(\boldsymbol{\theta}_0|M_\ell)\pi(\boldsymbol{\theta}_{\ell\backslash 0}|\boldsymbol{\theta}_0, M_\ell). \tag{5}$$

To illustrate various approaches to the construction of priors on parameters, we will use the variable selection problem in normal linear regression models as a running important example. In this case, model M_ℓ is specified by

$$\boldsymbol{Y}|\boldsymbol{X}_\ell, \boldsymbol{\beta}_\ell, \sigma^2, M_\ell \sim N_n(\boldsymbol{X}_\ell \boldsymbol{\beta}_\ell, \sigma^2 \boldsymbol{I}_n), \tag{6}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is the vector of responses, $\boldsymbol{X}_\ell$ is a known $n \times (p_\ell + 1)$ design matrix (p_ℓ covariates plus the intercept), $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{\beta}_\ell$ is a $(p_\ell + 1)$-vector of regression coefficients, and σ^2 is the error variance, common to all models. Therefore each model M_ℓ has parameters $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2)$ of size $d_\ell = p_\ell + 2$. With M_0 we denote the null model having the intercept only, with parameters $\boldsymbol{\theta}_0 = (\beta_0, \sigma^2)$, and with M_F the full model with all p covariates under consideration. For model M_ℓ we write $\boldsymbol{\beta}_\ell = (\beta_0, \boldsymbol{\beta}_{\ell\backslash 0}^T)^T$ and $\boldsymbol{X}_\ell = [\boldsymbol{X}_0, \boldsymbol{X}_{\ell\backslash 0}]$, where $\boldsymbol{X}_0$ is the n-dimensional unit vector. All matrices $\boldsymbol{X}_\ell$ are assumed to be of full rank. Moreover, in the case of variable selection, it is common to substitute the model indicator M_ℓ by a vector of binary indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ that identify which covariates are included in the model (George and McCulloch, 1993).

3.2 Posterior measures of evidence

A natural tool for comparing model M_0 versus M_ℓ is the posterior odds (Jeffreys, 1961) defined by

$$PO_{0\ell} = \frac{\pi(M_0|\boldsymbol{y})}{\pi(M_\ell|\boldsymbol{y})} = \frac{m_0(\boldsymbol{y})}{m_\ell(\boldsymbol{y})} \times \frac{\pi(M_0)}{\pi(M_\ell)}, \tag{7}$$

where $\pi(M_k)$ is the prior probability of model M_k, $k \in \{0, \ell\}$, while $m_k(\boldsymbol{y})$ is the "marginal" likelihood (also called Bayesian "evidence") of M_k given by $m_k(\boldsymbol{y}) =$

$\int_{\Theta_k} f(\boldsymbol{y}|\theta_k, M_k)\pi(\theta_k|M_k)d\theta_k$, with $\pi(\theta_k|M_k)$ denoting the prior distribution of θ_k under model M_k, $k \in \{0, \ell\}$. The ratio of the marginal likelihoods of the two models is called the Bayes factor (BF)

$$B_{0\ell} = \frac{m_0(\boldsymbol{y})}{m_\ell(\boldsymbol{y})}. \tag{8}$$

From (7) it appears that the BF is the multiplicative term, or factor, which updates the prior odds $\pi(M_0)/\pi(M_\ell)$ to the posterior odds $PO_{0\ell}$. The terminology is due to Good (1958), and the initial use of the BF can be attributed both to Jeffreys and Turing who introduced it independently around the same time (Kass and Raftery, 1995). Notice that if equal prior model probabilities are assumed (prior indifference between models), the posterior odds reduce to the Bayes factor. The BF does not depend on the prior model probabilities; however it depends on the prior densities $\pi(\theta_k|M_k)$, which in general must be proper. Notice that in some cases improper priors are allowed. For instance, Berger et al. (1998) proved a remarkable result which states that, in situations characterized by a group structure leading to invariance considerations, right Haar priors are perfectly legitimate to be used for computing BFs. Additionally, use of improper priors is common in nested scenarios, dating back to Jeffreys (1961); see also Kass and Raftery (1995). Improper priors may be used, although not in a direct way, for computing BFs; see Subsection 3.4 for more details.

Posterior model odds (and BFs) are directly related to posterior model probabilities $\pi(M_\ell|\boldsymbol{y})$ because

$$\pi(M_\ell|\boldsymbol{y}) = \frac{m_\ell(\boldsymbol{y})\pi(M_\ell)}{\sum_{M_k \in \mathcal{M}} m_k(\boldsymbol{y})\pi(M_k)} = \frac{PO_{\ell 0}}{\sum_{M_k \in \mathcal{M}} PO_{k0}}, \tag{9}$$

for any model $M_\ell, M_0 \in \mathcal{M}$. If $M_\ell, M_0 \in \mathcal{M}$ are the only two models under consideration and they have the same prior probabilities, then $\pi(M_\ell|\boldsymbol{y}) = 1/(1 + B_{0\ell})$. The posterior model probability (9) is often interpreted as the probability that M_ℓ is the "true" data generating model. Notice however that this interpretation is meaningful only under the $\mathcal{M}$-closed view, wherein it is assumed that the true model is included in the set of models under consideration, and provided that the induced Bayesian procedure is consistent (see Section 3.3 for details). In most real life problems, the $\mathcal{M}$-closed view is unrealistic. Nevertheless, measures of Bayesian model comparison support models (in $\mathcal{M}$) that are close in Kullback–Leibler divergence to the true generating mechanism; see for details Walker et al. (2004), Clyde and Iversen (2013), Chib and Kuffner (2016). A disadvantage of using $\pi(M_\ell|\boldsymbol{y})$, as opposed to posterior odds or BFs, is the "dilution" of the posterior probability over the space of models (George, 1999), which becomes spread out over many similar models. Dilution increases as more models are considered, so that posterior model probabilities, even for the maximum a-posteriori (MAP) model, decrease. For this reason it is advised to report, besides the posterior probability of each model, also its posterior odds or BF against the MAP model.

For the variable selection problem, we may further calculate the posterior inclusion probabilities for each covariate X_j given by

$$\pi(\gamma_j = 1|\boldsymbol{y}) = \sum_{M_\xi \in \mathcal{M}_j} \pi(M_\xi|\boldsymbol{y}) = \frac{\sum_{M_\xi \in \mathcal{M}_j} PO_{\xi 0}}{\sum_{M_k \in \mathcal{M}} PO_{k0}},$$

where $\mathcal{M}_j = \{M \in \mathcal{M} : \gamma_j = 1\} \subset \mathcal{M}$ is the set of all models in $\mathcal{M}$ with variable X_j included in the model formulation. Posterior inclusion probabilities (George and McCulloch, 1993) represent an accumulated measure of evidence in favor of a covariate being present in a model structure, and have been used as an informal, empirical measure of evidence for many years. Their usefulness was highlighted in the work by Barbieri and Berger (2004) where it was proved that the median probability (MP) model, defined as that model containing only covariates whose posterior inclusion probabilities exceeds the value 0.5, has better predictive properties than the MAP model in specific cases. Posterior inclusion probabilities do not generally suffer from the phenomenon of posterior dilution because they can be written as

$$\pi(\gamma_j = 1|\boldsymbol{y}) = 1/(1 + O_j) \text{ with } O_j = \frac{\sum_{M_\xi \in \mathcal{M}_j} PO_{\xi 0}}{\sum_{M_k \in \overline{\mathcal{M}}_j} PO_{k0}},$$

where $\overline{\mathcal{M}}_j = \{M \in \mathcal{M} : \gamma_j = 0\} \subset \mathcal{M}$ is the complementary set of $\mathcal{M}_j$. In the above expression, the numerator and the denominator of O_j are sums of 2^{p-1} elements making this quantity robust when we decide to increase the number of covariates under evaluation. Similarly, when using any tool of model exploration in large model spaces, posterior inclusion probabilities are more reliably and quickly estimated than individual posterior model probabilities due to the large number of models with small but non-zero probability involved in the denominator of (9).

There is a growing interest in applying posterior measures of evidence in empirical research. For instance the *Journal of Mathematical Psychology* recently devoted a whole issue to this topic; see the introductory editorial page by Mulder and Wagenmakers (2016). One reason might be the acute dissatisfaction with current frequentist testing methods, also related to lack of reproducibility in scientific investigations; see Johnson (2013) and the recent statement by the American Statistical Association (Wasserstein and Lazar, 2016). Benjamin et al. (2017) is the outcome of a concerted effort by a large group of statisticians and scientists to define more stringent statistical standards of evidence for claiming new discoveries in many fields of science.

We close this subsection by presenting a variety of viewpoints on the issue of Bayesian model comparison from an objective standpoint. First of all it is worth pointing out that the use of the BF is not undisputed even within the OB community. Bernardo and Rueda (2002) consider testing a null model nested into a larger one. They argue that a testing problem should be regarded as a formal decision problem on whether or not to use the null model. Accordingly a loss function should be specified to take into account the amount of information which would be lost if the null model were used. Objectivity comes into the picture through the use of a reference prior on the parameter space. Dawid and Musio (2015) address the problem of the indeterminacy of the marginal likelihood of a model in the presence of an improper prior, and solve

it by replacing the marginal log-likelihood with a homogeneous proper scoring rule, which is insensitive to the arbitrary scaling constant of the prior. They also show that, when suitably applied, their proposal will typically enable consistent selection of the true model. Kamary et al. (2014) propose to view the model selection enterprise as a problem in mixture modeling. Specifically the models under investigation are viewed as components of a mixture model, so that the original testing problem is transformed into an estimation problem, and accordingly the posterior probability of a model or an hypothesis is evaluated through the posterior distribution of the weights of a mixture of the models under comparison. Again improper priors can be used, although some care must be exercised. In order to perform OB methods for testing or selection, other authors rely on an unconventional use of the BF. Johnson (2005) proposes a test-based BF (TBF) for two nested models which is defined through a test statistic, rather than individual observations. The main idea is that the distribution of a test statistic does not depend on unknown model parameters under the null, so that some of the subjectivity that is normally associated with the definition of Bayes factors is eliminated. It remains to compute the marginal likelihood under the alternative model: this can be obtained through a prior or using a marginal maximum likelihood estimate. Further aspects are examined in Hu and Johnson (2009), while Held et al. (2015) relate BF's based on g-priors (discussed in Section 3.4) to TBF's. Finally Johnson (2013) introduces the concept of a uniformly most powerful Bayesian test (UMPBT) for testing a null model nested in a larger alternative one. A UMPBT is such that the prior under the alternative hypothesis is determined so as to maximize the probability that a Bayes factor against the null exceeds a specified threshold for each possible value of the true parameter belonging to the alternative set.

3.3 Principles for objective model comparison

Criteria for objective Bayesian model choice

Bayarri et al. (2012) developed criteria (*desiderata*) to be satisfied by objective prior distributions for Bayesian model choice. A number of these criteria are applicable only in nested model comparisons. Notice that this represents a distinctive innovation relative to previous attempts in the literature which typically proposed, based on intuition or otherwise, reasonable priors which were subsequently evaluated in terms of their properties. Here the paradigm is turned upside down: first criteria meaningful for priors tailored to objective model selection are set out, and then priors satisfying them are derived. These criteria are grouped into four classes: *basic*, *consistency*, *predictive matching* and *invariance*. The *basic criterion (C1)* states that the prior of each model specific parameter, conditionally on the common ones, $\pi(\boldsymbol{\theta}_{\ell\setminus 0}|\boldsymbol{\theta}_0, M_\ell)$ should be proper, so that Bayes factors do not contain different arbitrary normalizing constants across distinct models.

Model selection consistency (C2) has been widely used as a crucial criterion for objective model selection priors. The criterion implies that if data have been generated by M_ℓ, then the posterior probability of M_ℓ should converge to one as the sample size diverges. Although consistency is an important requirement, it might not be enough to

differentiate between several priors, all satisfying (C2). Hence the need to better investigate the rate of convergence to the true model. Current research in high-dimensional models, on which we report in Section 4, is precisely devoted to this issue; see in particular Castillo and Misner (2018) and Ročková and George (2018). An additional consistency criterion is *information consistency (C3)*: if there exists a sequence of datasets with the same sample size n such that the likelihood ratio between M_ℓ and M_0 goes to infinity, then the corresponding sequence of Bayes factors should also go to infinity. Information inconsistency was first discovered by Berger and Pericchi (2001) in the case of conjugate priors for location when the scale is unknown and was further studied by Liang et al. (2008). It represents a severe lack of robustness to highly specific sample information. When some aspects of the model, sample size and, to some extent, also of the observations, affect model selection priors, it is desirable that such features should disappear as n grows, leading to a limiting proper prior. This requirement is named *intrinsic consistency criterion (C4)*.

Predictive matching (C5) is viewed as the most crucial aspect for objective model selection priors. Informally, with a *minimal sample size*, one should not be able to discriminate between two models, so that the BF should be close to one, for all samples of minimal size. In particular, *exact predictive matching* occurs if the BF equals one. The minimal sample size n^* is defined as the smallest sample size with a finite nonzero marginal density for the combination of models and priors; i.e. $0 < m(\boldsymbol{y}^*|M_\ell) < \infty$ for all observations $\boldsymbol{y}^*$ of size n^*, and for all models M_ℓ under the prior $\pi(\boldsymbol{\theta}_\ell|M_\ell)$. Bayarri et al. (2012) elaborate further on the notion of predictive matching, but we omit details for the sake of conciseness.

The last two criteria are in terms of *invariance arguments*. *Measurement invariance (C6)* broadly states that answers should not be affected by changes of measurement units. A special type of invariance arises when the families of sampling distributions of models under consideration are such that the model structures are invariant to group transformations. The *group invariance criterion (C7)* states that if models M_ℓ and M_0 are invariant under a group of transformations G_0, then the conditional priors $\pi(\boldsymbol{\theta}_{\ell\setminus 0}|\boldsymbol{\theta}_0, M_\ell)$ should be chosen in such a way that the conditional marginal distribution $f(\boldsymbol{y}|\boldsymbol{\theta}_0, M_\ell)$ is also invariant under G_0. This means that if models exhibit an invariance structure, this should be preserved after marginalization. Note that G_0 is a group of transformations relevant to the null model M_0, and therefore the group invariance criterion can be understood as a formalization of the Jeffreys' requirement that the prior for a non-null parameter should be "centered at the simplest model." Another use of invariance is to find priors on common parameters.

Remarkably, Bayarri et al. (2012) accomplished the goal of finding a prior satisfying all their *desiderata* within the framework of normal linear regression models, which they called *robust prior*. Under model M_ℓ, as in (6), the prior takes the form

$$\pi^R(\boldsymbol{\beta}_{\ell\setminus 0}, \boldsymbol{\beta}_0, \sigma|M_\ell) \propto \sigma^{-1} \int_0^{+\infty} N_{p_\ell - p_0}(\boldsymbol{\beta}_{\ell\setminus 0}|\mathbf{0}, g\boldsymbol{\Sigma}_{\ell\setminus 0})\pi^R(g)dg, \tag{10}$$

where $\boldsymbol{\Sigma}_{\ell\setminus 0} = \sigma^2(\boldsymbol{V}_{\ell\setminus 0}^T \boldsymbol{V}_{\ell\setminus 0})^{-1}$, $\boldsymbol{V}_{\ell\setminus 0} = (\boldsymbol{I}_n - \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T)\boldsymbol{X}_{\ell\setminus 0}$, and

$$\pi^R(g) = a\,[\rho_\ell(b+n)]^a\,(g+b)^{-(a+1)} 1_{\{g > \rho_\ell(b+n)-b\}}, \quad a, b > 0 \text{ and } \rho_\ell \geq \frac{b}{b+n}.$$

While the result holds for a general matrix of common predictors $\boldsymbol{X}_0$, note that, if $\boldsymbol{X}_0 = \mathbf{1}$ (i.e. when M_0 contains only the intercept), then $\boldsymbol{V}_{\ell\setminus 0} = \boldsymbol{Z}_{\ell\setminus 0}$, with $\boldsymbol{Z}_{\ell\setminus 0}$ denoting the column-wise centered version of $\boldsymbol{X}_{\ell\setminus 0}$.

