Bayesian Analysis (2020) **15**, Number 4, pp. 1345–1423

On a Class of Objective Priors from Scoring Rules (with Discussion)

Fabrizio Leisen[*], Cristiano Villa[†], and Stephen G. Walker[‡]

Abstract. Objective prior distributions represent an important tool that allows one to have the advantages of using a Bayesian framework even when information about the parameters of a model is not available. The usual objective approaches work off the chosen statistical model and in the majority of cases the resulting prior is improper, which can pose limitations to a practical implementation, even when the complexity of the model is moderate. In this paper we propose to take a novel look at the construction of objective prior distributions, where the connection with a chosen sampling distribution model is removed. We explore the notion of defining objective prior distributions which allow one to have some degree of flexibility, in particular in exhibiting some desirable features, such as being proper, or log-concave, convex etc. The basic tool we use are proper scoring rules and the main result is a class of objective prior distributions that can be employed in scenarios where the usual model based priors fail, such as mixture models and model selection via Bayes factors. In addition, we show that the proposed class of priors is the result of minimising the information it contains, providing solid interpretation to the method.

Keywords: calculus of variation, differential entropy, Euler–Lagrange equation, Fisher information, invariance, objective Bayes, proper scoring rules.

1 Introduction

With the ever increasing popularity of Bayesian methods, attributable largely to the advent of Markov chain Monte Carlo methods and other sampling techniques, the need for default, otherwise known as objective or noninformative, priors is also in demand. Model based objective priors, such as the reference prior (Berger et al., 2009) and Jeffreys prior (Jeffreys, 1961), are commonly used when available. However, as models become larger and more complex, so it is that such priors are becoming more difficult to obtain, if not altogether unavailable. Indeed, it is our contention that model based objective priors have now reached their natural ceiling with little progress or advances in recent years.

Our observation is that limits to the progress in the research on objective priors is connected to their improperness. In fact, with very few exceptions, objective priors are improper. Although this may not represent a problem, as long as the posterior is

[*]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK, F.Leisen@kent.ac.uk

[†]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK, C.Villa-88@kent.c.uk

[‡]Department of Mathematics, University of Texas at Austin, USA, s.g.walker@math.utexas.edu

 https://doi.org/10.1214/19-BA1187

proper, it causes severe limitations to the use of objective prior distributions. Indeed, the improperness of objective priors is the main motivation which brought us to investigate a novel approach to derive objective priors. A thorough discussion of the problems that improper priors cause can be found in Kass and Wasserman (1996), where they illustrate issues such as incoherence, strong inconsistencies and nonconglomerability, dominating effect of the prior, inadmissibility, marginalisation paradoxes and improper posteriors. Although all points are undoubtedly important, it is probably the last issue that requires careful consideration. The main concern is that, as of today, general results that allow one to assess if a given improper prior yields a proper posterior are yet to be found. Research has progressed on a case by case basis; for example, see Ibrahim and Laud (1991) for the use of Jeffreys prior in generalised linear models, extended to overdispersed models of the same kind by Dey et al. (1993), Natarajan and McCulloch (1995), Berger and Strawderman (1993), and Yand and Chen (1995). More recently, Rubio and Steel (2018) describe general conditions to use improper priors for linear mixed models with longitudinal and survival data. As one would expect, the task of assessing posterior properness becomes more onerous the more complex the model is. But, even for simple models, the risk is high; as, for example, the discussion in Vallejos and Steel (2013) about the use of the Jeffreys rule prior for the Student-t regression model, derived in Fonseca et al. (2008), shows.

In this paper, we investigate constructing objective prior distributions that are not model dependent and based on the sole knowledge of the parameter space, say, Θ. As such, the connection between the prior distribution, $p(\theta)$, and the likelihood function, $f(\cdot|\theta)$, is limited to the common parameter space only.

Therefore, the prior $p(\theta)$ using only Θ loses the connection with the subjective component $f(\cdot|\theta)$ and could be argued as a consequence to be *more objective*. Conversely, model based priors, such as the Jeffreys prior (Jeffreys, 1946, 1961) or the reference prior (Berger et al., 2009), necessarily include the subjective choice of the model. In fact, models are by and large misspecified and, consequently, model based priors are propagating this misspecification. So, while a model based prior reinforces the connection between the misspecified model and the prior itself, a prior that depends on the parameter space loses only the connection.

The method we propose to derive a novel class of objective priors is based on defining a scoring rule as a combination of the *log-score* and of the *Hyvärinen* score (Parry et al., 2012). We then seek a prior such that the above score, say $S(\theta, p)$, is constant, that is

$$S(\theta, p) = \text{constant} \qquad \forall\, \theta \in \Theta. \tag{1}$$

We show in Section 3 that the density p satisfying (1) identifies a class of objective priors. Furthermore, we show in Section 4 that the density $p(\theta)$ solving (1) minimises the information in the prior as measured by a combination of the Shannon information and the Fisher information. As a result of these equivalent approaches, we show that the objective prior $p(\theta)$ is obtained by solving the following differential equation:

$$\frac{p''(\theta)}{p(\theta)} - \frac{1}{2}\left\{\frac{p'(\theta)}{p(\theta)}\right\}^2 = 1 + \log p(\theta). \tag{2}$$

Our prior is the class of solutions to the differential equation (2). The class allows us to include constraints such as properness, convexity and being decreasing, among others.

The development of objective priors is thoroughly reviewed in Consonni et al. (2018). The idea is that in a scenario where prior elicitation is not feasible, or not desirable, a prior distribution can be formed through structural or formal rules (Kass and Wasserman, 1996). The most popular objective prior is Jeffreys prior (Jeffreys, 1946, 1961), who proposed a prior distribution for continuous parameter spaces which is invariant for one-to-one transformations of the parameter space. Although in scenarios where there is only one parameter of interest, Jeffreys prior yields sensible posterior distributions. However, in cases where the parameter space has a dimension of two or more, the prior is known to yield posteriors with poor performance (sometimes giving paradoxical results, such as the marginalisation paradox, see Stone and Dawid (1972)). In such cases the priors are taken to be independent. Although other more general invariance priors have been proposed, such as in Dawid (1983), Hartigan (1964) and Jaynes (1968), the reference prior of Bernardo (Berger et al., 2009) represents an alternative to Jeffreys prior. A limitation of reference priors is sensitivity to the order of importance of parameters; this issue and possible solutions have been discussed in Berger et al. (2015). Other objective priors proposed include that of Box and Tiao (1973), based on data-translated likelihoods, and maximum entropy priors, see, for example, Jaynes (1957, 1968). As discussed in Kass (1990), these priors turn out to be very restrictive. Another important class of objective priors are the probability matching priors, first proposed in Welch and Peers (1963). The aim is to obtain a prior distribution under which the posterior probabilities of certain regions coincide with their coverage probabilities, either exactly or approximately. Recent developments of this method can be found in Sweeting et al. (2006) and Sweeting (2008). A different method, based on information theoretical concepts, has been proposed by Zellner and Min (1993), giving the so called maximal data information prior. Possibly, the most recent development in defining prior distributions, although not strictly in an objective sense, is discussed in Simpson et al. (2017). The idea is to identify the parts in a complex model that require subjective input, while the remaining parts can be associated with non-informative priors. It is important to point out that objective priors derived from scoring rules have been proposed in the recent work of Giummolè et al. (2018). A final consideration is reserved for discrete parameter spaces, whose systematic discussion can be seen to be generated by the paper of Rissanen (1983). The lack of general methods, due to the challenges that discreteness imposes, has been filled by Berger et al. (2012) first, and by Villa and Walker (2015) later.