Regarding the hyperparameters of the above prior distribution, the default values recommended by Bayarri et al. (2012) are $a = 1/2$, $b = 1$ and $\rho_\ell^{-1} = p_\ell + 1$. Under the robust prior, the resulting Bayes factors have closed form expressions in terms of the hypergeometric function. Finally, the hyper-g-prior (Liang et al., 2008), discussed in Section 3.4, using the recommended value of 3 for its hyperparameter, is a particular case of the robust prior with $a = 1/2$, $b = 1$ and $\rho_\ell^{-1} = n + 1$; similarly, the hyper-g/n-prior (Liang et al., 2008), using the recommended value of 3 for its hyperparameter, may be obtained from the robust prior by setting $a = 1/2$, $b = n$ and $\rho_\ell^{-1} = 2$.

Compatibility of priors

When dealing with model choice, a prior on the parameter space under each model should incorporate not only uncertainty but also features which are germane to the comparison setting. One important feature is *compatibility* of priors across models; see Dawid and Lauritzen (2011) and Consonni and Veronese (2008). Informally this means that priors should be related across models, although in principle they need not be, each being conditional on a given model. Compatibility is usually applied to nested models, with parameter spaces having different dimensions, but it can be extended to more general setups whenever we can identify a benchmark model (often the null model), which is nested into every other model under consideration (encompassing from below), so that compatibility is realized between each model and the benchmark, and thus indirectly between any pair of models. Compatibility was initially proposed to lessen the sensitivity of model comparison to prior specifications, and also to facilitate the task of multiple prior elicitations when several models are entertained. However it also underlies some approaches to the construction of objective priors for Bayesian testing, e.g. the expected posterior prior (Pérez, 1998) (see Section 3.4), wherein the prior under each model is anchored to a common base measure. Another version of prior compatibility across models, named *matching*, was examined at the beginning of Section 3.3 within a more general theoretical setup.

Validation of Bayesian approaches

The desiderata of Bayarri et al. (2012) refer to the desirable properties of prior distributions and the induced model selection procedures. Nevertheless, when more general methods with Bayesian motivation are used (e.g. the intrinsic and the fractional Bayes factors; see Section 3.4) then an additional important property should be satisfied. According to Principle 1 of Berger and Pericchi (1996), "methods that correspond to use of plausible default (proper) priors are preferable to those that do not correspond to any possible actual Bayesian analysis". Thus an acceptable Bayesian procedure should correspond, at least asymptotically, to a prior which makes sense in the context where it is applied.

Methods with good frequentist properties

A popular alternative to the standard objective Bayes techniques is to use prior distributions that lead to good frequentist performances. This trend is especially notable in high-dimensional settings as we discuss in Section 4. For instance, priors are selected based on the coverage of posterior intervals and false discovery rates (FDR). The former focuses on estimation (Castillo and van der Vaart, 2012; van der Pas et al., 2017), and is further discussed in Section 4, while the control of FDR is tailored to multiple comparisons and prior model probability specification (Tansey et al., 2018); see Section 3.6.

3.4 Methods for constructing objective prior distributions

Unit information principle

The unit information principle has its origin in the work of Kass and Wasserman (1995) who investigated the use of the Schwarz (1978) criterion (or BIC) as an approximation of the Bayes factor. Informally, a unit information prior (UIP) has an information content equivalent to a sample of size one. For a dataset of size n, the observed Fisher information matrix under model M_ℓ divided by n can be interpreted as an estimate of the average amount of information in one data point. If $\boldsymbol{\theta}_\ell \in \mathbb{R}^{d_\ell}$ one way to construct a UIP is as follows

$$\boldsymbol{\theta}_\ell | M_\ell \sim \mathrm{N}_{d_\ell}\Big(\boldsymbol{\mu}_{\theta_\ell},\, n\left[\mathcal{J}_\ell^n\big(\boldsymbol{\mu}_{\theta_\ell}\big)\right]^{-1}\Big), \tag{11}$$

where $\mathcal{J}_\ell^n(\cdot)$ is the negative of the Hessian matrix of the log-likelihood. Under this prior the logarithm of the BF is asymptotically equivalent to the Schwarz criterion (BIC). In this way the unit information prior provides a Bayesian interpretation for the BIC model selection procedure.

There exist specifications of UIPs alternative to (11); for instance one could replace $\boldsymbol{\mu}_{\theta_\ell}$ with the maximum-likelihood estimate. In the same spirit, Ntzoufras (2009) proposed a simplified version by considering independent prior distributions with means and variances equal to the corresponding posterior means and the variances (multiplied by n) obtained using a flat improper prior. The posterior model probabilities under this approach can be used as an initial yardstick for comparisons with other objective Bayes approaches.

The unit information principle can be easily combined with the power-prior approach described shortly below. Under this setting, the prior mean $\boldsymbol{\mu}_{\theta_\ell}$ can be specified by "prior", or "imaginary", data. A sensible choice, for nested model comparisons, is to generate the latter under the null model. Examples of priors based on the unit information principle can be found in Ntzoufras et al. (2003) for binary response models, in Overstall and Forster (2010) for generalized linear mixed models, in Sabanés Bové and Held (2011) for generalized linear models, and in Ntzoufras and Tarantola (2008) for contingency tables.

The unit information principle rests on the notion of sample size which is straightforward for i.i.d. observations, but requires careful considerations in other settings, such as non-i.i.d. observations or in hierarchical models. In Bayarri et al. (2014) the concept

of effective sample size is analyzed in detail, and applied to the construction of priors for model selection in a variety of statistical setups.

Training samples

This subsection by itself does not represent a direct method for constructing priors: its goal is rather to motivate the use of intrinsic priors which are described in the subsequent paragraph.

The difficulties in computing the Bayes factor under improper priors, mentioned in Section 3.2, have generated a few proposals that try to address them. One line of research rests on the use of training samples and led to the *intrinsic Bayes factor* (IBF) proposed by Berger and Pericchi (1996). The IBF employs a subset of the data, of size n^* (the training sample) to convert the improper baseline prior to a proper posterior, and then uses the remaining data to calculate the Bayes factor. Next, a summary, e.g. median, arithmetic or geometric mean, of the Bayes factors over the set of all possible sub-samples of size n^* can be reported, resulting in the median, arithmetic or geometric intrinsic Bayes factors respectively. Under the IBF approach, *minimal training samples* are often employed in order to minimize the loss of data utilized for building the prior distribution. These samples are defined such that their size is "as small as possible, subject to yielding proper posteriors" (Berger and Pericchi, 1996). The IBF has the disadvantage that in principle one should consider all possible sub-samples having a minimal sample size, and then take averages. This can be computationally costly. A way to overcome this difficulty is to resort to *intrinsic priors* which we describe below.

A related method is the *fractional Bayes factor* (FBF) proposed by O'Hagan (1995), which however does not require training samples. In order to compute the marginal likelihood of a given model using an improper prior, the prior is "trained" using a fraction of the full sample likelihood, that is raising the full likelihood to a power. Next the calculation of the marginal likelihood is implemented using the complementary fraction of the likelihood together with the newly trained prior. The FBF is appealing because of its simplicity, and has been used to address challenging statistical problems involving model comparison. In particular, we mention here two areas: multivariate time series models (Corander and Villani, 2004, 2006; Villani, 2001), and graphical models (Carvalho and Scott, 2009; Consonni and La Rocca, 2012; Altomare et al., 2013; Leppä-aho et al., 2016; Consonni et al., 2017). Recent theoretical work on Bayesian fractional posteriors (Bhattacharya et al., 2016), while not directly motivated by OB methodologies and having a much broader scope, may provide useful results for further investigation into properties of FBF.

Intrinsic priors Intrinsic prior distributions were originally introduced by Berger and Pericchi (1996) in order to provide a proper Bayesian interpretation for intrinsic Bayes factors, according to the principle that a good Bayesian procedure should correspond to the use (at least asymptotically) of a sensible prior; see Section 3.3.

The intrinsic prior can be obtained by equating the limit (as $n \to \infty$) of the arithmetic intrinsic Bayes factor with the corresponding Bayes factor obtained by using the

intrinsic prior resulting in two intrinsic equations for every pair of models under comparison. For any two nested models under comparison M_ℓ and M_0, the two equations coincide. Although the intrinsic prior distributions always exist for nested model comparisons (Sansó et al., 1996), the intrinsic equations do not collapse into a single equation in non-nested cases. Therefore, the existence of the intrinsic priors is not ensured, and when they exist, we obtain a class of intrinsic prior distributions rather than a single solution (Moreno, 2005). Berger and Pericchi (1996) prove that in nested situations, the arithmetic, but not the geometric, IBF corresponds to a proper prior under the "alternative" when the "null" is simple, or when the baseline prior under the "null" is proper.

Consider the comparison of a "null" model $M_0 = \{f(\cdot|\boldsymbol{\theta}_0, M_0), \pi^N(\boldsymbol{\theta}_0|M_0)\}$ nested in model $M_\ell = \{f(\cdot|\boldsymbol{\theta}_\ell, M_\ell), \pi^N(\boldsymbol{\theta}_\ell|M_\ell)\}$. The baseline priors in each model are assumed to be objective, typically improper, and the superscript "N" stands for "noninformative." In this part of the paper only, we depart somewhat from the notation employed in Section 3.1 because both $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_\ell$ are meant to be model specific parameters without assuming that $\boldsymbol{\theta}_0$ is a 'common' parameter. If we assume that the intrinsic priors are limit of proper intrinsic priors then it can be shown (Moreno et al., 1998) that the pair

$$\begin{aligned} \pi^I(\boldsymbol{\theta}_0|M_0) &= \pi^N(\boldsymbol{\theta}_0|M_0), \\ \pi^I(\boldsymbol{\theta}_\ell|M_\ell) &= \pi^N(\boldsymbol{\theta}_\ell|M_\ell)\mathbb{E}_{\boldsymbol{Y}(l)|\boldsymbol{\theta}_\ell, M_\ell}\left[B^N_{0\ell}(\boldsymbol{Y}(l))\right], \end{aligned} \tag{12}$$

are the unique limit of proper intrinsic priors, where $B^N_{0\ell}(\boldsymbol{y}(l)) = \frac{m^N(\boldsymbol{y}(l)|M_0)}{m^N(\boldsymbol{y}(l)|M_\ell)}$ is the Bayes factor of model M_0 versus M_ℓ evaluated at the training sample $\boldsymbol{y}(l)$, and $m^N(\boldsymbol{y}(l)|M_\ell) = \int f(\boldsymbol{y}(l)|\boldsymbol{\theta}_\ell, M_\ell)\pi^N(\boldsymbol{\theta}_\ell|M_\ell)d\boldsymbol{\theta}_\ell$, with a similar expression holding for $m^N(\boldsymbol{y}(l)|M_0)$. Hence, the intrinsic prior under model M_ℓ is the baseline prior $\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$ adjusted by the expected Bayes factor of M_0 against M_ℓ with respect to the distribution of $\boldsymbol{Y}(l)$ under model M_ℓ.

If the prior $\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$ is improper, so that its expression is unique up to a constant c_ℓ, an important feature of the intrinsic prior is that it is free from c_ℓ. Indeed $\pi^I(\boldsymbol{\theta}_\ell|M_\ell)$ only depends on the constant c_0 of the (improper) prior $\pi^N(\boldsymbol{\theta}_0|M_0)$ under the null model M_0. However, if the latter is nested into every M_ℓ, meaning that M_0 can be taken as a null, or baseline, model in all pairwise comparisons, c_0 will appear as a multiplicative constant in the intrinsic prior distribution of *each* model M_ℓ, and therefore will cancel out in the ensuing Bayes factors, causing no indeterminacy problem in the resulting model comparison procedure based on intrinsic priors.

As in Berger and Pericchi (1996), also in Moreno et al. (1998) it has been proved that in nested model comparisons, if the baseline prior for the reference model M_0 is proper, then $\pi^I(\boldsymbol{\theta}_\ell|M_\ell)$ is also proper and unique under mild conditions. However, additionally, Moreno et al. (1998) constructed a limiting intrinsic procedure for the case where $\pi^N(\boldsymbol{\theta}_0|M_0)$ is not proper. General theory for intrinsic tests and comparisons between nested models or hypotheses can be found in Moreno (1997) while for non-nested comparisons results are available in Berger and Mortera (1999) and in Cano et al. (2004). Cano and Salmerón (2013) generalized the intrinsic prior formulae, for non-nested situations, by iteration.

Objective model comparison and hypothesis testing based on intrinsic priors have been implemented in a variety of problems. Here we can only list a subset of them which have appeared in the more recent years: analysis of variance models with heteroscedastic errors (Bertolino and Racugno, 2000), survival analysis models (Kim and Sun, 2000), tests for the selection of the number of mixture components (Moreno and Liseo, 2003), one-sided hypothesis tests (Moreno, 2005), test for the equality of regression coefficients with heteroscedastic errors (Moreno et al., 2005), changepoint problems (Girón et al., 2007), one-way random effects models (Garcia-Donato and Sun, 2007), the equality of two correlated proportions (Consonni and La Rocca, 2008), two-way contingency tables (Casella and Moreno, 2009), comparisons in multivariate normal regression models (Torres-Ruiz et al., 2011), Hardy Weinberg equilibrium models (Consonni et al., 2011), and comparison of constrained ANOVA models (Consonni and Paroli, 2017). Finally in Pérez et al. (2017) a sensible prior to substitute the inverted gamma prior for scales is found as an intrinsic prior, and shown to generate by marginalization the horseshoe prior described in Section 4.

Moreover, intrinsic priors have been successfully used for variable selection in normal regression (Casella and Moreno, 2006), multivariate regression (Torres-Ruiz et al., 2011) and probit models (Leon-Novelo et al., 2012). For normal regression models with a finite number of predictors, a variety of priors, including the intrinsic, leads to a consistent variable selection procedure (Casella et al., 2009). For models whose dimension grows with the sample size n, Moreno et al. (2010) show that the Bayes factor for nested models under the intrinsic prior is consistent when the size of the model grows as $O(n^b)$ for $b < 1$, and this holds also for the BIC selection procedure. When $b = 1$, the Bayes factor under the intrinsic prior is still consistent, except for a small set of alternative larger models which they characterize. Finally consistency of intrinsic posterior distributions both under model selection and model averaging is studied in Womack et al. (2014). Moreno and Girón (2008) provide a comparison between two different types of encompassing in each pairwise model comparison: "from below", so that the null model is nested into each of the remaining ones and acts as the baseline model, and "from above", considering each model as baseline when compared to the full one; only the former however guarantees the rather obvious coherency requirement that $B_{0\ell}((\boldsymbol{y})/B_{0k}(\boldsymbol{y}) = B_{k\ell}(\boldsymbol{y})$. For a concise review of the intrinsic prior methodology we refer the readers to the recent publication of Moreno and Pericchi (2014).