It is known that for a mixture model Jeffreys prior can only be found under specific conditions, as studied in Grazian and Robert (2018), for reasons identified, in part, by Titterington et al. (1985). Also, standard objective priors are problematic in model comparison since Bayes factors depend on the arbitrary normalizing constants of improper prior. Though, special solutions have been proposed, for example, by O'Hagan (1995) and Berger and Pericchi (1996). Further discussion of these contexts is contained in Sections 6.1 and 6.2, where solutions provided by our approach are explored. There are other examples where the use of improper priors is challenging. For example, Kass and Wasserman (1996) discuss some issues related to their use in hierarchical modelling. Also, paradoxical results may appear, such as the marginalisation paradox (Stone and

Dawid, 1972) or the Stein's paradox (Bernardo and Smith, 1994). For more examples, see Stone (1976) and Syversveen (1998).

The paper is structured as follows. In Section 2 we give an outline of the idea and discuss a simple introductory example. In Section 3 we introduce the foundations of the proposed prior on the basis of scoring rules and their properties. An interesting aspect of the prior based on scoring rules is its interpretation in terms of the information content carried by the prior itself. This aspect is explored in Section 4. In Section 5 we present the objective priors concentrating on $\Theta = (0, 1)$, $(0, \infty)$, $(-\infty, +\infty)$ and $(-M, M)$, for a finite M, as well as on a multidimensional space. The implementation of the prior for some specific applications is presented in Section 6 where, in addition, we explore some properties of the proposed prior. Finally, Section 7 is dedicated to some concluding remarks.

2 Outline of the idea

Here we illustrate two key motivating examples, in particular in relation to the use of improper priors, and give an intuitive description of the core ideas of this work.

The key to our idea is to consider a loss function $l(\theta, p(\theta))$ which penalizes for each $\theta \in \Theta$ a choice of a prior density $p(\theta)$. The objective criterion is then based on the idea of finding the class of p which makes $l(\theta, p(\theta))$ constant. For obvious reasons, the loss function should have the following property

$$\int l(\theta, p(\theta))q(\theta)\,\mathrm{d}\theta \geq \int l(\theta, q(\theta))q(\theta)\,\mathrm{d}\theta, \tag{3}$$

for all q's representing a density for the θ. In other words, if a "true" density for θ exists, the expected loss should be minimised when such a density is chosen. The condition in (3) identifies a particular class of loss functions, known as *proper scoring rules*. One way of interpreting (proper) scoring rules is as loss functions that measure the quality of a quoted density p for an uncertain quantity θ; see, for example, Parry et al. (2012). We indicate a proper scoring rule by $S(\theta, p)$, and we ask it to be constant for all $\theta \in \Theta$. So we set

$$S(\theta, p) = \text{constant} \qquad \forall\ \theta \in \Theta,$$

and the densities satisfying the above equality identify a class of objective priors. We set the constant to 1 and show later that this choice is without loss of generality. The criterion defining this class of priors is clearly objective, for if the scoring rule were not constant, some parts of the space Θ would be given preference above others.

As discussed in Parry et al. (2012) a scoring rule is defined as *local* if it depends on the density function $p(\cdot)$ only through its value at θ, that is $p(\theta)$. Holding the above definition, we have that any *proper local* scoring rule is equivalent to the *log score*, $-\log p(\theta)$, also known as the self information loss function. However, the above scoring rule is not suitable to us, for if we set $-\log p(\theta) = \text{constant}$, we only achieve $p(\theta) \propto 1$, and the result is not interesting. Hence, we extend the score function to include the first and second derivatives of $p(\theta)$; a natural extension, in our opinion.

Thus, we consider additionally the Hyvärinen scoring rule (Hyvärinen, 2005) which makes use of the first two derivatives of p, written as p' and p''. We then have a scoring

rule $S(\theta, p)$ which has two components; the log score and the Hyvärinen score. Finding solutions to $S(\theta, p) = 1$ will now involve solving a second order differential equation and we obtain the class of prior through the two constants connected with the two derivatives. Multivariate versions of the scoring rule criterion we are proposing, and corresponding solutions, are discussed in Section 3.

Parry et al. (2012) describe a larger class of score based on the first two derivatives; see (39) in their paper. In the second order case these are based on an Euler–Lagrange equation which itself is based on a measure of information of the form $\int [p'(\theta)]^k / [p(\theta)]^{k-1}\, d\theta$ for $k = 2, 3, \ldots$. The only widely recognised information occurs with $k = 2$; the Fisher information and the corresponding score coincides with the Hyvärinen score. Hence, we use this. A connection between our scoring rule criterion and information is made through an alternative formulation through variational methods, which is discussed in Section 4.

3 Priors from scoring rules

Let us consider a quantity of interest, θ, which can take values in the space $\Theta \subset \mathbb{R}^k$. The fundamental argument behind objective prior distributions is that they should represent a state of actual or alleged prior ignorance about the true value of θ. Several criteria have been proposed to select such a prior, all of which assume that a probabilistic model generating the data (given θ) has been chosen. What we propose is to avoid this choice and derive a prior depending on Θ only. The idea is to measure the quality of the prior p with a proper scoring function, say $S(\theta, p)$, and require it to be constant, as discussed in the Introduction.

Definition 1. *A density p with respect to the Lebesgue measure on Θ, is* objective *(in accordance with commonly accepted meaning of the expression) if $S(\theta, p) =$ constant for all $\theta \in \Theta$, where S is a proper scoring rule.*

The constant here is unimportant and we can take it as 0. Any constant works – effectively it becomes 0 once the normalizing constant for p has been established.