Intrinsic priors, as virtually all commonly used priors for testing, result in pairwise model comparison procedures with unbalanced learning rates under the two rival hypotheses/models. Specifically, if M_0 is nested within M_ℓ, the BF in favor of M_0 decreases as a power of n if M_ℓ holds; on the other hand, the BF in favor of M_ℓ decreases exponentially fast in the sample size when M_0 holds; see Dawid (2011). To alleviate this imbalance, one can resort to *non-local priors* (Johnson and Rossell, 2010), which we briefly discuss at the end of this subsection. An intrinsic version of non-local priors was implemented for the first time in Consonni et al. (2013) with an application to the comparison of nested models for discrete observations. Alternatively, as one referee pointed out, the imbalance in the learning rate can be also managed by considering "objective" losses that naturally arise in specific problems; see Goutis and Robert (1998), Plummer (2008) and Dawid and Musio (2015) for examples.

Similarly to intrinsic priors, fractional priors have been introduced in the objective Bayes community by Moreno (1997) in order to identify a Bayesian procedure that approximates the results obtained by the FBF. De-Santis and Spezzaferri (1997) derived formulae for the calculation of intrinsic priors of the FBF.

Imaginary observations

One of the main approaches used to construct prior distributions for objective Bayes methods is the concept of *imaginary observations*. The basic idea (whose origin can be traced back to the work of Good, 1950) is to consider a thought experiment with an appropriate dataset that will be used to specify the normalizing constants involved in the Bayes factors when using improper priors (Spiegelhalter and Smith, 1982). The main pathway here was to adopt the "local" principle, where the imaginary dataset fully supports the null hypothesis in nested model comparisons. In order to make the induced methods minimally informative, the notions of minimal training sample and the UIP principles were used in several occasions. A "non-local" alternative has been introduced by Spitzner (2011) who used the notion of "neutral" imaginary samples which result in posterior model odds that do not support either of the two hypotheses; see also Section 3.2 of Spitzner (2011) for details concerning the connection of this approach with the "non-local" priors for a simple hypothesis test. We further distinguish between fixed and random imaginary observations.

Fixed imaginary data

In this subsection, we will focus on three main approaches. We start with the description of power priors, because of their wider scope. We then continue with g priors, and mixture thereof, which are very popular choices in variable selection problems.

Power priors Ibrahim and Chen (2000) and Chen et al. (2000) introduced power priors as a resourceful probabilistic procedure for the elicitation of prior information in the form of additional prior data whose importance is weighted by a power parameter. Although the primary use of the power priors was in subjective Bayes approaches, using historical data to build the prior, they can be used (in combination with the notions of unit information priors) also to build meaningful prior distributions for objective Bayesian analysis through the device of "fixed imaginary data" (Spiegelhalter and Smith, 1980).

Consider model M_ℓ in (4), and let $\pi_\ell^N(\boldsymbol{\theta}_\ell|M_\ell)$ be an objective noninformative prior typically used for estimation purposes. Then for a set of imaginary data $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_{n^*}^*)^T$ of size n^*, a sensible prior for the model parameters can be obtained by the following expression

$$\pi(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, a_0, M_\ell) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)^{a_0}\pi^N(\boldsymbol{\theta}_\ell|M_\ell)\,. \tag{13}$$

The parameter $0 \leq a_0 \leq 1$ controls the weight that the imaginary data contribute to the final posterior distribution of $\boldsymbol{\theta}_\ell$, since

$$\pi(\boldsymbol{\theta}_\ell|\boldsymbol{y},\boldsymbol{y}^*,a_0,M_\ell) \propto f(\boldsymbol{y}|\boldsymbol{\theta}_\ell,M_\ell)f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell,M_\ell)^{a_0}\pi^N(\boldsymbol{\theta}_\ell|M_\ell)\,.$$

For $a_0 = 1$, the prior (13) is exactly equal to the posterior distribution of $\boldsymbol{\theta}_\ell$ after observing the imaginary data $\boldsymbol{y}^*$. Usually, when limited prior information is available, we let $a_0 = 1/n^*$ inducing contribution of the imaginary data to the overall posterior which is equivalent to one data point; i.e. the prior has a unit information interpretation. Moreover, the imaginary data can be generated from the simplest model (when available) under comparison in order to *a priori* support more parsimonious models. This specification can serve as a sensible default choice to conduct Bayesian analysis in a minimally informative way.

g-priors Zellner's (1986) g-prior is one of the standard choices of prior distributions for variable selection in the normal linear regression models. It has been widely used due to its computational convenience, direct interpretation and its connection to the widely used BIC. Its original formulation is given by

$$\boldsymbol{\beta}_\ell|\sigma^2,M_\ell \sim N_{p_\ell+1}\Big(\boldsymbol{\mu}_{\beta_\ell},\, g(\boldsymbol{X}_\ell^T\boldsymbol{X}_\ell)^{-1}\sigma^2\Big) \quad \text{and} \quad \pi\big(\sigma^2\big|M_\ell\big) \propto 1/\sigma^2\ , \tag{14}$$

suppressing dependence on $\boldsymbol{X}_\ell$. Up to the term g the prior variance-covariance matrix of $\boldsymbol{\beta}_\ell$ coincides with that of the maximum likelihood estimator of $\widehat{\boldsymbol{\beta}}_\ell$. Formula (14) reports the original specification, wherein the improper prior for σ^2 is meant to provide no information about the error variance; however some researchers extend the term g-prior to more informative settings with σ^2 having a normal-inverse gamma distribution. An alternative version of g-prior has been widely used in literature, see for example Liang et al. (2008). In this approach, after centering all covariates, the intercept is treated as a "common" parameter, and the g-prior takes the form

$$\boldsymbol{\beta}_{\ell\setminus 0}\big|\beta_0,\sigma^2,M_\ell \sim \mathrm{N}_{p_\ell}\Big(\mathbf{0}, g\big(\boldsymbol{Z}_{\ell\setminus 0}^T\boldsymbol{Z}_{\ell\setminus 0}\big)^{-1}\sigma^2\Big) \quad \text{and} \quad \pi(\beta_0,\sigma^2|M_\ell) \propto 1/\sigma^2\ , \tag{15}$$

with $\boldsymbol{\beta}_{\ell\setminus 0}$ denoting the sub-vector of $\boldsymbol{\beta}_\ell$ without the common parameter β_0 and $\boldsymbol{Z}_{\ell\setminus 0}$ denoting the column-wise centered version of $\boldsymbol{X}_{\ell\setminus 0}$.

The g-prior in (14), with $\boldsymbol{\mu}_{\beta_\ell} = \mathbf{0}$, can be interpreted as a power prior with fixed imaginary data $\boldsymbol{y}^* = \mathbf{0}$ of size n and imaginary design matrix $\boldsymbol{X}_\ell$ (same as the sample design matrix), power parameter equal to $a_0 = 1/g$, and a flat baseline prior distribution for $\boldsymbol{\beta}_\ell$ conditionally on σ^2. Similarly, the conditional distribution $\boldsymbol{\beta}_{\ell\setminus 0}|\beta_0,\sigma^2,M_\ell$ in (15) can be interpreted as a power prior with all imaginary data set equal to a pre-specified value.

The g-prior has been widely used in practice for several reasons, among which: (a) analytical tractability for posterior inference; (b) connection to readily available variable selectors such as BIC; (c) ease of prior elicitation, because there is only one unspecified prior hyperparameter, namely g. With regard to (c), notice that g has an interpretation similar to the inverse of the power parameter a_0 in the power prior setup. Therefore it determines the amount of prior information relative to the empirical or imaginary data. The information introduced by the prior can be measured by the ratio n/g and can be considered in terms of the effective sample size of the prior. Hence for the default choice

$g = n$, the prior information will be equivalent to adding one observation in our analysis, while for $g = 1$, the prior information will be equivalent to adding n observations in our analysis. The prior mean of $\boldsymbol{\beta}_\ell$ is usually set equal to zero, also to favor shrinkage of parameter values towards to zero, especially for those components which are not especially relevant. Alternative choices of g have been proposed in the literature; see for example Foster and George (1994) and Fernández et al. (2001). Empirical Bayes approaches have been also proposed for the specification of g; see for example George and Foster (2000), Hansen and Yu (2001) and Liang et al. (2008). Both versions (14) and (15) of g-priors with $g = n$ asymptotically lead to a BIC based variable selection procedure.

Zellner's g-prior leads to a consistent variable selection method; however it suffers from an "information paradox" (Liang et al., 2008). In response to this criticism, Zellner (2008) argued that a Bayesian procedure which places a high posterior model probability (but not equal to one), even on a limiting perfectly fitted model, is a reasonable answer, in line with the philosophy of Box ("all models are wrong"), and with Jeffreys (1961) who claimed that there is always an infinite number of models that can perfectly fit the data. Finally, the posterior model probability eventually converges to one as the sample size increases, which again is a plausible behavior because uncertainty progressively reduces as data information is accumulated.

Using power prior setups, extensions of g-priors have been introduced for binary response models (Ntzoufras et al., 2003; Fouskakis et al., 2009), for generalized linear models (Sabanés Bové and Held, 2011) and, more recently, for zero-inflated Poisson models (Malesios et al., 2017).

Mixtures of g-priors A natural extension of g-priors can be obtained by considering a hyper-prior $\pi(g)$ in order to "let the data decide" about the value of g. Although Zellner (1986) had already suggested such an extension, no solid scientific arguments existed before the work of Liang et al. (2008), which justified theoretically the use of hyper-priors. Since g is nothing but the power parameter as described in the previous paragraph, any mixture of g-priors can be considered as a power-prior with fixed imaginary data and a hyper-prior placed on a_0, that controls the amount of prior information which is fed into the posterior.

Within the normal linear regression model formulation, Cui and George (2008) and Liang et al. (2008) introduce in (15) the hyper-g prior which places a beta prior on the shrinkage parameter $g/(g+1)$ with hyperparameters 1 and $a/2-1$, leading to a mean equal to $2/a$. The induced hyperprior for g has density function $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}$, for $g > 0$. Liang et al. (2008) suggested the value of $a = 4$ (uniform prior), or $a = 3$ with prior mean shrinkage equal to 2/3. Another sensible choice is $a = 2(1 + 1/n)$, so that $\mathbb{E}[g/(g+1)] = n/(n+1)$, which corresponds to the shrinkage of the unit-information setup of the g-prior (i.e. for $g = n$). Generally, any choice $2 < a \leq 4$ leads to robust answers (Dellaportas et al., 2012) except for choices extremely close to 2 which eventually activate the Jeffreys–Lindley–Bartlett paradox. A practical disadvantage of the hyper-g variable selection method is that, for non-important covariates, it results in posterior covariate inclusion probabilities which are inflated towards 1/2 in comparison with other methods; for examples and discussion see Dellaportas et al. (2012).

Under the hyper-g prior, the induced variable selection method is consistent in terms of prediction, model selection (*C2*) for any true model except the null, and information consistent (*C3*). Model selection consistency under the null is achieved under the hyper-g/n prior, whose density is $\pi(g) = \frac{a-2}{2n}(1+g/n)^{-a/2}$, for $g > 0$. Alternatively, one can consider the reparametrization $g = ng^*$ and place a hyper-g prior on g^*. The reciprocal of the variance multiplier $1/g^* = n/g$ measures the units of information in data points added in the analysis *via* the prior. Under this parametrization, a $Beta(1, a/2-1)$ prior is assigned to the factor $g^*/(g^*+1) = g/(g+n)$. In a similar manner, Ley and Steel (2012) use a Beta distribution with hyperparameters b and c on $g/(g+n)$ (they also use a more specific horseshoe type of prior for the same shrinkage factor). Computations in normal linear regression models are relatively straightforward because the marginal likelihoods involved in all model comparisons require the computation of one-dimensional integrals.

Mixtures of g-priors include the Cauchy prior of Zellner and Siow (1980) which can be re-expressed as a mixture of g-priors with an inverse gamma hyper-prior with parameters $1/2$ and $n/2$ (Liang et al., 2008), the approaches by Maruyama and George (2011) and George and Maruyama (2014), and the robust prior of Bayarri et al. (2012). Maruyama and George (2011) propose to use a Beta-prime distribution for g under which $g/(1+g)$ has a Beta prior with hyperparameters b and c and proposed values $c = 1/4$ and $b = (n - p_\ell - 1)/2 - (1-c)$ for model M_ℓ when the number of covariates p_ℓ is lower than $n-1$. Therefore, this prior uses model specific hyperparameters: a feature that was not adopted in the original formulation of Liang et al. (2008).

Extensions to generalized linear models have been introduced by Sabanés Bové and Held (2011), Li (2013) and by Li and Clyde (2016), where calculations of the posterior probabilities can be based on Laplace approximations or on trans-dimensional MCMC methods. Additional articles related to mixtures of g-priors include the work of Malesios et al. (2017) in which hyper-g variable selection is implemented for zero-inflated Poisson epidemic models for sheep-pox incidences, and the work of Sabanés Bové et al. (2015) where they implement hyper-g priors in generalized additive models with penalized splines. Mukhopadhyay and Minerva (2017) propose a mixture of g-priors for variable selection when the number of regressors increases with the sample size. Som et al. (2015) introduce the block hyper-g priors in order to avoid undesirable behaviors appearing when one coefficient is much larger than the rest. Wetzels et al. (2012) apply the hyper-g priors in ANOVA designs while Wang (2017) study the behavior of hyper-g priors on ANOVA models when the number of parameters is growing with the sample size.

Building on the seminal ideas of Jeffreys (1961) and with the goal to generalize the priors developed by Zellner and Siow (1980), Bayarri and García-Donato (2008) propose divergence based (DB) priors for general testing purposes in an objective framework. A DB prior for the comparison of two models is a function of a unitary symmetrized Kullback–Leibler divergence between the two models. This function is chosen so that the resulting prior has a desirable tail behavior. They apply their methodology in challenging scenarios such as irregular models and mixture models, showing that DB priors are well defined and enjoy appealing properties.

Random imaginary data

We proceed with the more recent introduction of prior distributions that treat imaginary data as stochastic components. The idea was independently introduced by Pérez and Berger (2002) and Neal (2001), while the power version of this prior was later introduced by Fouskakis et al. (2015) in order to alleviate the amount of information introduced by the size of the training dataset.

Expected posterior priors Pérez and Berger (2002) have developed priors for Bayesian hypothesis testing, through the utilization of the device of "imaginary training samples" (Good, 1950; Spiegelhalter and Smith, 1980; Iwaki, 1997). The expected posterior prior (EPP) for the parameter under a given model is the expectation of the posterior distribution given imaginary observations $\boldsymbol{y}^*$ of size n^*, where the expectation is taken with respect to a suitable probability measure $m^*(\boldsymbol{y}^*|M_*)$ under a reference model M_*, while the posterior distribution is computed *via* Bayes's theorem starting from a baseline, typically improper, prior. Specifically, consider model M_ℓ with distribution $f(\cdot|\boldsymbol{\theta}_\ell, M_\ell)$ and baseline prior $\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$. The EPP is given by

$$\pi^{EPP}(\boldsymbol{\theta}_\ell|M_\ell) = \int \pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, M_\ell) m^*(\boldsymbol{y}^*|M_*) d\boldsymbol{y}^*, \tag{16}$$

where $\pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, M_\ell) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$ is the posterior distribution of $\boldsymbol{\theta}_\ell$ under model M_ℓ conditionally on the imaginary data $\boldsymbol{y}^*$ for the given baseline prior $\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$. Consider now the comparison of several models having the same structure. There will typically exist a model M_0 which is nested into each of the remaining models (the simplest model). In this case setting M_* to M_0 is a reasonable choice, under the "local" principle described previously in this section. Accordingly $m^*(\boldsymbol{y}^*|M_*)$ will be the prior-predictive distribution under M_0, namely

$$m^*(\boldsymbol{y}^*|M_*) = m^N(\boldsymbol{y}^*|M_0) = \int f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, M_0)\pi^N(\boldsymbol{\theta}_0|M_0) d\boldsymbol{\theta}_0, \tag{17}$$

where $f(\cdot|\boldsymbol{\theta}_0, M_0)$ is the distribution under model M_0, with model specific parameter $\boldsymbol{\theta}_0$ and $\pi^N(\boldsymbol{\theta}_0|M_0)$ is the baseline prior under M_0. Notice that $m^*(\boldsymbol{y}^*|M_*)$ may be improper; this will occur in (17) whenever $\pi^N(\boldsymbol{\theta}_0|M_0)$ is improper. If $M_* = M_0$, then it is straightforward to show that the EPP for the parameter $\boldsymbol{\theta}_\ell$ reduces to the intrinsic prior for nested model comparison because

$$\begin{aligned} \pi^{EPP}(\boldsymbol{\theta}_\ell|M_\ell) &= \pi^N(\boldsymbol{\theta}_\ell|M_\ell) \int \frac{m^N(\boldsymbol{y}^*|M_0)}{m^N(\boldsymbol{y}^*|M_\ell)} f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell) d\boldsymbol{y}^* \\ &= \pi^N(\boldsymbol{\theta}_\ell|M_\ell) \mathbb{E}_{\boldsymbol{Y}^*|\boldsymbol{\theta}_\ell, M_\ell} \left[\frac{m^N(\boldsymbol{Y}^*|M_0)}{m^N(\boldsymbol{Y}^*|M_\ell)} \right] = \pi^I(\boldsymbol{\theta}_\ell|M_\ell). \end{aligned}$$

Additionally, it is immediate to verify that $\pi^{EPP}(\boldsymbol{\theta}_0|M_0) = \pi^N(\boldsymbol{\theta}_0|M_0)$, so the EPP and the intrinsic prior for $\boldsymbol{\theta}_0$ also coincide. Pérez and Berger (2002, Eq. 2.1) provide conditions for the existence of the EPP; namely that $\pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, M_\ell)$ is proper and that the expectation in (16) is positive and finite.

EPPs offer the same advantages of intrinsic priors, among which: i) impropriety of baseline priors causes no indeterminacy in the resulting Bayes factor; ii) an effective way of establishing compatibility of priors across models, as already mentioned in Section 3.3, because all priors are anchored to the same baseline measure $m^*(\cdot)$. On the downside, EPPs rely on features of the imaginary training sample, such as the size n^*, or, in variable selection problems, the choice of the imaginary design matrices $\boldsymbol{X}_\ell^*$ for each competing model. The selection of a *minimal* training sample size n^* has been proposed (Berger and Pericchi, 2004), to make the information content of the prior as small as possible, and this is an appealing idea. But even under this setup, the resulting prior can be influential when the sample size n is not much larger than the total number of parameters under the full model; see Fouskakis et al. (2015) for a discussion of the difficulties associated with the implementation of the EPP with particular reference to variable selection.

Under the variable selection problem in normal linear regression models, Womack et al. (2014) and Fouskakis et al. (2017a) show that the EPP prior, using M_0 as the reference model, minimal training sample of size $n^* = p_\ell + 2$ and default baseline priors, can be expressed as a mixture of g-priors

$$\pi^{EPP}(\boldsymbol{\beta}_{\ell\backslash 0}, \boldsymbol{\beta}_0, \sigma \mid M_\ell) \propto \sigma^{-1} \int_0^1 N_{p_\ell - p_0}\big(\mathbf{0}, \tfrac{1}{t}\boldsymbol{\Sigma}^*_{\ell\backslash 0}\big)\mathrm{Beta}\big(t \mid \tfrac{1}{2}, \tfrac{1}{2}\big)dt, \tag{18}$$

where $\mathrm{Beta}(t|a,b)$ denotes the density of the Beta distribution with parameters a and b evaluated at t, $\boldsymbol{\Sigma}^*_{\ell\backslash 0} = \sigma^2(\boldsymbol{V}^{*T}_{\ell\backslash 0}\boldsymbol{V}^*_{\ell\backslash 0})^{-1}$, $\boldsymbol{V}^*_{\ell\backslash 0} = (\boldsymbol{I}_n - \boldsymbol{X}^*_0(\boldsymbol{X}^{*T}_0\boldsymbol{X}^*_0)^{-1}\boldsymbol{X}^{*T}_0)\boldsymbol{X}^*_{\ell\backslash 0}$, $\boldsymbol{X}^*_0$ is an $(p_\ell + 2) \times (p_0 + 1)$ imaginary design matrix under model M_0 and $\boldsymbol{X}^*_\ell = [\boldsymbol{X}^*_0, \boldsymbol{X}^*_{\ell\backslash 0}]$ is an $(p_\ell + 2) \times (p_\ell + 1)$ imaginary design matrix under model M_ℓ. Imaginary design matrices are formed by suitably subsetting the original full imaginary design matrix.

Power expected posterior priors Fouskakis et al. (2015) and Fouskakis and Ntzoufras (2016b) introduced the power-expected-posterior (PEP) prior and the power-conditional-expected-posterior (PCEP) prior respectively, as generalized versions of the EPPs by combining ideas from the power prior method of Ibrahim and Chen (2000) and the unit information prior approach of Kass and Wasserman (1995). The goal is to produce a minimally informative prior, and at the same time to diminish the effect of training samples within the EPP methodology. In practice, the PEP methodology is sufficiently insensitive to the size n^* of the training sample, because PEPs are constructed using unit information ideas, so that one may even take $n^* = n$.

Under the PEP methodology, as a first step, the likelihoods involved in the EPP distribution are raised to the power $\frac{1}{\delta}$ ($\delta \geq 1$) and then they are density-normalized. The power parameter δ could be set equal to n^*, to represent information equal to one data point. For $\delta = 1$ the PEP prior is equivalent to the EPP. Regarding the size n^* of the training sample, Fouskakis et al. (2015) set it equal to n; this choice gives rise to significant advantages, for example for the variable selection problem it leads to setting the imaginary design matrix equal to the observed one, and therefore the selection of a training sample of covariates and its effects on the posterior model comparison is avoided, while still holding the prior information content equivalent to one data point.

Here is an outline of the PEP method. Suppose we wish to compare model M_0 and M_ℓ with M_0 nested in M_ℓ. Assuming $M_* = M_0$, the PEP prior is defined by the following equation

$$\pi^{PEP}(\boldsymbol{\theta}_\ell|\delta, M_\ell) = \int \pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, \delta, M_\ell) m^N(\boldsymbol{y}^*|\delta, M_0) d\boldsymbol{y}^*, \tag{19}$$

with

$$\begin{aligned} \pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, \delta, M_\ell) &\propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, \delta, M_\ell)\pi^N(\boldsymbol{\theta}_\ell|M_\ell), \\ f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, \delta, M_\ell) &= \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)^{1/\delta}}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)^{1/\delta} d\boldsymbol{y}^*}, \\ m^N(\boldsymbol{y}^*|\delta, M_0) &= \int f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, \delta, M_0)\pi^N(\boldsymbol{\theta}_0|M_0) d\boldsymbol{\theta}_0 \ . \end{aligned}$$

When the density normalized power likelihood is not a distribution of a known form, one can resort to a suitable extension of the above method, as illustrated in Fouskakis et al. (2017b).

Under the variable selection problem in normal linear regression models Fouskakis et al. (2017a) show that the PEP prior, using M_0 as the reference model, a training sample size equal to n, the default baseline priors and $\delta = n$, can be expressed as a mixture of g-priors

$$\pi^{PEP}(\boldsymbol{\beta}_{\ell\setminus 0}, \boldsymbol{\beta}_0, \sigma \big| M_\ell) \propto \sigma^{-1} \int_0^1 N_{p_\ell - p_0}\big(\mathbf{0}, \tfrac{n}{t}\boldsymbol{\Sigma}_{\ell\setminus 0}\big) \mathrm{Beta}\big(t \,\big|\, \tfrac{n-p_\ell-1}{2}, \tfrac{n-p_\ell-1}{2}\big) dt \ , \tag{20}$$

where $\boldsymbol{\Sigma}_{\ell\setminus 0}$ is defined in analogy with $\boldsymbol{\Sigma}^*_{\ell\setminus 0}$ in (18) based on the sample design matrix.

Empirical Bayes approaches

Empirical Bayes (EB) approaches have been traditionally used to alleviate prior elicitation in multi-parameter setups (e.g. hierarchical models) by settings some prior hyperparameters equal to the corresponding sample estimates. The main criticism against EB is the obvious double use of the data which violates a basic principle of Bayesian theory. This can however be mitigated by combining EB with other ideas described in the previous section, such as the unit information principle, in order to minimize the re-use of the data especially in cases when the sample size is not large.

EB methods in model selection usually focus on the specification of the prior for a small number of parameters, typically those causing the sensitivity of the Bayes factor. Estimates of hyperparameters are obtained either by maximizing a suitable integrated likelihood, see for example George and Foster (2000), or by controlling the false discovery rates (Tansey et al., 2018). With regard to the variable selection problem, EB methods have been used to specify (a) the parameter g in the g-prior (George and Foster, 2000; Liang et al., 2008); (b) the prior inclusion probability (George and Foster, 2000; Scott and Berger, 2010; Castillo and Misner, 2018); (c) the shrinkage parameter under the lasso setting (Yuan and Lin, 2005).

Finally we note that empirical versions of EPP and PEP can be produced by using the empirical distribution of the actual data to specify the predictive measure under the reference model, see for example Pérez and Berger (2002).

Non-local priors

Recall that criterion *C7* described in Section 3.3 can be understood as a formalization of Jeffreys' criterion for comparing nested models. This says that the prior for the specific parameter of the larger model (the alternative hypothesis) should be "centered at the simplest model". In practice this has been implemented by assigning a continuous prior having mode at the parameter value specified by the null model. These type of priors are called *local priors*. On the other hand, Johnson and Rossell (2010) proposed the use of *non-local priors* in order to improve convergence rates in favor of the true null hypothesis. Such priors have densities which vanish on the null subspace. Example of such priors are the moment prior and the inverse moment prior; see for details Johnson and Rossell (2010). In a discussion of Consonni and La Rocca (2011), Rousseau and Robert suggest to cast the testing problem in a decision-theoretic setup and use the well-known duality between prior and loss function (Rubin, 1987) to replace non-local priors with suitable loss functions that take into account the distance from the null.

3.5 Comparison of priors for Bayesian variable selection in normal linear models

For the variable selection problem in normal linear regression models, most of the priors discussed in the previous sections can be expressed as mixtures of g-priors. Table 1 provides a summary. Save for the first three, all the remaining priors are mixtures of g-priors. Moreover, with the exception of the EPP and Maruyama and George prior, they can be written in the general form of the robust prior (10) with $\pi^R(g)$ replaced by a specific distribution as detailed in Table 1. The robust prior fulfills all the desiderata of Bayarri et al. (2012). Regarding the rest of the priors in Table 1, we have the following results with respect to the seven criteria.

- All priors satisfy the basic criterion (*C1*).
- All priors lead to consistent model selection procedures (criterion *C2*); for the g-prior see Fernández et al. (2001); for the Cauchy, the hyper-g and hyper-g/n see Liang et al. (2008) (with the hyper-g only to suffer from model selection inconsistency when the true model is the null model); for the Maruyama and George prior see Maruyama and George (2011); for the EPP see Casella et al. (2009) and finally for the PEP prior see Fouskakis et al. (2015) and Fouskakis and Ntzoufras (2016a).
- Liang et al. (2008) showed that the g-prior suffers from information inconsistency; while the Cauchy, the hyper-g and hyper-g/n priors satisfy the criterion *C3* of information consistency. Finally, Fouskakis and Ntzoufras (2017) proved that model selection under PEP is free from information inconsistency.

Prior distribution (reference)	Value of g	
Unit Information g-Prior (Zellner, 1986; Kass and Wasserman, 1995)	n	
Risk Inflation Criterion Prior (Foster and George, 1994)	p^2	
Benchmark Prior (Fernández et al., 2001)	$\max\{n, p^2\}$	
Prior distribution (reference)	**Prior $\pi^R(g) \propto$**	
Cauchy Prior (Zellner and Siow, 1980)	$g^{-3/2}e^{-n/2g}$	
Hyper-g-Prior (Liang et al., 2008)	$(1+g)^{-a/2}$,	$a > 2,\ g > 0$
Hyper-g/n-Prior (Liang et al., 2008)	$(1+g/n)^{-a/2}$,	$a > 2,\ g > 0$
Maruyama and George Prior (Maruyama and George, 2011)	$g^b(1+g)^{-(a+b+2)}$,	$a > 1,\ b > -1,\ g > 0$
Robust Prior (Bayarri et al., 2012)	$(b+g)^{-(a+1)}$,	$a, b > 0,\ g > \frac{(b+n)}{\rho_\ell^{-1}} - b,\ \rho_\ell \geq \frac{b}{b+n}$
EPP ($n^* = p_\ell + 1$) (Pérez and Berger, 2002)	$g^{-1}(g-1)^{-1/2}$,	$g > 1$
PEP-Prior ($n^* = n = \delta$) (Fouskakis et al., 2015)	$g^{-(n-p_\ell-1)} \times (g-n)^{(n-p_\ell-1)/2-1}$,	$g > n$

Table 1: Different mixtures of g-priors.

- All priors in Table 1 belong to a more general class of conditional priors

$$\pi(\boldsymbol{\beta}_0, \boldsymbol{\beta}_{\ell\setminus 0}, \sigma | M_\ell) \propto \sigma^{-1-(p_\ell - p_0)} h\Big(\frac{\boldsymbol{\beta}_{\ell\setminus 0}}{\sigma} \Big| M_\ell\Big), \tag{21}$$

  where $h(\cdot|M_\ell)$ is a proper density with support $\mathbb{R}^{p_\ell - p_0}$. Bayarri et al. (2012) prove that the predictive matching criterion ($C5$) and the group invariance criterion ($C7$) hold if the priors are of the form (21) with $h(\cdot|M_\ell)$ symmetric around zero. Further results on matching properties apply by specializing (21).