Before proceeding we provide a brief discussion on scoring rules. Scoring rules are *proper* if $\int_\Theta S(\theta, p)\, q(\theta)\, \mathrm{d}\theta$ is minimized at $p = q$ and *local* if it depends on p only through the value $p(\theta)$. The unique proper local scoring rule is the log score, defined as

$$S_L(\theta, p) = -\log p(\theta).$$

Parry et al. (2012) extend the local property to m-local, in that now $S(\theta, p)$ depends also on the l-derivative $p^{(l)}(\theta)$, for $0 \le l \le m$. In particular, for $m = 2$, there is the Hyvärinen scoring rule, (Hyvärinen, 2005), given by

$$S_H(\theta, p) = \Delta \log p(\theta) + \frac{1}{2} |\nabla \log p(\theta)|^2.$$

In Section 4 we will illustrate the connection of the Hyvärinen scoring rule to Fisher information. Our choice of scoring rule following our reasoning in Section 2, and with

the weighting factor, is

$$\begin{aligned} S(\theta, p) &= w\, S_L(\theta, p) + S_H(\theta, p) \\ &= -w\, \log p(\theta) + \sum_{j=1}^{k} \frac{\partial^2 p/\partial\theta_j^2}{p}(\theta) - \frac{1}{2}\sum_{j=1}^{k}\left(\frac{\partial p/\partial\theta_j}{p}(\theta)\right)^2. \end{aligned}$$

That this is a proper scoring rule is derived from the fact that it is the sum of two proper scoring rules. It is also clearly 2-local. Previously, priors have been sought based solely on $\log p$; for example, the reference prior, and the math becomes unnatural as a consequence. On the other hand, including higher derivatives yields well defined solutions to optimization procedures. That we set this score to 1 for all θ is done without loss of generality, as we shall see later on. That we understand this to be an objective procedure is evident from the fact that no part of Θ is being given preference; the *loss* at θ for our choice of $p(\theta)$ is the same for all θ. For, if $S(\theta, p)$ did depend on θ then we argue that this could only be driven by *information*; i.e. parts of Θ space are preferential to others.

Predominantly, throughout the paper, we will be using the choice of $w = 1$. After all the value of w is a calibration issue between the two scores; i.e. to put them on a comparable scale. The reason for $w = 1$ is that for the benchmark standard normal density function, i.e. $p(\theta) \propto e^{-\frac{1}{2}\theta^2}$, the difference between the scores S_L and S_H is a constant (i.e. does not depend on θ), and so one does not end up dominating the other, only for $w = 1$.

Hence, we see that the objective prior, $p(\theta) \propto \exp\{-u(\theta)\}$, is obtained by solving the following differential equation:

$$\sum_{j=1}^{k} \frac{\partial^2 u}{\partial\theta_j^2} - \frac{1}{2}\sum_{j=1}^{k}\left(\frac{\partial u}{\partial\theta_j}\right)^2 = wu. \tag{4}$$

To derive the solution, we have the following result.

Theorem 3.1. *The solution to (4) is given by, for $j = 1, \ldots, k$,*

$$\frac{\partial u}{\partial\theta_j} = \pm\sqrt{c_j e^{u(\theta)} - 2w(1 + u(\theta))/k}, \tag{5}$$

for some suitable constants $c = (c_j)$ and a specified value of u at some point; e.g. $u(0, \ldots, 0)$.

Proof. Solving the differential equation (4) is equivalent to solving the following differential equation;

$$\sum_{j=1}^{k} v_j \frac{\partial v_j}{\partial u} = \frac{1}{2}\sum_{j=1}^{k} v_j^2 + uw, \tag{6}$$

having defined $v_j = \partial u/\partial\theta_j$. It is now seen the solution is given by (5). This follows since

$$v_j \frac{\partial v_j}{\partial u} = v_j\, \frac{1}{2}\frac{1}{v_j}\,(c_j e^u - 2w/k)$$

and note that $c_j e^u - 2w/k = v_j^2 + 2wu/k$. Therefore (6) holds. □

The missing pieces in (5) are $c = (c_j)$ and say $u(0)$, the constants of integration. Note that, the initial value $u(0)$ (together with the constant c) is required to ensure the existence and uniqueness of the solution. We will see how to complete these when we look at illustrations in Section 4. In general, as the solution depends on the above arbitrary constants, our method provides a class of solutions, where some are proper and some are improper and, more general, where the priors will have some assigned properties via specification of $(c, u(0))$.

We also note here that we do not need the normalizing constant for p and neither do we need to find an explicit solution for u, and p, beyond (5). The reason for this is that we can find an accurate solution via numerical methods; i.e. if we have $u(\theta)$ at a particular θ value, then we can evaluate

$$u(\theta + \varepsilon) = u(\theta) + \varepsilon' \frac{\partial u}{\partial \theta}(\theta) + \frac{1}{2}\varepsilon' \frac{\partial^2 u}{\partial \theta^2}(\theta)\varepsilon + o(|\varepsilon|^2)$$

for small ε, and the $\partial u/\partial\theta$ and $\partial^2 u/\partial\theta^2$ are available explicitly, combined with the ease of obtaining higher derivatives if needed. From here we can evaluate $p(\theta)$.

To ease the reader into the proposed prior, we illustrate the following simple example, where the parameter space is $\Theta = (-M, +M)$. Here we discuss an explicit solution to the equation $u'(\theta) = \pm\sqrt{ce^{u(\theta)} - 2(1 + u(\theta))}$, which is (5) where we have set $w = 1$ and, as the parameter space is unidimensional, $k = 1$. If we set $c = 0$ then for a solution to exist we must have $1 + u$ to be negative. Consider a prior on $\Theta = (-M, +M)$ for some finite M. With $c = 0$ we have the solution for u in the form $1 + u(\theta) = -\frac{1}{2}(\theta - \mu)^2$ for some μ. Hence, $p(\theta) \propto \exp\left\{\frac{1}{2}(\theta - \mu)^2\right\}$, which will provide a proper density on Θ.

A more general solution $p(\theta) \propto \exp\left\{\frac{1}{2}\, w(\theta - \mu)^2\right\}$ arises when we take the more general form of score function; i.e. $S(\theta, p) = w\, S_L(\theta, p) + S_H(\theta, p)$, providing interpretation for the score weighting parameter in this case. Plots of such $p(\theta)$ depending on w are presented in Figure 1, with $M = 2$.

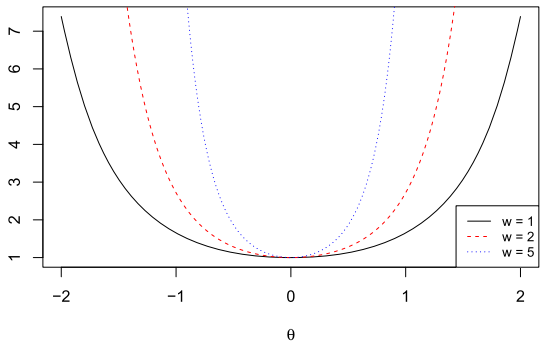


Figure 1: Prior for the parameter space $(-2, 2)$ with weights $w = 1$ (continuous black line), $w = 2$ (dashed red line) and $w = 5$ (dotted blue line).

4 Variational problems and solutions

Here we provide an alternative derivation of (4) using information theory, specifically entropy information and Fisher information. We show that the p solving (4) can also be regarded as a density carrying minimal local information. This material then is to provide support for the solution to (4) being an objective prior.

The entropy information (negative entropy) of a density function p is given by $I_E(p) = \int p(\theta) \log p(\theta)\, \mathrm{d}\theta$, which is related to Shannon's entropy and is equal to negative the expected self-information loss. In addition to $I_E(p)$, we consider a measure of the information in the density p known as Fisher information, given by

$$I_F(p) = \int \frac{p'(\theta)^2}{p(\theta)}\, \mathrm{d}\theta = \int p(\theta) \left\{ \frac{\partial}{\partial \theta} \log p(\theta) \right\}^2 \mathrm{d}\theta.$$

See, for example, Bobkov et al (2014).

Now consider $I(p) = I_E(p) + \frac{1}{2} I_F(p)$ and the aim is to find the p which minimizes $I(p)$. Recalling variational methods (Rustagi, 1976), if we wish to minimise $\int_a^b L(\theta, p, p')\, \mathrm{d}\theta$, a necessary condition for a local extremum of the integral of the Lagrangian $L(\theta, p, p')$ is that

$$\frac{\partial L}{\partial p} = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{\partial L}{\partial p'}. \tag{7}$$

Minimising $\int_a^b L(\theta, p, p')\, \mathrm{d}\theta$ reduces to the classical calculus of variation problem where we want to extremize the integral of the function

$$L(\theta, p, p') = \frac{1}{2} \frac{p'(\theta)^2}{p(\theta)} + p(\theta)\, \log p(\theta).$$

The solution to the extremal problem, if it exists, is obtained from the Euler–Lagrange equation, given by (7). According to page 44 of Rustagi (1976), if $L(p, p')$ is strictly convex on $(0, \infty) \times (-\infty, +\infty)$, and p satisfies the Euler equation, then p is a minimum of $\int_a^b L(\theta, p, p')\, \mathrm{d}\theta$. Now $L(p, p')$ is strictly convex if the matrix

$$H = \begin{pmatrix} \partial^2 L/\partial p^2 & \partial^2 L/\partial p \partial p' \\ \partial^2 L/\partial p \partial p' & \partial^2 L/\partial (p')^2 \end{pmatrix}$$

is positive definite.

Theorem 4.1. *A minimum satisfying the Euler–Lagrange equations is given by the p solving the differential equation $p' = \pm p \sqrt{c/(e\, p) + 2 \log p}$, for some suitable c.*

Proof. Calculations give

$$H = \frac{1}{p} \begin{pmatrix} \kappa^2 + 1 & -\kappa \\ -\kappa & 1 \end{pmatrix},$$

where $\kappa = p'/p$. This is easily seen to be a positive definite matrix; the eigenvalues are given by

$$\frac{1}{p}\left[\frac{1}{2}(2+\kappa^2) \pm \sqrt{\frac{1}{4}(2+\kappa^2)^2 - 1}\right],$$

which are positive.

Then (7), after some elementary algebra and differentiation, leads to the differential equation (2), which we report here for convenience,

$$\frac{p''(\theta)}{p(\theta)} - \frac{1}{2}\left\{\frac{p'(\theta)}{p(\theta)}\right\}^2 = 1 + \log p(\theta),$$

which is the same as (4). □

This differential equation has the solution derived in the previous section. It is interesting that the Euler–Lagrange equations are solved by precisely the same p solving (4).

It might be thought that we would need the constraint $\int p(\theta)\, d\theta = 1$, that is to consider $L(\theta, p, p') + \lambda(p)$, i.e. to include a Langrange multiplier. However, any ensuing differences are covered by our note in Section 5.

5 Illustrations

Before proceeding with some illustrations and applications, it is important that we discuss two aspects of the proposed class of priors within the boundaries of *Objective Bayes*: i.e. uniqueness and invariance.

It is important to discuss uniqueness and flexibility associated with objective priors. It is widely acknowledged that a prior representing total ignorance is elusive, and it might not even be possible to obtain in principle, see Bernardo and Smith (1994). As a consequence, any prior distribution, objective or not, must out of necessity provide some knowledge about something, and this "something" is not necessarily unique. For example, given a particular problem, the corresponding objective prior over a given parameter space could be proper or improper; differentiable everywhere or not; convex; log-concave; etc. In other words, a prior can be objective and exhibit desirable features of choice without impinging on subjective components relating to information.

So while we will be introducing a Bayesian objective prior criterion, it does not lead to a unique prior, rather to a class of priors, where some desirable features may or may not be included. We believe that this level of flexibility is a point of strength of the proposed approach, making it adaptable to different scenarios, including those where model based priors do not work.

Another fundamental point of discussion about prior distributions and, in particular, objective prior distributions, is invariance. Indeed, Jeffreys' rule to derive a prior distribution for the parameters of a given model is based on an invariance requirement,

in particular on invariance under one-to-one reparameterisations. Also, other common objective priors, such as reference priors, have been shown to be invariant and the same apply, for example, to the priors in Simpson et al. (2017).

Here we discuss invariance from two opposite perspectives: that it is not important, and that it is important. Before discussing this apparent contradiction, we need to point out that we define the objective prior by setting the scoring rule equal to a constant, that is $S(\theta, p(\theta)) = \text{constant}$, is invariant under location transformations.

The question is whether lack of invariance has any practical implications given that only one parameterisation will be used. Current model based objective procedures are bound to throw away some coherence properties to achieve invariance, see Kass and Wasserman (1996). However, our point is that there is no practical consequence of any relevance arising from the lack of invariance, given that, as mentioned above, a single parameterisation will be used. For example, in the case the chosen model is the normal density, one either considers the precision parameter or the variance parameter, not both. And whichever parameterisation is used, our claim is that the corresponding objective prior is adequate for the purpose to which it has been assigned.