3.6 Objective priors on model space

Within the $\mathcal{M}$-closed view of model selection (i.e. the true model is included in $\mathcal{M}$), the default choice to express ignorance or indifference between two or more models under comparison was, for many years, the uniform distribution on the model space $\mathcal{M}$, that is $\pi(M_\ell) = 1/|\mathcal{M}|$ for all $M_\ell \in \mathcal{M}$, where $|\mathcal{M}|$ denotes the cardinality of $\mathcal{M}$. For variable selection problems, letting p denote the potential number of predictors beyond those which must be present in all models, the uniform prior distribution $\pi(M_\ell) = 2^{-p}$ is obtained by assuming that each predictor enters the model independently with inclusion probability 1/2. In recent years, this choice has become progressively less popular, because it does not account for structural features, notably sparsity, dimensionality, and collinearity of predictors. In particular Chipman et al. (2001) and George (2010) discuss how to construct dilution priors which are uniform over neighborhoods of models which are regarded to be similar according to some criterion. Scott and Berger (2010) argue that prior model probabilities should take into consideration multiplicity issues inherent in model comparisons. When applied to variable selection problems, this principle can

be implemented by assuming that, conditionally on a random probability of inclusion ω, each predictor can enter a model independently, so that $\pi(M_\ell|\omega) = \omega^{p_\ell}(1-\omega)^{n-p_\ell}$. Next, a hyper-prior is assigned to ω; in particular if $\omega \sim Beta(a_\omega, b_\omega)$, the resulting prior becomes

$$\pi(M_\ell) = \frac{B(a_\omega + p_\ell, b_\omega + p - p_\ell)}{B(a_\omega, b_\omega)}, \tag{22}$$

which is commonly known as the beta-binomial prior on model space. The default choice $a_\omega = b_\omega = 1$ results in a uniform distribution for ω. Under this specification, (22) reduces to

$$\pi(M_\ell) = \frac{1}{p+1}\binom{p}{p_\ell}^{-1}, \tag{23}$$

which induces a uniform prior on model size:

$$\pi\big(\{M_\ell \in \mathcal{M} : p_\ell = d\}\big) = 1/(p+1) \text{ for } d = 0, 1, \ldots, p.$$

The choice of a uniform prior on ω provides more support to individual models having either low or high dimensionality and does not penalize for complexity. Wilson et al. (2010) propose $a_\omega = 1$ and $b_\omega = \lambda p$, where λ is a positive constant, resulting in a prior on model-dimension having expectation $1/\lambda$, and a behavior similar to a geometric distribution for low values of the dimension. This prior also corresponds to an approximate penalization equal to $\log(\lambda + 1)$ in log-odds scale for each additional covariate added to the model.

Castillo et al. (2015) investigate high-dimensional linear regression models under sparsity constraints. Conditionally on the size of the set of predictors, the prior on the regression parameter is a mixture of point masses at zero and continuous distributions. Assuming the prior and the design matrix satisfy some conditions, they show a variety of contraction properties for the posterior distribution; including the correct selection of at least the coefficients that are significantly different from zero. Further results of their approach are reported in Section 4. Womack et al. (2015) take a geometric approach, and argue, using isometry considerations on model space, that the appropriate distribution on model size is a truncated Poisson, while the prior probability of models having the same size is uniform. This provides a consistent model selection procedure. Another usual way to specify Bayesian procedures which account for multiple testing is via the control of false discovery rate (FDR); see for example in Storey (2003).

We close this section with two alternative treatments of the specification of the prior on the model space. The first approach, introduced by Dellaportas et al. (2012), argues that we should *jointly* specify the prior on the model parameters and the model space; see Robert (1993) for related ideas. The key point is that, by relating the two aspects, sensitivity of posterior model probabilities to the prior variance of the model coefficients can be avoided by suitable specification of prior model probabilities $\pi(M_\ell)$, $M_\ell \in \mathcal{M}$. For example in the g-prior setup it is straightforward to see that setting $\pi(M_\ell) \propto g^{(p_\ell+1)/2}$ in (14) or $\pi(M_\ell) \propto g^{p_\ell/2}$ in (15) will eliminate any dependence of the posterior model probability $\pi(M_\ell|\boldsymbol{y})$ on the prior variance multiplier g. To illustrate the method, consider the modified g-prior specification (15), conditional on the intercept

and error variance. Dellaportas et al. (2012) propose to use prior model probabilities with the structure

$$\pi(M_\ell) \propto p(M_\ell) \left(\frac{g}{n}\right)^{\frac{p_\ell}{2}} \; M_\ell \in \mathcal{M},$$

where $p(M_\ell)$ is some baseline model weight, and should reflect prior features of the model not related to the prior distribution on the model parameters, such as model dimension or complexity, or sparsity preferences. They note that setting $p(M_\ell) \propto 1$ will result in posterior model probabilities "which are asymptotically equivalent to those implied by BIC". Alternative choices of $p(M_\ell)$ can be obtained by matching the log-posterior model probabilities to suitable information criteria, although $p(M_\ell)$ should not change according to the sample size. The approach based on the joint specification on model and parameter spaces not only avoids the sensitivity of the posterior model probabilities to the prior uncertainty of model parameters, but also produces Bayesian model averaging estimators which do not suffer from the Jeffreys-Lindleys-Bartlett paradox.

The second approach to the specification of prior model probabilities is proposed by Villa and Walker (2015b) and it is strictly related to the method for obtaining objective prior in models with discrete parameter space, already discussed in Section 2. The basic idea is that each model M_ℓ has a *worth*, which only depends on how "close" in KL-divergence M_ℓ is to its nearest neighbor in the collection of models under consideration (the smaller the divergence, the smaller the worth, because it means that M_ℓ can be excluded with a small loss). Since the worth depends on no other considerations, the method can claim to fall within the objective methodology. This leads to the following specification

$$\pi(M_\ell) \propto \exp\left(\mathbb{E}_{\boldsymbol{\theta}_\ell | M_\ell}\left\{ \inf_{\boldsymbol{\theta}_k,\, k \neq \ell} D_{KL}(f(\boldsymbol{y}|\boldsymbol{\theta}_\ell, M_\ell) \,||\, f(\boldsymbol{y}|\boldsymbol{\theta}_k, M_k)) \right\}\right), \tag{24}$$

where D_{KL} is the KL-divergence, see Section 2. This approach has been illustrated in a variety of simple model comparisons (nested and non-nested) in Villa and Walker (2015b), and in Villa and Walker (2017) for the testing setup described in Lindley (1957). Villa and Lee (2015) have extended the method for variable selection in normal linear regression models. In such problems, (24) is proportional to one, for all models, which induces the uniform prior on model space. To resolve this issue, Villa and Lee (2015) introduced an additional loss function based on the dimensionality/complexity of the model.

Finally, Spitzner (2011) introduced the idea of "neutral" data which support neither of the two hypothesis/models under consideration. This idea can be naturally accommodated for the construction of "objective" priors on the model space.

4 High-dimensional models

Current applications of statistical methods often deal with high-dimensional models, wherein the derivation of an objective prior, defined according to a well established formal rule, like Jeffreys' or reference prior, is virtually impossible; see also Section 2. In regression settings, common default priors such as the g-prior and its extensions to

random g, are not defined when the number of predictors p is larger than the sample size n, save for the generalized g-prior of Maruyama and George (2011). The "robust" prior of Bayarri et al. (2012) suffers from the same problem because it requires the existence of the maximum likelihood estimator for each model under consideration. Similarly the intrinsic, or more generally the Expected Posterior prior (EPP), methodology would require a training sample size n^* bigger than n. This means that the training design matrix $\boldsymbol{X}^*$ should be taller than the observed $\boldsymbol{X}$ matrix, with extra rows that would need to be fixed exogenously. This raises inevitable concerns for the OB approach, although they could be mitigated through a suitable discounting factor within the PEP methodology. More generally, high-dimensional problems pose new challenges that need be addressed through novel methodologies.

1. *Sparsity.* Consider the sparse normal means problem, that is

$$y_i|\theta_i, \sigma \overset{\text{ind}}{\sim} N(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \tag{25}$$

   where n is typically very large. Let $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0n})$ be the true mean value. Under sparsity, in the near-black sense, the number p_n of θ_{0i}'s different from zero (signals) is allowed to grow with n but at a slower rate, so that $p_n = o(n)$. The goal is estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, distinguishing signal from noise.

2. *Shrinkage.* Bayesian methods are ideally suited for creating suitable shrinkage in many dimensions, as has been recognized for many decades, starting from the seminal work of Stein (1956). Indeed sparsity and shrinkage, though distinct, are closely related as we look for priors that do shrink strongly on noise components. On the other hand, strong signals should be clearly picked-up, and model estimates of the corresponding parameters should undergo negligible shrinkage. Priors which achieve this goal are often named, in this context, *robust*.

3. $p >> n$ situations. High-dimensionality often means that the number of parameters p exceeds the sample size n, a situation which is routinely found today in many applications. Improper priors cannot deal with these cases, and accordingly suitable proper priors need be developed.

A large body of research has been deployed to develop default proper priors for high dimensional models. Typically the performance of these priors is assessed in relation to: 1) computational efficiency; 2) frequentist assessment, especially in terms of the speed of concentration of the posterior parameter distribution, or functionals thereof, to the true value, and in terms of coverage of credible sets; 3) ease of interpretation, so that tuning hyperparameters (when present) can be readily set in specific applications.

The number of papers dealing with the above topics has literally mushroomed in the last decade, and we cannot even try to provide a reasonably exhaustive review of the various contributions. Accordingly, we shall merely present a highly selective account in order to provide the interested reader with some useful signposts. A general point to make is that, in these situations, the typical use of proper priors makes the distinction between

objective priors for estimation and testing redundant. Most of the proposals can be collected under two broad categories: 1) *spike-and-slab* priors and 2) *global-local* priors.

The spike-and-slab prior (George and McCulloch, 1993) for θ_i is a two-point mixture of distributions, one being absolutely continuous and heavy-tailed (the slab), and the other a Dirac measure at zero. More formally, conditionally on a latent binary random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$, one has

$$p(\theta_i|\gamma_i, \lambda_1) = (1 - \gamma_i)\delta_0(\theta_i) + \gamma_i \psi(\theta_i|\lambda_1), \tag{26}$$

where $\delta_0(\cdot)$ is the Dirac delta function while $\psi(\cdot|\lambda_1)$ is the slab distribution possibly depending on a fixed hyperparameter λ_1. The latent vector $\boldsymbol{\gamma}$ in turn is assigned a distribution $\pi(\boldsymbol{\gamma}|\nu)$. Castillo and van der Vaart (2012) show that, under the prior (26) and a suitably chosen value for ν, or a suitable beta-prior $\pi(\nu)$, the whole posterior distribution concentrates on the true value at the minimax rate. The same result holds for several posterior estimators, under a convex loss, targeted to both location and spread parameters. Castillo et al. (2015) provide contraction results in a Gaussian regression setup under a family of joint distributions for the size of the active covariates (signals) and the regression parameter which includes the spike-and-slab prior. A remarkable result is that the product of Laplace priors for the individual regression coefficients, whose mode is the popular lasso estimator, produces a posterior distribution which fails to contract at the same speed as the mode.

Several elaborations of (26) have been considered, with special emphasis on continuous relaxations, that is replacing $\delta_0(\cdot)$ with a peaked continuous density (George and McCulloch, 1993; Ishwaran and Rao, 2005). The motivation is twofold: to enhance flexibility and to make the ensuing Bayesian analysis amenable to fast deterministic computation (Ročková and George, 2014). In particular, Ročková and George (2018) introduce the spike-and-slab lasso (SS-LASSO) prior where both components of the mixture are Laplace distributions, so that the resulting prior can be viewed as a compromise between the theoretical benchmark (26) and the (computationally convenient) single Laplace prior. A thorough theoretical evaluation of the SS-LASSO priors is undertaken in Ročková (2018), where connections with current penalized likelihood methods are established in order to enhance interpretation, and risk results are proved for estimators not only of functionals of the posterior distribution of θ_i (especially the mode) but, importantly, for the whole posterior distribution. Castillo and Misner (2018) provide convergence results of the posterior distribution associated to a variety of spike and slab prior distributions when the key sparsity hyperparameter is calibrated via marginal maximum likelihood empirical Bayes.

An alternative approach, which is easy to implement using generic sampling tools, and is typically fully automatic, is represented by continuous scale mixture priors. Among the many existing proposals, and limiting ourselves to the general set-up exhibited in (25), we mention the normal-exponential-gamma prior (Griffin and Brown, 2010), and the very popular *horseshoe prior* (Carvalho et al., 2010; Polson and Scott, 2012b) which is hierarchically specified as

$$\theta_i|\lambda_i, \tau \overset{\text{ind}}{\sim} N(0, \sigma^2\tau^2\lambda_i^2), \quad \lambda_i|\tau, \overset{\text{iid}}{\sim} \text{Cauchy}^+(0, \tau), \quad \tau \sim g(\tau), \quad i = 1, \ldots, n; \tag{27}$$

that is the θ_i's are conditionally independent given the local parameters λ_i's, which in turn are conditionally i.i.d. given the global parameter τ. An interesting representation of the above priors is obtained by considering $\kappa_i = (1+\tau^2\lambda_i^2)^{-1}$, $i = 1, \ldots, n$. Then the marginal posterior mean of θ_i, conditionally on τ, is

$$\mathbb{E}\,(\theta_i|y_i, \tau) = y_i - \mathbb{E}\,(\kappa_i|y_i, \tau)\, y_i. \tag{28}$$

Thus $\kappa_i \in [0, 1]$ operates as a local shrinkage factor for the i-th component of the model. On the other hand τ acts as a global parameter. The horseshoe prior is thus a *global-local* shrinkage prior because it is able to combine robustness control on the tails as well as sparsity. The resulting conditional prior for κ_i has a U-shape, depending on τ, whence the name *horseshoe* given to the entire prior structure.

The horseshoe prior approach has to be completed with the choice of a prior on τ. This is the most sensitive issue and no clear default choices exist, although the common proposal is to adopt a half-Cauchy prior (Polson and Scott, 2012a). This issue is deeply discussed in Piironen and Vehtari (2017a, 2017b), who propose an intuitive way of formulating the prior for τ based on prior assumptions on the effective number of nonzero parameters. Further elaborations on horseshoe priors are provided in Polson and Scott (2012a), Polson and Scott (2012b) and Bhadra et al. (2016).

The frequentist properties of the horseshoe priors have been analyzed in a series of papers; see for instance Datta and Ghosh (2013) who consider the asymptotic properties of the multiple testing rule induced by the estimator (28), and van der Pas et al. (2017) who consider the frequentist coverage of posterior intervals of the location parameters, and discuss the irreconcilability between adaptivity and honesty when the level of sparsity is unknown.

In a manner similar to the horseshoe prior, but with the aim of studying the posterior asymptotic behavior (in particular contraction rates) of the joint vector $(\theta_1, \ldots, \theta_n)$, Bhattacharya et al. (2015) have proposed a novel class of global-local shrinkage priors, named Dirichlet-Laplace, defined as

$$\begin{aligned} &\theta_i|\varphi_i, \tau \overset{\text{ind}}{\sim} \text{Double Expon}(\tau\varphi_i), \quad (\varphi_1, \ldots \varphi_n) \sim \text{Dirichlet}(a, \ldots a), \\ &\qquad \tau \sim g(\tau), \quad i = 1, \ldots, n. \end{aligned} \tag{29}$$

Compared with (27) with $\sigma = 1$, the Dirichlet-Laplace prior models independently the global parameter τ and the local parameters φ_i's.

An alternative way to modeling, with proper priors, the scale parameters in a hierarchical setting, is given in Pérez et al. (2017). Instead of assuming the usual conjugate inverse Gamma or the half-Cauchy (Gelman, 2006), the authors suggest to consider a Gamma mixture of Gamma densities, which is named Scaled Beta2 (SB2). It was previously derived in Girón et al. (2006) as an intrinsic prior for the scale parameter in a linear model. The two parameters of the mixing Gamma determine the behavior of the marginal density around zero and for large values, respectively, and make the SB2 family quite appealing for its flexibility. Additionally, the Cauchy-Scaled Beta2 is shown to represent an explicit horseshoe distribution.

Finally non-local priors can also be represented as mixtures; in this case the mixing parameter is a latent truncation. Rossell and Telesca (2017) thoroughly investigate their behavior in high-dimensional settings showing their good performance both in terms of model selection and estimation.

5 Discussion

Objective Bayesian analysis is here to stay, and so is the search for priors that allow its efficient implementation in a great variety of situations. Although we presented many such priors, we also tried to highlight principles and methods behind them. Paraphrasing a Reviewer of our paper: there is a galaxy of stars (priors) out there, but fortunately we also have categories to study, evaluate and organize them into meaningful systems.

Below we report on a few of outstanding issues which are worth of further consideration.

- **OB priors for estimation and model selection.** This distinction was posited at the very beginning of our review, because the conceptual framework underlying the construction of priors for estimation is different from that leading to priors for model selection, with the latter largely influenced by the approach initiated by Jeffreys (1961); see for instance the *desiderata* illustrated in Section 3.3.

  Consider however a setting where prediction under model uncertainty is the goal, so that model averaging (Hoeting et al., 1996) techniques are employed. In this case one is potentially confronted with two separate priors on the parameter space of the same model: one to determine the model posterior probability, and another one to compute predictions (conditionally on a given model). This dichotomy is however hardly discussed explicitly. Typically the prior employed for model selection is also used to carry out estimation/prediction; see for instance Pérez and Berger (2002, Sect. 6) with regard to expected posterior priors, but the motivation is mostly pragmatic and confined to a specific data analysis. Interestingly, in the area of Bayesian experimental design, it is not uncommon to entertain two distinct priors for the same parameter of a given model, because one distinguishes between a prior for design and a prior for inference; see Han and Chaloner (2004) and earlier references therein.

- **Priors for high-dimensional models.** Our account of this body of research, in this article, is clearly too limited, especially with regard to important technical results on: i) sparsity conditions; ii) assumptions on the priors and features of the underlying model; iii) posterior contraction rates for several notions of recovery of the true model; iv) new computational tools, also alternative to traditional MCMC algorithms. We believe that a review paper devoted to default priors in high-dimensional settings will be a useful gift to the Bayesian community.

  In this connection, a point we would like to raise concerns methods for evaluating the performance of priors in high-dimensions. Currently this is measured in terms of rates of contraction of the posterior distribution (or functionals thereof) to the

underlying true values. Among the *desiderata* that we laid out in Section 3.3, it seems that only properness of the prior and model selection consistency are taken into account. Actually consistency becomes a rather weak property to evaluate priors, while rates with which such consistency is achieved become more crucial. However, as one Reviewer pointed out, insistence only on frequentist properties is open to criticism, as one would like to embrace a "more Bayesian" perspective, possibly along the lines of newly formulated *desiderata*.

- **Computational aspects.** Computation aspects are becoming increasingly important for evaluating any statistical methodology. This is of course the case in high-dimensional settings where scalability of a procedure is an obvious concern. From this perspective, Section 4 does not even come close to providing a reasonably complete account of current technology and trends, although some of the papers we reference contain substantial material on computation; see e.g. Ročková and George (2014) on leveraging the EM algorithm for variable selection. As already hinted above we expect that a full treatment of this topic is better left to a specific review paper.

  On a related point, we note that complex models pose challenges even with regard to traditional objective priors, such as the reference, and often the Jeffreys, priors, which are hard to obtain in a closed form. On the other hand, it is also true that often the exact knowledge of the functional form of the prior is not strictly necessary. Nowadays, the vast majority of applications of Bayesian methods rely on the use of Monte Carlo, or other simulation methods, where the evaluation of the prior, rather than its form, is important. Also, it is often the case that, from a mathematical perspective, the hard step in computing the prior is the evaluation of an expected value. In this context, it is reasonable to include the algorithm for evaluating the prior within the general simulation method. This approach has been discussed in Lafferty and Wasserman (2013), and only sporadically mentioned in other papers (Berger and Sun, 2008; Berger et al., 2009).

- **Priors for model selection based on the *desiderata* of** Bayarri et al. (2012). The general methodology was illustrated in Section 3.3, and in our opinion it represents a major conceptual innovation which deserves to be carefully considered. We still see some outstanding difficulties:

  i) *Non-nested models.* The method is currently predicated on the comparison between two nested models. This of course is not a major drawback if one can find a null model which is nested into every other model under consideration, as we mentioned in Section 3.3. However, when this is not the case, the problem remains open, unless some other forms of encompassing are implemented. Notice that the comparison of non-nested models is also problematic for other more specific approaches, such as the intrinsic, or the EP, prior.

  ii) *Scope.* The implementation of the methodology within normal linear regression models represents a major accomplishment; yet it remains to be seen whether the general idea can be extended to other substantive statistical settings.

References

Altomare, D., Consonni, G., and La Rocca, L. (2013). "Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors." *Biometrics*, 69: 478–487. MR3071066. doi: https://doi.org/10.1111/biom.12018. 644

Arima, S., Datta, G. S., and Liseo, B. (2012). "Objective Bayesian analysis of a measurement error small area model." *Bayesian Analysis*, 7: 363–383. MR2934955. doi: https://doi.org/10.1214/12-BA712. 636

Barbieri, M. and Berger, J. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32: 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 639

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 640, 641, 642, 650, 654, 655, 658, 662

Bayarri, M. J., Berger, J. O., and Pericchi, L. R. (2014). "The Effective Sample Size." *Econometric Reviews*, 33(1–4): 197–217. MR3170846. doi: https://doi.org/10.1080/07474938.2013.807157. 643

Bayarri, M. J. and García-Donato, G. (2007). "Extending conventional priors for testing general hypotheses in linear models." *Biometrika*, 94: 135–152. MR2367828. doi: https://doi.org/10.1093/biomet/asm014. 628

Bayarri, M. J. and García-Donato, G. (2008). "Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70: 981–1003. MR2530326. doi: https://doi.org/10.1111/j.1467-9868.2008.00667.x. 650

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Jones, J. H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L. R., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Zandt, T. V., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017). "Redefine Statistical Significance." *Nature Human Behavior*. 639

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis, 2nd ed.*. Springer, New York. MR0804611. doi: https://doi.org/10.1007/978-1-4757-4286-2. 630

Berger, J. O. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1: 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 628, 632

Berger, J. O. and Bernardo, J. M. (1989). "Estimating a product of means: Bayesian analysis with reference priors." *Journal of the American Statistical Association*, 84: 200–207. MR0999679. 632

Berger, J. O. and Bernardo, J. M. (1992). "Ordered group reference priors with application to the multinomial problem." *Biometrika*, 79: 25. MR1158515. doi: https://doi.org/10.1093/biomet/79.1.25. 632

Berger, J. O., Bernardo, J. M., and Mendoza, M. (1989). "On priors that maximize expected information." In Klein, J. P. and Lee, J. C. (eds.), *Recent Developments in Statistics and their Applications*, 1–20. Seoul: Freedom Academy Publishing. 632

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *Annals of Statistics*, 37: 905–938. MR2502655. doi: https://doi.org/10.1214/07-AOS587. 662

Berger, J. O., Bernardo, J. M., and Sun, D. (2012). "Objective priors for discrete parameter spaces." *Journal of the American Statistical Association*, 107: 636–648. MR2980073. doi: https://doi.org/10.1080/01621459.2012.682538. 632, 633

Berger, J. O., Bernardo, J. M., and Sun, D. (2015). "Overall Objective Priors (with discussion)." *Bayesian Analysis*, 10: 189–221. MR3420902. doi: https://doi.org/10.1214/14-BA915. 632, 636

Berger, J. O., De Oliveira, V., and Sansò, B. (2001). "Objective Bayesian Analysis of Spatially Correlated Data." *Journal of the American Statistical Association*, 96: 1361–1374. MR1946582. doi: https://doi.org/10.1198/016214501753382282. 631

Berger, J. O. and Mortera, J. (1999). "Default Bayes Factors for Nonnested Hypothesis Testing." *Journal of the American Statistical Association*, 94: 542–554. MR1702325. doi: https://doi.org/10.2307/2670175. 645

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 629, 642, 644, 645

Berger, J. O. and Pericchi, L. R. (2001). "Objective Bayesian methods for model selection: Introduction and comparison." In *Model Selection. Institute of Mathematical Statistics Lecture Notes, Monograph Series 38*, 135–207. IMS, Beachwood, OH. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 629, 641

Berger, J. O. and Pericchi, L. R. (2004). "Training Samples in Objective Model Selection." *The Annals of Statistics*, 32: 841–869. MR2065191. doi: https://doi.org/10.1214/009053604000000238. 652

Berger, J. O., Pericchi, L. R., and Varshavsky, J. (1998). "Bayes factors and marginal distributions in invariant situations." *Sankhya*, 60: 109–122. MR1718789. 638

Berger, J. O., Strawderman, W., and Tang, D. (2005). "Posterior Propriety and Admissibility of Hyperpriors in Normal Hierarchical Models." *The Annals of Statistics*, 33:

606–646. MR2163154. doi: https://doi.org/10.1214/009053605000000075. 634, 635

Berger, J. O. and Strawderman, W. E. (1996). "Choice of hierarchical priors: Admissibility in estimation of normal means." *The Annals of Statistics*, 24: 931–951. MR1401831. doi: https://doi.org/10.1214/aos/1032526950. 634

Berger, J. O. and Sun, D. (2008). "Objective priors for the bivariate normal model." *The Annals of Statistics*, 36: 963–982. MR2396821. doi: https://doi.org/10.1214/07-AOS501. 662

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference (with discussion)." *Journal of the Royal Statistical Society B*, 2: 113–147. MR0547240. 628, 632

Bernardo, J. M. and Rueda, R. (2002). "Bayesian Hypothesis Testing: A Reference Approach." *International Statistical Review*, 70: 351–372. 639

Bernardo, J. M. and Smith, A. (1994). *Bayesian Theory*. Chichester, UK: Wiley. MR1274699. doi: https://doi.org/10.1002/9780470316870. 628

Bertolino, F. and Racugno, W. (2000). "Bayesian model selection approach to analysis of variance under heteroscedasticity." 49: 503–517. 646

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969. MR3620450. doi: https://doi.org/10.1093/biomet/asw041. 660

Bhattacharya, A., Pati, D., Pillai, N., and Dunson, D. (2015). "Dirichlet-Laplace priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110: 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 660

Bhattacharya, A., Pati, D., and Yang, Y. (2016). "Bayesian fractional posteriors." *arXiv:1611.01125*. 644

Bodnar, O., Link, A., and Elster, C. (2016). "Objective Bayesian inference for a generalized marginal random effects model." *Bayesian Analysis*, 11: 25–45. MR3447090. doi: https://doi.org/10.1214/14-BA933. 636

Branco, M. D., Genton, M. G., and Liseo, B. (2013). "Objective Bayesian analysis of skew-t distributions." *Scandinavian Journal of Statistics. Theory and Applications*, 40: 63–85. MR3024032. doi: https://doi.org/10.1111/j.1467-9469.2011.00779.x. 636

Brown, L. D. (1971). "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems." *The Annals of Mathematical Statistics*, 42: 855–903. MR0286209. doi: https://doi.org/10.1214/aoms/1177693318. 635

Bush, C. A., Lee, J., and MacEachern, S. N. (2010). "Minimally informative prior distributions for non-parametric Bayesian analysis." *Journal of the Royal Statistical Society B*, 72: 253–268. MR2830767. doi: https://doi.org/10.1111/j.1467-9868.2009.00735.x. 635

Cano, J. and Salmerón, D. (2013). "Integral Priors and Constrained Imaginary Training Samples for Nested and Non-Nested Bayesian Model Comparison." *Bayesian Analysis*, 8: 361–380. MR3066945. doi: https://doi.org/10.1214/13-BA812. 645

Cano, J. A., Kessler, M., and Moreno, E. (2004). "On Intrinsic Priors for Nonnested Models." *TEST*, 13: 445–463. MR2154008. doi: https://doi.org/10.1007/BF02595781. 645

Carvalho, C., Polson, N., and Scott, J. (2010). "The horseshoe estimator for sparse signal." *Biometrika*, 97: 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 659

Carvalho, C. and Scott, J. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96: 497–512. MR2538753. doi: https://doi.org/10.1093/biomet/asp017. 644

Casella, G., Girón, F. J., Martínez, M. L., and Moreno, E. (2009). "Consistency of Bayesian Procedures for Variable Selection." *The Annals of Statistics*, 37: 1207–1228. MR2509072. doi: https://doi.org/10.1214/08-AOS606. 646, 654

Casella, G. and Moreno, E. (2006). "Objective Bayesian Variable Selection." *Journal of the American Statistical Association*, 101: 157–167. MR2268035. doi: https://doi.org/10.1198/016214505000000646. 637, 646

Casella, G. and Moreno, E. (2009). "Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables." *Journal of the American Statistical Association*, 104: 1261–1271. MR2750249. doi: https://doi.org/10.1198/jasa.2009.tm08106. 646

Castillo, I. and Misner, R. (2018). "Empirical Bayes analysis of spike and slab posterior distributions." *arXiv:1801.01696v1*. 641, 653, 659

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 656, 659

Castillo, I. and van der Vaart, A. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40: 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 643, 659

Chen, M., Ibrahim, J. G., and Shao, Q. M. (2000). "Power prior distributions for generalized linear models." *Journal of Statistical Planning and Inference*, 84: 121–137. MR1747500. doi: https://doi.org/10.1016/S0378-3758(99)00140-8. 647

Chib, S. and Kuffner, T. A. (2016). "Bayes factor consistency." *arXiv:1607.00292*. 638

Chipman, H., George, E., and McCulloch, R. (2001). "The practical implementation of Bayesian model selection." *Model Selection, IMS Lecture Notes – Monograph Series*, 38: 67–116. MR2000752. doi: https://doi.org/10.1214/lnms/1215540964. 655

Choirat, C. and Seri, R. (2012). "Estimation in discrete parameter models." *Statistical Science*, 27: 278–293. MR2963996. doi: https://doi.org/10.1214/11-STS371. 634

Clyde, M. and Iversen, E. (2013). "Bayesian Model Averaging in the M-open framework." In *Bayesian Theory and Applications*, in P. Damien, P. Dellaportas, N.G. Polson, and D.A. Stephens, eds., 484–498. Oxford University Press. MR3221178. doi: https://doi.org/10.1093/acprof:oso/9780199695607.003.0024. 638

Consonni, G., Forster, J. J., and La Rocca, L. (2013). "The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data." *Statistical Science*, 28: 398–423. MR3135539. doi: https://doi.org/10.1214/13-STS433. 646

Consonni, G. and La Rocca, L. (2008). "Tests Based on Intrinsic Priors for the Equality of Two Correlated Proportions." *Journal of the American Statistical Association*, 103: 1260–1269. MR2462897. doi: https://doi.org/10.1198/016214508000000436. 646

Consonni, G. and La Rocca, L. (2011). "On moment priors for Bayesian model choice with applications to directed acyclic graphs." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics*, 119–144. Oxford University Press. MR3204004. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0004. 654

Consonni, G. and La Rocca, L. (2012). "Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models." *Scandinavian Journal of Statistics*, 39: 743–756. MR3000846. doi: https://doi.org/10.1111/j.1467-9469.2011.00785.x. 644

Consonni, G., La Rocca, L., and Peluso, S. (2017). "Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection." *Scandinavian Journal of Statistics*, 44: 741–764. MR3687971. doi: https://doi.org/10.1111/sjos.12273. 644