The above points of discussion are concerned with the perspective that invariance is not important. To consider the opposite point of view let us assume that there is a canonical parameterisation for the model $f(\cdot|\theta)$. Certainly, for most models the set of parameters for which priors would be assigned is obvious. For example, the exponential family has

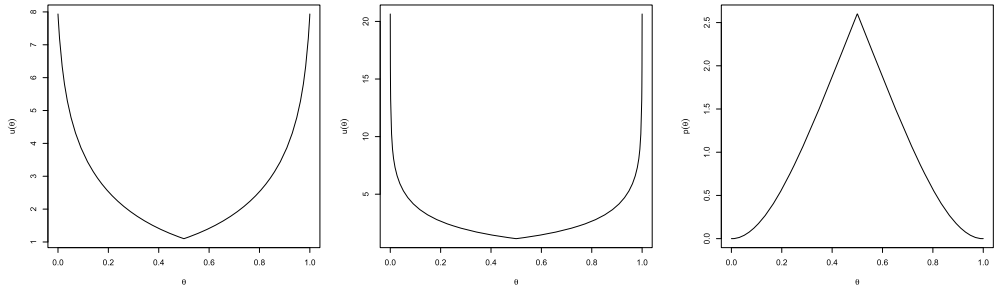
$$f(x|\theta) = h(x)\, \exp\left\{ \sum_{j=1}^{p} \theta_j\, T_j(x) - A(\theta) \right\}$$

with $\theta = (\theta_1, \ldots, \theta_p)$ being the canonical parameterisation. We can then define the canonical objective prior for statistical model $f(\cdot|\theta)$, $\theta \in \Theta$, as $p_\Theta(\theta) = \prod_{j=1}^{p} p_{\Theta_j}(\theta_j)$, where $\Theta = \otimes_{j=1}^{p} \Theta_j$. Then, any transformed prior can be obtained in the usual way involving variable transformations; that is $p(\phi) = |J|\, p_\Theta(\theta(\phi))$, where J is the Jacobian matrix for the transformation.

To illustrate the proposed method we consider three common parameter spaces in the unidimensional case and one in the multidimensional case. In particular, we consider the space for a parameter representing a probability, that is $\Theta = (0, 1)$, the space $\Theta = (0, \infty)$, usually representing the support of scale parameters, and the support for (location) parameters $\Theta = (-\infty, +\infty)$. The space $\Theta = (-M, M)$, for some finite M, has been illustrated in Section 3. As an illustration for the multidimensional case, we consider the bidimensional parameter space $(0, \infty)^2$.

5.1 One dimensional parameter space

The aim here is to solve (7) for particular motivated choices of $(c, u(0))$, equivalently, $(c, p(0))$ or $(p'(0), p(0))$. In fact, the solutions to the Euler–Lagrange equations are many, and the choice of the two constants $(c, u(0))$ will then determine a unique solution.



(a) (b) (c)

Figure 2: Plot of numerical solutions for u, panels (a) and (b) with, respectively, $u(\frac{1}{2}) = 1.1$ and $u(\frac{1}{2}) = 1.14$; plot of the normalised prior density obtained by setting $u(\frac{1}{2}) = 1.14$ (c).

Now there is the flat solution for all Θ in (4) given by $p(\theta) \propto 1$. This is achieved by setting $c = 2$ and $u(0) = 0$. However, in each of the settings of Θ considered we can find alternate priors with particular features. So, e.g. for $\Theta = (0, 1)$ we ask that $p(0) = p(1) = 0$ and for $\Theta = (0, \infty)$ we ask that p is convex and decreasing.

Case $\Theta = (0, 1)$ Here we consider the u function, recall $p \propto e^{-u}$, and so for $p(0) = p(1) = 0$ we require $u(0) = u(1) = \infty$. For additional symmetry, we can take $u\left(\frac{1}{2}\right) > 0$ and taking $c = 2$ as the extremal value, we have

$$u' = \sqrt{2}\sqrt{e^u - 1 - u} \qquad \theta > \tfrac{1}{2},$$
$$u' = -\sqrt{2}\sqrt{e^u - 1 - u} \qquad \theta < \tfrac{1}{2}.$$

Note there is a discontinuity in the derivative of u at $\theta = \frac{1}{2}$. As $u(\frac{1}{2})$ increases it is that $u(0)$ and $u(1)$ got to ∞. A plot of u is given in Figure 2a for $u(\frac{1}{2}) = 1.1$ and for $u(\frac{1}{2}) = 1.14$ in Figure 2b. For these figures we used a grid of 1000 either side of $\theta = \frac{1}{2}$ to obtain the numerical solutions. In the latter case, the corresponding density for p is presented in the right plot of Figure 2c.

It is also possible to obtain a prior that mimics Jeffreys'; that is, a distribution that has spikes at $\theta = 0$ and $\theta = 1$ with the lowest value at $\theta = \frac{1}{2}$. This is done by simply inverting u: i.e. set $u = -u$ in the above prior.

Case $\Theta = (0, \infty)$ For a prior defined on the space $(0, \infty)$ we require a specific shape property for p (convex and decreasing) and then take extremal values for the c and $u(0)$. This property is common to most objective priors on $(0, \infty)$. Thus, since $p' < 0$ we require $u' > 0$ and so $u' = \sqrt{ce^u - 2(1 + u)}$, and for u' to exist for all u we must have $c \geq 2$. Thus, as an extremal value, we take $c = 2$.

For the prior to be convex we require $p'' \geq 0$. Now $p' = -u'p$ implying $p'' = p\left\{(u')^2 - u''\right\}$. Therefore, p is convex when $(u')^2 \geq u''$. From (6) and (5) we have $(u')^2 = ce^u - 2(1 + u)$ and $u'' = \frac{1}{2}c\,e^u - 1$, and hence we are interested in the $u(0)$ for

which $c = 2 \geq 2(1+2u)e^{-u}$ for all u; i.e. $(1+2u)e^{-u} \leq 1$ for all u. Given that the function $(1+2u)e^{-u}$ is maximum, with a value of 1.31, at $u = \frac{1}{2}$, and u is increasing, since $u' > 0$, we need $u(0) > \frac{1}{2}$ and $(1+2u(0))e^{-u(0)} = 1$, again as an extremal value. Solving this gives a value for $u(0)$ of approximately 1.31.

In the next result we show that u' is bounded away from 0, and this will have important consequences for the properness of p.

Lemma 1. *It is that u' is bounded away from 0.*

Proof. To show this we need to show that $e^u - 1 - u$ is bounded away from 0 for $u \geq u(0)$. This follows trivially since $e^u - 1 - u \geq \frac{1}{2}u^2 \geq \frac{1}{2}u(0)^2$. □

The result of Lemma 1 has also the implication that p is a proper density function. To show this, we require Gronwall's inequality (Gronwall, 1919). This inequality states that, if f and g are real valued functions on $\Theta = (0, \infty)$, g is differentiable on $\text{int}(\Theta)$, and $g'(t) \leq g(t)\, f(t)$ for all $t \in \Theta$ then $g(t) \leq g(0)\, \exp\{\int_0^t f(s)\, \mathrm{d}s\}$.

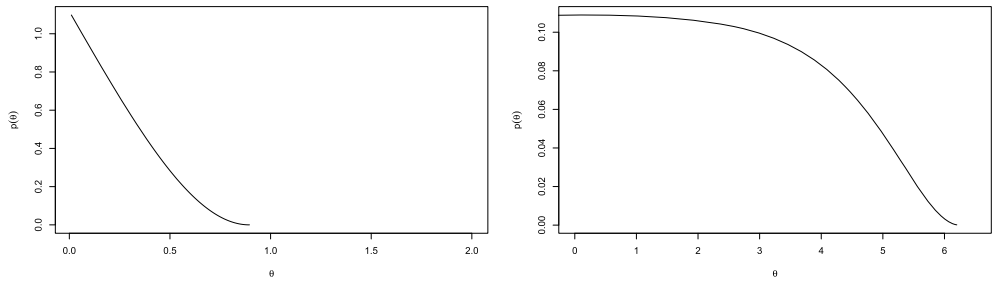
Lemma 2. *If $p(0) < \infty$, it is that $p(\theta) \leq p(0)\, e^{-\epsilon\theta}$, and hence p is proper.*

Proof. Since $p' = -u'\, p$ and we have $u' \geq \epsilon$ for some $\epsilon > 0$, it is that $p' \leq -\epsilon\, p$. From Gronwall's lemma, with $f(t) = -\epsilon$ and $g = p$, we have that

$$p(\theta) \leq p(0)\, \exp\left\{ -\int_0^\theta \epsilon\, \mathrm{d}s \right\},$$

and hence the proof is complete. □

To have a graphical image, in Figure 3a we plot the prior using the approximation available via a numerical solution to the differential equation for p. Note that this is the unnormalised p.



(a) (b)

Figure 3: Plots of the prior on $(0, \infty)$ with $c = 2$ and $u(0) = 1.31$ (a). Plot of the positive half of the prior on $(-\infty, \infty)$, obtained by symmetrising the first one and make it differentiable at the origin (b). Both densities are normalised.

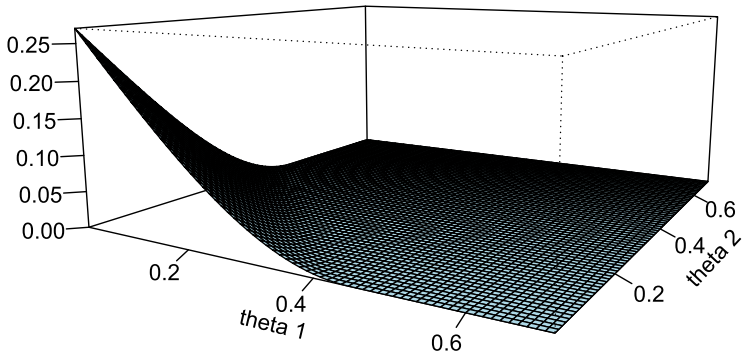


Figure 4: Surface plot of the prior for the bidimensional parameter space $\Theta = (0, \infty)^2$.

Case $\Theta = (-\infty, +\infty)$ A solution here is a symmetric version of the case $\Theta = (0, +\infty)$, which will represent a proper prior. On the other hand, if we ask that p is smooth at the origin, i.e. $p'(0) = 0$, then we need $u'(0) = 0$ and hence we must take $(c, u(0))$ to satisfy $c\, e^{u(0)} = 2 + 2u(0)$. If now we take $c = 2$, then $u(\theta)$ is a constant, resulting in a flat (improper) prior for p.

For a proper prior here one could take $u(0)$ to be small, say $u(0) = 0.01$ and then to take $c = 2\{1 + u(0)\}/\exp\{u(0)\}$. We computed numerically the right side; i.e. the $(0, \infty)$ side, for p, which is shown in Figure 3b.

To make the value of $u(0)$ more diverted, we could equally set a motivated choice for $p(0) = 1/(\sigma\sqrt{2\pi})$, corresponding to a normal density with zero mean and variance σ^2.

5.2 Multidimensional parameter space

Here we consider the bidimensional parameter space $\Theta = (0, \infty)^2$, thus $\boldsymbol{\theta} = (\theta_1, \theta_2)$. As discussed in Section 3, the prior will have the solution $p(\boldsymbol{\theta}) \propto \exp\{-u(\boldsymbol{\theta})\}$, with the partial derivatives as in (5) and additional parameters c_1 and c_2. It is sensible to expect the marginal priors $p_1(\theta_1)$ and $p_2(\theta_2)$ to be as the prior for the $\theta = (0, \infty)$ case discussed above. We can then set $u(0, 0) = (1.31, 1.31)$ and $c_1 = c_2 = 2$. The bidimensional prior is obtained using the bivariate Taylor expansion and, by applying the algorithm described in Appendix A of the Supplementary Material (Leisen et al., 2019), we obtain the prior for $(0, \infty)^2$, which surface is plotted in Figure 4.

6 Applications

The proposed class of priors is illustrated through a simulation study and the discussion of some practical implementation. Besides two initial simple examples, this section is dedicated to show the behaviour of the prior for cases where improper priors cannot be used (i.e. mixture models and model comparison via Bayes factors), while simula-

tion studies and a real data example are illustrated, respectively, in Appendix B and Appendix C in the Supplementary Material.

Although we do not have an explicit form for $p(\theta)$, we can use (5) to calculate it numerically quite easily. In particular, if we know $p(\theta)$ then we calculate $p(\theta+\delta\theta)$ for small $\delta\theta$, hence setting up the possibility of a posterior estimation process via Metropolis–Hastings sampling. The algorithm employed is detailed in Appendix A of the Supplementary Material.

To be specific, suppose we are currently at θ and the proposal value is θ'. The acceptance probability is

$$\alpha = \min\left\{1, \frac{l(\theta')}{l(\theta)}\frac{p(\theta')}{p(\theta)}\frac{q(\theta|\theta')}{q(\theta'|\theta)}\right\}, \tag{8}$$

where $l(\theta)$ is the likelihood function, and $q(\theta'|\theta)$ is the proposal density. The evaluation of $p(\theta')/p(\theta)$ in (8) does not represent any particular challenge. In fact, we have $p(\theta')/p(\theta) = \exp\{-[u(\theta') - u(\theta)]\}$, where u is the solution of the differential equation

$$u' = \sqrt{ce^u - 2(1+u)}. \tag{9}$$

Equation (9) allows us to evaluate $u(\theta') - u(\theta)$ numerically, via

$$u(\theta') \approx u(\theta) + (\theta' - \theta)u'(\theta) + \frac{(\theta'-\theta)^2}{2}u''(\theta) + \frac{(\theta'-\theta)^3}{6}u'''(\theta),$$

where the derivatives are $u'(\theta) = \sqrt{ce^{u(\theta)} - 2(1+u(\theta))}$, $u''(\theta) = \frac{1}{2}ce^{u(\theta)} - 1$, and $u'''(\theta) = \frac{1}{2}ce^{u(\theta)}\sqrt{ce^{u(\theta)} - 2(1+u(\theta))}$, and so on. Depending on how far θ' is from θ we can either use the direct approximation just given or otherwise get from θ to θ' using smaller step sizes.

The frequentist performances are illustrated in a thorough simulation study, which is presented in Appendix B of the Supplementary Material. There, we also show the complete analysis for two single i.i.d. samples, that is where we obtain data from a Poisson distribution and from a normal density with unknown mean and known variance.