Consonni, G., Moreno, E., and Venturini, S. (2011). "Testing Hardy-Weinberg equilibrium: An objective Bayesian analysis." *Statistics in Medicine*, 30: 62–74. MR2758860. doi: https://doi.org/10.1002/sim.4084. 646

Consonni, G. and Paroli, R. (2017). "Objective Bayesian Comparison of Constrained Analysis of Variance Models." *Psychometrika*, 82: 589–609. MR3688962. doi: https://doi.org/10.1007/s11336-016-9516-y. 646

Consonni, G. and Veronese, P. (2008). "Compatibility of Prior Specifications Across Linear Models." *Statistical Science*, 23: 332–353. MR2483907. doi: https://doi.org/10.1214/08-STS258. 642

Corander, J. and Villani, M. (2004). "Bayesian assessment of dimensionality in reduced rank regression." *Statistica Neerlandica*, 58: 255–270. MR2157005. doi: https://doi.org/10.1111/j.1467-9574.2004.00108.x. 644

Corander, J. and Villani, M. (2006). "A Bayesian Approach to Modelling Graphical Vector Autoregressions." *Journal of Time Series Analysis*, 27: 141–156. MR2235152. doi: https://doi.org/10.1111/j.1467-9892.2005.00460.x. 644

Cui, W. and George, E. I. E. (2008). "Empirical Bayes vs. fully Bayes variable selection." *Journal of Statistical Planning and Inference*, 138: 888–900. MR2416869. doi: https://doi.org/10.1016/j.jspi.2007.02.011. 649

Datta, G. and Rao, J. (2010). "The Choice of Nonsubjective Priors on Hyperparameters for Hierarchical Bayes Models." In Chen, M., Mueller, P., Sun, D., Ye, K., and Dey, D. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of Jim Berger*, 237–246. Springer, New York. MR2766461. doi: https://doi.org/10.1007/978-1-4419-6944-6. 636

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Lecture Notes in Statistics. Sringer, New York. MR2053794. doi: https://doi.org/10.1007/978-1-4612-2036-7. 630, 631

Datta, J. and Ghosh, J. K. (2013). "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis*, 8: 111–132. MR3036256. doi: https://doi.org/10.1214/13-BA805. 660

Dawid, A. (1982). "Intersubjective Statistical Models." In *Exchangeability in Probability and Statistics*, in G. Koch and F. Spizzichino eds., 217–232. Amsterdam, North Holland. MR0675977. 628

Dawid, A., Stone, M., and Zidek, J. (1973). "Marginalization Paradoxes in Bayesian and Structural Inference (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 35: 189–233. MR0365805. 631

Dawid, A. P. (2006). "Invariant Prior Distributions." In *Encyclopedia of Statistical Sciences*. Wiley, New York. 630

Dawid, A. P. (2011). "Posterior Model Probabilities." In Bandyopadhyay, P. S. and Forster, M. (eds.), *Philosophy of Statistics*, 607–630. Elsevier. MR3295937. doi: https://doi.org/10.1016/B978-0-444-51862-0.50001-0. 646

Dawid, A. P. and Lauritzen, S. (2011). "Compatible prior distributions." In George, E. I. (ed.), *Bayesian Methods with Applications to Science, Policy and Official Statistics. Proceedings of the 6th World Meeting*, 109–118. International Society for Bayesian Analysis, Office for Official Publications of the European Communities. 642

Dawid, A. P. and Musio, M. (2015). "Bayesian Model Selection Based on Proper Scoring Rules." *Bayesian Analysis*, 10: 479–499. MR3420890. doi: https://doi.org/10.1214/15-BA942. 639, 646

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prunster, I., and Ruggiero, M. (2015). "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Tranactions on Pattern Analysis and Machine Intelligence*, 37,2: 803–821. 635

De-Santis, F. and Spezzaferri, F. (1997). "Alternative Bayes factors for model selection." *Canadian Journal of Statistics*, 25: 503–515. MR1614347. doi: https://doi.org/10.2307/3315344. 647

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2012). "Joint Specification of Model Space and Parameter Space Prior Distributions." *Statistical Science*, 27: 232–246. MR2963994. doi: https://doi.org/10.1214/11-STS369. 649, 656, 657

Dette, H., Ley, C., and Rubio, F. (2017). "Natural (Non-)Informative Priors for Skew-symmetric Distributions (to appear)." *Scandinavian Journal of Statistics*. 636

Fernández, C., Ley, E., and Steel, M. F. J. (2001). "Benchmark Priors For Bayesian Model Averaging." *Journal of Econometrics*, 100: 381–427. MR1820410. doi: https://doi.org/10.1016/S0304-4076(00)00076-2. 649, 654, 655

Fonseca, T., Ferreira, M., and Migon, H. (2008). "Objective Bayesian analysis for the Student-t regression model." *Biometrika*, 95: 325–333. MR2521587. doi: https://doi.org/10.1093/biomet/asn001. 634

Foster, D. P. and George, E. I. (1994). "The Risk Inflation Criterion for Multiple Regression." *Annals of Statistics*, 22: 1947–1975. MR1329177. doi: https://doi.org/10.1214/aos/1176325766. 649, 655

Fouskakis, D. and Ntzoufras, I. (2016a). "Limiting behavior of the Jeffreys Power-Expected-Posterior Bayes Factor in Gaussian Linear Models." *Brazilian Journal of Probability and Statistics*, 30: 299–320. MR3481105. doi: https://doi.org/10.1214/15-BJPS281. 654

Fouskakis, D. and Ntzoufras, I. (2016b). "Power-Conditional-Expected Priors: Using g-priors with Random Imaginary Data for Variable Selection." *Journal of Computational and Graphical Statistics*, 25: 647–664. MR3533631. doi: https://doi.org/10.1080/10618600.2015.1036996. 652

Fouskakis, D. and Ntzoufras, I. (2017). "Information Consistency of the Jeffreys Power-Expected-Posterior Prior in Gaussian Linear Models." *Metron*, 75: 371–380. MR3736668. doi: https://doi.org/10.1007/s40300-017-0110-6. 654

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). "Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care." *The Annals of Applied Statistics*, 3: 663–690. MR2750677. doi: https://doi.org/10.1214/08-AOAS207. 649

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). "Power-Expected-Posterior Priors for variable selection in Gaussian Linear Models." *Bayesian Analysis*, 10: 75–107. MR3420898. doi: https://doi.org/10.1214/14-BA887. 651, 652, 654, 655

Fouskakis, D., Ntzoufras, I., and Pericchi, L. R. (2017a). "Priors via Imaginary Training Samples of Sufficient Statistics for Objective Bayesian Model Comparison." *(submitted); Technical Report, Dept. of Mathematics, National Technical University of Athens*. 652, 653

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2017b). "Power-Expected-Posterior Priors for Generalized Linear Models." *Bayesian Analysis (to appear)*. 653

Garcia-Donato, G. and Sun, D. (2007). "Objective Priors for Hypothesis Testing in One-Way Random Effects Models." *The Canadian Journal of Statistics*, 35: 303–320. MR2393611. doi: https://doi.org/10.1002/cjs.5550350207. 646

Gelman, A. (2006). "'Prior distributions for variance parameters in hierarchical models'." *Bayesian Analysis*, 1: 515–533. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 660

Gelman, A. and Hennig, C. (2017). "Beyond subjective and objective in statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4): 967–1033. MR3723784. 629

George, E. I. (1999). "Discussion of "Bayesian Model Averaging and Model Search Strategies" by Clyde M." In *Bayesian Statistics, Vol. 6*, in J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., 175–177. Oxford University Press. MR1723497. 638

George, E. I. (2010). "Dilution priors: Compensating for model space redundancy." *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown, IMS Collections*, 6: 158–165. MR2798517. 655

George, E. I. and Foster, D. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87: 731–747. MR1813972. doi: https://doi.org/10.1093/biomet/87.4.731. 649, 653

George, E. I. and Maruyama, Y. (2014). "Posterior Odds with a Generalized Hyper-g-Prior." *Econometric Reviews*, 33: 251–269. MR3170848. doi: https://doi.org/10.1080/07474938.2013.807181. 650

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 637, 639, 659

Ghosh, M. (2011). "'Objective Priors: An Introduction for Frequentists." *Statistical Science*, 26: 187–202. MR2858380. doi: https://doi.org/10.1214/10-STS338. 629

Girón, F. J., Martínez, M. L., Moreno, E., and Torres, F. (2006). "Objective testing procedures in Linear Models: Calibration of the p-values." *Scandinavian Journal of Statistics*, 33: 765–784. MR2300915. doi: https://doi.org/10.1111/j.1467-9469.2006.00514.x. 660

Girón, F. J., Moreno, E., and Casella, G. (2007). "Objective Bayesian analysis of multiple changepoints for linear models." In *Bayesian Statistics 8, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds).*, 227–252. Oxford University Press, Oxford. MR2433195. 634, 646

Good, I. (1950). *Probability and the Weighting of Evidence*. London, UK: Charles Griffin. MR0041366. 647, 651

Good, I. (1958). "Significance Tests in Parallel and in Series." *Journal of the American Statistical Association*, 53: 799–813. MR0103560. 638

Goutis, C. and Robert, C. (1998). "Model Choice in Generalised Linear Models: A Bayesian Approach Via Kullback–Leibler Projections." *Biometrika*, 85(1): 29–37. MR1627250. doi: https://doi.org/10.1093/biomet/85.1.29. 646

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5: 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 659

Guillotte, S. and Perron, F. (2012). "Bayesian estimation of a bivariate copula using the Jeffreys prior." *Bernoulli*, 18: 496–519. MR2922459. doi: https://doi.org/10.3150/10-BEJ345. 636

Han, C. and Chaloner, K. (2004). "Bayesian Experimental Design for Nonlinear Mixed-Effects Models with Application to HIV Dynamics." *Biometrics*, 60(1): 25–33. MR2043615. doi: https://doi.org/10.1111/j.0006-341X.2004.00148.x. 661

Hansen, M. H. and Yu, B. (2001). "Model selection and the principle of minimum description length." *Journal of the American Statistical Association*, 96: 746–774. MR1939352. doi: https://doi.org/10.1198/016214501753168398. 649

Held, L., Sabanés Bové, D., and Gravestock, I. (2015). "Approximate Bayesian Model Selection with the Deviance Statistic." *Statistical Science*, 30: 242–257. MR3353106. doi: https://doi.org/10.1214/14-STS510. 640

Hoeting, J., Madigan, D., and Raftery, A. (1996). "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression." *Computational Statistics and Data Analysis*, 22: 251–270. 661

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). "On adaptive posterior concentration rates." *The Annals of Statistics*, 43: 2259–2295. MR3396985. doi: https://doi.org/10.1214/15-AOS1341. 635

Hu, J. and Johnson, V. E. (2009). "Bayesian model selection using test statistics." *Journal of the Royal Statistical Society B*, 71: 143–158. MR2655527. doi: https://doi.org/10.1111/j.1467-9868.2008.00678.x. 640

Ibrahim, J. G. and Chen, M. H. (2000). "Power Prior Distributions for Regression Models." *Statistical Science*, 15: 46–60. MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 647, 652

Ishwaran, H. and Rao, J. (2005). "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies." *The Annals of Statistics*, 33: 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 659

Iwaki, K. (1997). "Posterior Expected Marginal Likelihood for Testing Hypotheses." *Journal of Economics, Asia University*, 21: 105–134. 651

Jaynes, E. (2003). *Probability Theory*. Cambridge University Press. MR1992316. doi: https://doi.org/10.1017/CBO9780511790423. 630

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press. MR0187257. 631, 637, 638, 649, 650, 661

Johnson, V. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 646, 654

Johnson, V. E. (2005). "Bayes factors based on test statistics." *Journal of the Royal Statistical Society B*, 67: 689–701. MR2210687. doi: https://doi.org/10.1111/j.1467-9868.2005.00521.x. 640

Johnson, V. E. (2013). "Uniformly most powerful Bayesian tests." *The Annals of Statistics*, 41: 1716–1741. MR3127847. doi: https://doi.org/10.1214/13-AOS1123. 639, 640

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). "Testing hypotheses via a mixture estimation model." *arXiv:1412.2044*. 640

Kass, R. and Raftery, A. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 638

Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90: 928–934. MR1354008. 627, 629, 643, 652, 655

Kass, R. E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, 91: 1343–1369. 629, 631, 632

Kim, S. W. and Sun, D. (2000). "Intrinsic priors for model selection using an encompassing model with applications to censored failure time data." *Lifetime Data Analysis*, 6: 251–269. MR1786801. doi: https://doi.org/10.1023/A:1009641709382. 646

Lafferty, J. D. and Wasserman, L. A. (2013). "Iterative Markov Chain Monte Carlo Computation of Reference Priors and Minimax Risk." *CoRR*, abs/1301.2286. 662

Lee, J., MacEachern, S. N., Lu, Y., and Mills, G. B. (2014). "Local-mass preserving prior distributions for nonparametric Bayesian models." *Bayesian Analysis*, 9: 307–330. MR3216998. doi: https://doi.org/10.1214/13-BA857. 635

Leisen, F., Villa, C., and Walker, S. (2017). "On a Global Objective Prior from Score Rule." *(submitted); Arxiv*: https://arxiv.org/pdf/1706.00599.pdf. 628

Leon-Novelo, L., Moreno, E., and Casella, G. (2012). "Objective Bayes model selection in probit models." *Statistics in Medicine*, 31: 353–365. MR2879809. doi: https://doi.org/10.1002/sim.4406. 646

Leppä-aho, J., Pensar, J., Roos, T., and Corander, J. (2016). "Learning Gaussian Graphical Models With Fractional Marginal Pseudo-likelihood." *arXiv:1602.07863*. MR3614245. doi: https://doi.org/10.1016/j.ijar.2017.01.001. 644

Ley, E. and Steel, M. (2012). "Mixtures of g-priors for Bayesian Model Averaging with Economic Applications." *Journal of Econometrics*, 171: 251–266. MR2991863. doi: https://doi.org/10.1016/j.jeconom.2012.06.009. 650

Li, Y. (2013). "Bayesian Hierarchical Models for Model Choice." Ph.D. thesis, Department of Statistical Science, Duke University, USA. MR3187290. 650

Li, Y. and Clyde, M. A. (2016). "Mixtures of g-priors in generalized linear models." *arXiv:1503.06913v2*. MR3213874. doi: https://doi.org/10.1007/s11425-014-4815-1. 650

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 641, 642, 648, 649, 650, 653, 654, 655

Lindley, D. (1957). "A Statistical Paradox." *Biometrika*, 44: 187–192. MR0087273. doi: https://doi.org/10.1093/biomet/44.1-2.179. 657

Liseo, B. and Macaro, C. (2013). "Objective priors for causal AR(p) with partial autocorrelations." *Journal of Statistical Computation and Simulation*, 83: 1613–1628. MR3169260. doi: https://doi.org/10.1080/00949655.2012.667409. 636

Malesios, C., Demiris, N., Kalogeropoulos, K., and Ntzoufras, I. (2017). "Bayesian spatio-temporal epidemic models with applications to sheep pox." *Statistics in Medicine (to appear)*. 649, 650

Maruyama, Y. and George, E. I. (2011). "Fully Bayes factors with a generalized g-prior." *The Annals of Statistics*, 39: 2740–2765. MR2906885. doi: https://doi.org/10.1214/11-AOS917. 650, 654, 655, 658

M'lan, C. E. and Chen, M.-H. (2015). "Objective Bayesian inference for bilateral data." *Bayesian Analysis*, 10: 139–170. MR3420900. doi: https://doi.org/10.1214/14-BA890. 636