6.1 Mixture models

In this section we discuss the application of the proposed method to a scenario where objective priors have been notoriously challenging, namely mixture models. Due to their flexibility, mixtures of probability distributions allow models suitable for complex data by building on simple components. As an example, consider a mixture of normal densities,

$$f(x) = \sum_{j=1}^{k} \omega_j N(x|\mu_j, \sigma_j^2), \tag{10}$$

where k is a positive integer, including ∞, and the $(\omega_j, \mu_j, \sigma_j)$ are the collection of parameters. Even under the scenario when k is known, the reference prior for model (10)

has yet to be derived, and Jeffreys prior can only be obtained under specific conditions; see Grazian and Robert (2018). Furthermore, this type of model is subject to other issues related to non-identifiability and unbounded likelihoods, among others. The issues mainly arise from the fact that improper priors may not be appropriate as we might not observe outcomes from every component of the mixture (Titterington et al., 1985). For example, Grazian and Robert (2018) show that Jeffreys prior is suitable for mixtures of normal densities only in certain circumstances; that is, when the unknown parameters are the weights. If the unknown parameters are the means or the variances, then using Jeffreys prior may lead to improper posteriors. In particular, if the unknown parameters are the means only, proper posteriors exist only when the number of mixture components is at most two; while, if the unknown parameters are the variance, or the mean and the variances, then Jeffreys prior is not suitable for inference. The above issues can be generalised to apply to any type of mixture model.

Given that the objective prior we propose is proper, it allows to make inference on the parameters of a mixture density as the yielded posteriors will be proper. As an illustration, we consider a mixture of three normal densities, where the weights and the parameters of the components are unknown. In particular, we sample from the following model

$$f(y|\omega_1,\omega_2,\omega_3,\mu_1,\mu_2,\mu_3,\sigma_1^2,\sigma_2^2,\sigma_3^2)=\sum_{i=1}^{3}\omega_i N(\mu_i,\sigma_i^2), \tag{11}$$

with weights $\omega_1 = 0.25$, $\omega_2 = 0.35$ and $\omega_3 = 0.40$, means $\mu_1 = -3.5$, $\mu_2 = 0$ and $\mu_3 = 2.5$, and variances $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.1$ and $\sigma_3^2 = 1.2$. Note that we have chosen mixture components that are reasonably distant, so not to be forced to impose any constraint to overcome identifiability, as the focus of the paper is not in this sense. However, the implementation of constraints in that sense is straightforward. For the parameters we have chosen prior independence, where each prior is the prior on the space $(0,1)$ for the weights, on the space $(-\infty,\infty)$ for the means and on the space $(0,\infty)$ for the variances, in agreement with Section 5. The prior on $(-\infty,\infty)$ is the symmetrised version from $\Theta=(0,\infty)$. To ensure properness of the priors for the means and variances, we have set $c = 2$ and $u(0) = 1.31$ (as discussed in Section 5).

We have performed the analysis on two data sets of size $n = 100$ and $n = 250$. The whole details, including histograms of the sample data, description of the algorithm implemented, as well as convergence diagnostics are reported in Appendix D of the Supplementary Material.

6.2 Model comparison

Another simple case where objective priors are problematic is in model comparison (or selection) via Bayes factors. So, if we wish to compare model $M_1 = \{f_1(x|\boldsymbol{\theta}_1), p_1(\boldsymbol{\theta}_1)\}$ to model $M_2 = \{f_2(x|\boldsymbol{\theta}_2), p_2(\boldsymbol{\theta}_2)\}$, where both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vector of parameters with some elements not in common, then the Bayes factor

$$BF_{12}=\frac{\int f_1(x|\boldsymbol{\theta}_1)p_1(\boldsymbol{\theta}_1)\,d\boldsymbol{\theta}_1}{\int f_2(x|\boldsymbol{\theta}_2)p_2(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2},$$

is, in general, meaningful if the priors assigned to non-common parameters are proper. If not, then the arbitrary multiplicative constant up to which they are defined do not cancel and the Bayes factor depends on an arbitrary constant. Solutions to the issue have been proposed, see, for example, O'Hagan (1995) and Berger and Pericchi (1996), however, the resulting procedures are still quite tedious to implement and are limited to simple models. By and large the above issue stays; however, Berger et al. (1998) give an exception of the issue. Furthermore, in Dawid and Musio (2015) and Dawid et al. (2017) the authors propose to make use of homogeneous scoring rules that circumvent the problem of using improper priors on the parameters.

An example on the application of the proposed prior in model selection is discussed in Appendix E of the Supplementary Material. There, a Poisson distribution and a geometric distribution are compared, where the priors are, respectively, of a parameter space $(0, \infty)$ and $(0, 1)$. As properness is necessary in this context, we have set $c = 2$ and $u(0) = 1.31$.

Nested models

When models under comparison are nested, there are particular considerations which are needed to be taken into account; see, for example, Consonni et al. (2013). The point is that a diffuse type prior for the larger model will end up lacking focus so that the mass assigned to the smaller model is too much. However, our argument is that if two nested models are under comparison, it is essential, at least from a coherent point of view, to center the larger prior on the fixed part of the smaller one. Let us elaborate.

Suppose $f(y|\theta)$ for $\theta \in \Theta_1$ is the larger model and the smaller one is given by $\theta \in \Theta_0$ where $\Theta_0 \subset \Theta_1$. Typically Θ_1 will be a higher dimension to Θ_0 and to get the latter from the former one fixes a particular value in the higher dimension. To make this concrete, let us consider Example 2.1 from Consonni et al. (2013), where $M_0 : f(y|\theta_0)$ is binomial(n, θ_0), with $\theta_0 = 1/4$ fixed, and $M_1 : f(y|\theta)$ is binomial(n, θ), for which a prior for θ, $p(\theta)$, is required. Given the nature of the comparison it is our argument that $p(\theta)$ must be centered on θ_0.

We can adapt quite easily the prior obtained in Section 4, the $\Theta = (0, 1)$, to be centered on $1/4$ rather than $\frac{1}{2}$. Without repeating the mathematics, we can take $u(1/4) = w$ and $c = 2$ and then

$$u' = \sqrt{2}\sqrt{e^u - 1 - u} \qquad \theta > 1/4,$$
$$u' = -\sqrt{2}\sqrt{e^u - 1 - u} \qquad \theta < 1/4.$$

For the illustration of the prior $p(\theta)$, obtained numerically from u, in Figure 5 we took $w = 1.5$.