Moreno, E. (1997). "Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness." *Lecture Notes-Monograph Series*, 31. MR1833592. doi: https://doi.org/10.1214/lnms/1215454142. 645, 647

Moreno, E. (2005). "Objective Bayesian methods for one-sided testing." *TEST*, 14: 181–198. MR2203428. doi: https://doi.org/10.1007/BF02595402. 645, 646

Moreno, E., Bertolino, F., and Racugno, W. (1998). "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing." *Journal of the American Statistical Association*, 93: 1451–1460. MR1666640. doi: https://doi.org/10.2307/2670059. 645

Moreno, E. and Girón, F. J. (2008). "Comparison of Bayesian Objective Procedures for Variable Selection in Linear Regression." *TEST*, 17: 472–490. MR2470092. doi: https://doi.org/10.1007/s11749-006-0039-1. 646

Moreno, E., Girón, F. J., and Casella, G. (2010). "Consistency of objective Bayes factors as the model dimension grows." *Annals of Statistics*, 38: 1937–1952. MR2676879. doi: https://doi.org/10.1214/09-AOS754. 646

Moreno, E. and Liseo, B. (2003). "A default Bayesian test for the number of components in a mixture." *Journal of Statistical Planning and Inference*, 111: 129–142. MR1955877. doi: https://doi.org/10.1016/S0378-3758(02)00294-X. 646

Moreno, E. and Pericchi, L. R. (2014). "Intrinsic Priors for Objective Bayesian Model Selection." In *Bayesian Model Comparison*, 279–300. Emerald Group Publishing Limited. 646

Moreno, E., Torres, F., and Casella, G. (2005). "Testing equality of regression coefficients in heteroscedastic normal regression models." *Journal of Statistical Planning and Inference*, 131: 117–134. MR2137530. doi: https://doi.org/10.1016/j.jspi.2003.12.016. 646

Mukhopadhyay, M. and Minerva, T. (2017). "A mixture of g-priors for variable selection when the number of regressors grows with the sample size." *TEST*, 26: 377–404. MR3650532. doi: https://doi.org/10.1007/s11749-016-0516-0. 650

Mulder, J. and Wagenmakers, E.-J. (2016). "Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments"." *Journal of Mathematical Psychology*, 72: 1–5. MR3506020. doi: https://doi.org/10.1016/j.jmp.2016.01.002. 639

Müller, P. and Mitra, R. (2013). "Bayesian nonparametric inference—why and how." *Bayesian Analysis*, 8: 269–302. MR3066939. doi: https://doi.org/10.1214/13-BA811. 635

Neal, R. M. (2001). "Transferring Prior Information Between Models Using Imaginary Data." Technical Report 0108, Department of Statistics and Department of Computer Science University of Toronto, Canada. 651

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley Series in Computational Statistics. Hoboken, NJ: Wiley and Sons. 643

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). "Bayesian Variable and Link Determination for Generalized Linear Models." *Journal of Statistical Planning and Inference*, 111: 165–180. MR1955879. doi: https://doi.org/10.1016/S0378-3758(02)00298-7. 643, 649

Ntzoufras, I. and Tarantola, C. (2008). "Bayesian Analysis of Marginal Log-Linear Graphical Models for Three Way Contingency Tables." *arXiv:0807.1001*. 643

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society B*, 57: 99–138. MR1325379. 644

Overstall, A. M. and Forster, J. J. (2010). "Default Bayesian model determination methods for generalised linear mixed models." *Computational Statistics & Data Analysis*, 54: 3269–3288. MR2727751. doi: https://doi.org/10.1016/j.csda.2010.03.008. 643

Pérez, J. (1998). "*Development of Expected Posterior Prior Distribution for Model Comparisons*." Ph.D. thesis, Department of Statistics, Purdue University, USA. MR2699463. 642

Pérez, J. M. and Berger, J. O. (2002). "Expected-posterior Prior Distributions for Model Selection." *Biometrika*, 89: 491–511. MR1929158. doi: https://doi.org/10.1093/biomet/89.3.491. 651, 654, 655, 661

Pérez, M., Pericchi, L. R., and Ramirez, I. (2017). "The Scaled Beta2 Distribution as a Robust Prior for Scales)." *Bayesian Analysis*, 12: 615–637. MR3655869. doi: https://doi.org/10.1214/16-BA1015. 646, 660

Pericchi, L. R. (2005). "Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors." In Dey, D. and Rao, C. (eds.), *Bayesian Thinking Modeling and Computation*, volume 25 of *Handbook of Statistics*, 115–149. Elsevier. MR2490524. doi: https://doi.org/10.1016/S0169-7161(05)25004-6. 629

Piironen, J. and Vehtari, A. (2017a). "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS, PMLR 54, 905–913. 660

Piironen, J. and Vehtari, A. (2017b). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 660

Plummer, M. (2008). "Penalized loss functions for Bayesian model comparison." *Biostatistics*, 9(3): 523–539. 646

Polson, N. and Scott, J. (2012a). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7: 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 660

Polson, N. and Scott, J. G. (2012b). "Local shrinkage rules, Lévy processes and regularized regression." *Journal of the Royal Statistical Society B*, 74: 287–311. MR2899864. doi: https://doi.org/10.1111/j.1467-9868.2011.01015.x. 659, 660

Rivoirard, V. and Rousseau, J. (2012). "Posterior concentration rates for infinite dimensional exponential families." *Bayesian Analysis*, 7: 311–333. MR2934953. doi: https://doi.org/10.1214/12-BA710. 635

Robert, C. (1993). "'A Note on Jeffreys-Lindley Paradox'." *Statistica Sinica*, 3: 601–608. MR1243404. 656

Robert, C. (2007). *The Bayesian Choice*. 2nd ed. New York: Springer-Verlag. MR2723361. 630, 632

Robert, C. (2014). "Jeffreys prior with improper posterior[Blog post]." *retrieved from* https://xianblog.wordpress.com/2014/05/12/jeffreys-prior-with-improper-posterior/. 632

Ročková, V. (2018). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *Annals of Statistics*, 46(1): 401–437. MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 659

Ročková, V. and George, E. I. (2014). "EMVS: The EM Approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109: 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 659, 662

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association (accepted)*. 641, 659

Rossell, D. and Telesca, D. (2017). "Nonlocal Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 661

Rubin, H. (1987). "A weak system of axioms for "rational" behavior and the nonseparability of utility from prior." *Statistics & Decisions*, 5: 47–58. MR0886877. 654

Rubio, F. J. and Liseo, B. (2014). "On the independence Jeffreys prior for skew-symmetric models." *Statistics & Probability Letters*, 85: 91–97. MR3157886. doi: https://doi.org/10.1016/j.spl.2013.11.012. 636

Sabanés Bové, D. and Held, L. (2011). "Hyper-g Priors for Generalized Linear Models." *Bayesian Analysis*, 6: 387–410. MR2843537. doi: https://doi.org/10.1214/ba/1339616469. 643, 649, 650

Sabanés Bové, D., Held, L., and Kauermann, G. (2015). "Objective Bayesian Model Selection in Generalized Additive Models With Penalized Splines." *Journal of Computational and Graphical Statistics*, 24: 394–415. MR3357387. doi: https://doi.org/10.1080/10618600.2014.912136. 650

Sansó, B., Pericchi, L. R., and Moreno, E. (1996). "On the robustness of the intrinsic Bayes factor for nested models. (with discussion)." In *Bayesian Robustness 2*, In J. Berger, F. Ruggeri, and L. Wasserman (Eds.), 157–176. California, USA: IMS Monographs. MR1478685. doi: https://doi.org/10.1214/lnms/1215453066. 645

Savage, L. (1954). *The Foundations of Statistical Inference*. John Wiley. MR0063582. 629

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6: 461–464. MR0468014. 643

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 653, 655

Simpson, D., Rue, H., Riebler, A., Martins, T., and Sorbye, S. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 636

Som, A., Hans, C. M., and MacEachern, S. N. (2015). "Block Hyper-g Priors in Bayesian Regression." *arXiv:1406.6419v2*. MR3321977. 650

Sorbye, S. and Rue, H. (2017). "Penalised complexity priors for stationary autoregressive processes." *Journal of Time Series Analysis*, 38(6): 923–935. MR3714116. 636

Spiegelhalter, D. and Smith, A. (1980). "Bayes Factors for Linear and Log-linear Models with Vague Prior Information." *Journal of the Royal Statistical Society B*, 44: 377–387. MR0693237. 647, 651

Spiegelhalter, D. J. and Smith, A. F. M. (1982). "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44(3): 377–387. MR0693237. 647

Spitzner, D. (2005). "Risk-reducing shrinkage estimation for generalized linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67: 183–196. MR2136646. doi: https://doi.org/10.1111/j.1467-9868.2005.00495.x. 635

Spitzner, D. J. (2011). "Neutral-data comparisons for Bayesian testing." *Bayesian Analysis*, 6(4): 603–638. MR2869959. doi: https://doi.org/10.1214/11-BA623. 647, 657

Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate distribution." *Proceedings of the Third Berkeley Symposium Mathematical Statistics and Probability*, 1: 197–206. MR0084922. 658

Storey, J. (2003). "The positive false discovery rate: A Bayesian interpretation and the q-value." *The Annals of Statistics*, 31: 2013–2035. MR2036398. doi: https://doi.org/10.1214/aos/1074290335. 656

Sun, D., Tsutakawa, R. K., and He, Z. (2001). "Propriety of posteriors with improper priors in hierarchical linear mixed models." *Statistica Sinica*, 11: 77–95. MR1820002. 634

Tansey, W., Koyejo, O., Poldrack, R., and Scott, J. (2018). "False discovery rate smoothing." *Journal of the American Statistical Association (accepted)*. 643, 653

Torres-Ruiz, F., Moreno, E., and Girón, F. J. (2011). "Intrinsic priors for model comparison in multivariate normal regression." *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales – Serie A: Matematicas*, 105: 273–289. MR2826707. doi: https://doi.org/10.1007/s13398-011-0033-7. 646

Vallejos, C. A. and Steel, M. F. J. (2015). "Objective Bayesian survival analysis using shape mixtures of log-normal distributions." *Journal of the American Statistical Association*, 110: 697–710. MR3367258. doi: https://doi.org/10.1080/01621459.2014.923316. 636

van der Pas, S., Szabo, B., and var der Vaart, A. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12: 1221–1274. MR3724985. doi: https://doi.org/10.1214/17-BA1065. 643, 660

Villa, C. and Lee, J. E. (2015). "Model Prior Distribution for Variable Selection in Linear Regression Models." *arXiv:1512.08077*. 657

Villa, C. and Walker, S. (2014a). "A cautionary note on the discrete uniform prior for the binomial N: Comment." *Ecology*, 95: 2674–2677. 634

Villa, C. and Walker, S. (2014b). "Objective Prior for the Number of Degrees of Freedom of a Student t Distribution." *Bayesian Analysis*, 9: 197–220. MR3188305. doi: https://doi.org/10.1214/13-BA854. 633, 634

Villa, C. and Walker, S. (2015a). "An objective approach to prior mass functions for discrete parameter spaces." *Journal of the American Statistical Association*, 110: 1072–1082. MR3420685. doi: https://doi.org/10.1080/01621459.2014.946319. 633

Villa, C. and Walker, S. (2015b). "An Objective Bayesian Criterion to Determine Model Prior Probabilities." *Scandinavian Journal of Statistics*, 42: 947–966. MR3426304. doi: https://doi.org/10.1111/sjos.12145. 633, 634, 657

Villa, C. and Walker, S. (2017). "On The Mathematics of The Jeffreys-Lindley Paradox." *Communications in Statistics – Theory and Methods*, 46: 12290–12298. MR3740865. doi: https://doi.org/10.1080/03610926.2017.1295073. 657

Villani, M. (2001). "Fractional Bayesian Lag Length Inference in Multivariate Autoregressive Processes." *Journal of Time Series Analysis*, 22: 67–86. MR1816317. doi: https://doi.org/10.1111/1467-9892.00212. 644

Walker, S., Damien, P., and Lenk, P. (2004). "On priors with a kullback-leibler property." *Journal of the American Statistical Association*, 99(466): 404–408. MR2062826. doi: https://doi.org/10.1198/016214504000000386. 638

Wang, M. (2017). "Mixtures of g-priors for analysis of variance models with a diverging number of parameters." *Bayesian Analysis*, 12: 511–532. MR3620743. doi: https://doi.org/10.1214/16-BA1011. 650

Wasserstein, R. L. and Lazar, N. A. (2016). "The ASA's Statement on p-Values: CContext, Process, and Purpose." *The American Statistician*, 70: 129–133. MR3511040. doi: https://doi.org/10.1080/00031305.2016.1154108. 639

Wetzels, R., Grasman, R. P., and Wagenmakers, E.-J. (2012). "A Default Bayesian Hypothesis Test for ANOVA Designs." *The American Statistician*, 66: 104–111. MR2968006. doi: https://doi.org/10.1080/00031305.2012.695956. 650

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., and Schildkraut, J. M. (2010). "Bayesian model search and multilevel inference for SNP association studies." *Annals of Applied Statistics*, 4: 1342–1364. MR2758331. doi: https://doi.org/10.1214/09-AOAS322. 656

Womack, A. J., Fuentes, C., and Taylor-Rodriguez, D. (2015). "Model Space Priors for Objective Sparse Bayesian Regression." *arXiv:1511.04745*. 656

Womack, A. J., Leon-Novelo, L., and Casella, G. (2014). "Inference from Intrinsic Bayes' Procedures Under Model Selection and Uncertainty." *Journal of the American Statistical Association*, 109: 1040–1053. MR3265679. doi: https://doi.org/10.1080/01621459.2014.880348. 646, 652

Xu, C., Sun, D., and He, C. (2014). "Objective Bayesian analysis for a capture-recapture model." *Annals of the Institute of Statistical Mathematics*, 66: 245–278. MR3171405. doi: https://doi.org/10.1007/s10463-013-0413-1. 636

Ye, K. and Berger, J. O. (1991). "Noninformative Priors for Inferences in Exponential Regression Models." *Biometrika*, 78: 645–656. MR1130933. doi: https://doi.org/10.1093/biomet/78.3.645. 631

Yuan, M. and Lin, Y. (2005). "Efficient empirical Bayes variable selection and estimation in linear models." *Journal of the American Statistical Association*, 100: 1215–1225. MR2236436. doi: https://doi.org/10.1198/016214505000000367. 653

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis Using G-Prior distributions." In Goel, P. and Zellner, A. (eds.), *Bayesian Inference*

and Decision Techniques: Essays in Honor of Bruno de Finetti, 233–243. Amsterdam: North-Holland. MR0881437. 648, 649, 655

Zellner, A. (2008). "Comments on Mixtures of g-priors for Bayesian Variable Selection." *Unpublished Report, Graduate School of Business, University of Chicago*. 649

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypothesis (with discussion)." In Bernardo, J. M., DeGroot, M., Lindley, D., and Smith, A. (eds.), *Bayesian Statistics 1*, 585–606 & 618–647 (discussion). Oxford University Press. MR0862503. 650, 655

Acknowledgments

We thank an Editor and two Reviewers for highly constructive comments that led to a much better paper in terms of contents and presentation. We also thank the Editor in Chief for his unfailing support. GC was partially funded by UCSC (D1 research track). BL was partially funded by the Italian Ministry of Education, PRIN 2015, grant number 2015EASZFS – PE1.