The proposed prior, centered at $\theta_0 = 0.25$, is compared with the intrinsic prior in Consonni et al. (2013), that is

$$p^I(\theta|b, t) = \sum_{x=0}^{t} \text{Beta}(\theta|b + x, b + t - x)\text{Bin}(x|t, \theta_0),$$

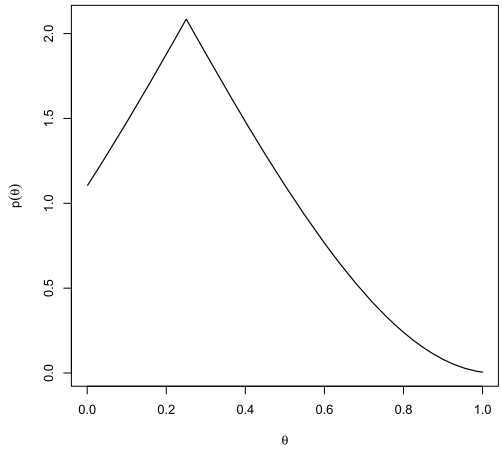


Figure 5: Numerical solution for normalized p with $w = 1.5$.

where $b = 1$ and $t = 8$. The intrinsic prior defined above is centered at $w\theta_0 + (1-w)\frac{1}{2}$, where $w = t/(2b+t)$, and has behaviour similar to the one in Figure 5, giving the Intrinsic Bayes Factor in favour of M_1

$$BF_{10}^I = \sum_{x=0}^{8} \frac{B(1+x+y, 1+t-x+n-y)}{B(1+x, 1+t-x)\theta_0^y(1-\theta_0)^{n-y}},$$

where $n = 12$. A relatively small sample size allows to better capture differences in the performance of the two priors. Figure 6 shows the posterior probability for M_1, i.e. $P(M_1|y) = (1 + 1/BF_{10}^I)^{-1}$. The priors yield model probabilities that are similar; in fact in both cases the lowest point is at $\theta = \theta_0$ and, the more θ moves away from θ_0 the higher the posterior probability for M_1.

6.3 Further properties of the prior

It is important to illustrate how the choices of the constraints $u(0)$ and c impact the prior; in particular, in terms of mean, variance and tail behaviour.

As a general result, if we need to center the prior at a particular θ_0, we can simply set

$$u' = \sqrt{ce^u - 2(1+u)} \qquad \theta > \theta_0,$$
$$u' = -\sqrt{ce^u - 2(1+u)} \qquad \theta < \theta_0.$$

See Section 6.2.1. In some cases, for example, when nested models are compared, one wishes to have a prior distribution with tails that are as heaviest as it is possible; i.e. suitable for alternative priors. This is achieved by aiming to have a prior for which u' is as small as possible, thus we can set c to the smallest value for which we can solve the equation, which is $c = 2$, if $u(0) < 0$, and $c = 2(1 + u(0))e^{-u(0)}$, if $u(0) > 0$.

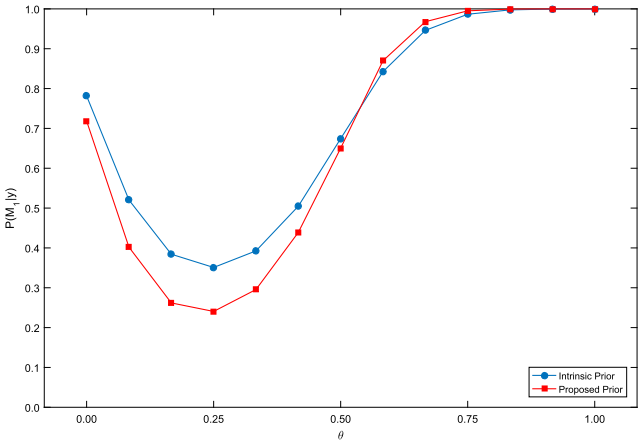


Figure 6: Small sample evidence for the Binomial example. The graph shows the posterior probability for model $f(y|\theta) = \text{Bin}(n = 12, \theta)$ using the proposed prior (squares) and the intrinsic prior (circles).

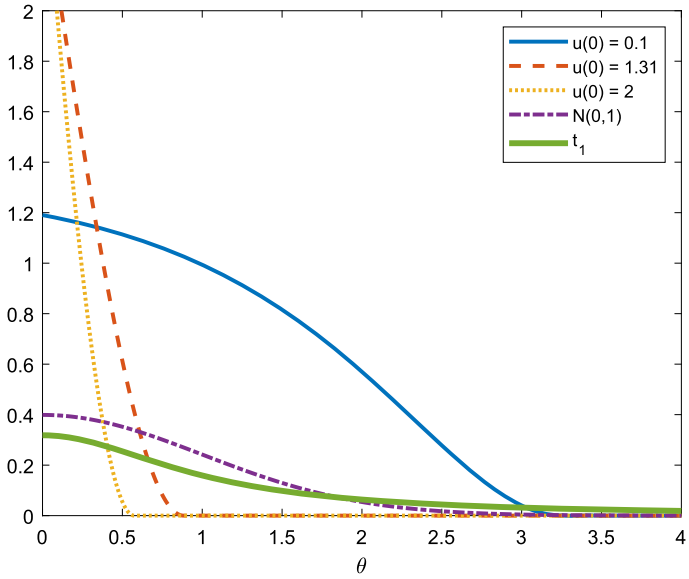


Figure 7: Comparison of the prior tail, with $c = 2$ and different values of $u(0)$, to the standard normal and the Student-t with 1 degree of freedom.

As an illustration, let us consider the case $\Theta = (-\infty, \infty)$, and compare the tail of the prior distribution to the tails of a normal density and a Student-t with 1 degree of freedom. For simplicity, we consider only the positive half of the real line. Figure 7 shows

the comparison of the prior based on scoring rules for $c = 2$ and for $u(0) = (0.1, 1.31, 2)$. When compare to both the standard normal and the Student-t with 1 degree of freedom, the proposed prior appears to have lighter tails for relatively large values of $u(0)$. In particular, the higher the value the quicker the prior drops towards 0. On the other hand, should we select a small value of the initial condition $u(0)$, then the proposed prior has a more gentle decrease to zero.

To understand the effect of the initial conditions on mean and variance of the prior, let us start by considering the case $\Theta = (0, \infty)$. We explore options where both $u(0)$ and c have increasing values; that is,

$$u(0) = (0.01, 0.05, 0.1, 0.5, 1, 1.31, 1.5, 2, 10),\ c = (2, 2.01, 2.05, 2.1, 2.5, 3, 5, 10, 20, 50).$$

The prior mean and variance are reported in, respectively, Table 1 and Table 2. We see that when the conditions increase, mean and variance of the prior tend to 0. This is in

	$u(0)$								
c	0.01	0.05	0.1	0.5	1	1.31	1.5	2	10
2	3.54	1.82	1.07	0.23	0.09	0.05	0.04	0.02	5.59E-06
2.01	1.63	1.23	0.90	0.23	0.09	0.05	0.04	0.02	5.57E-06
2.05	0.89	0.77	0.64	0.21	0.08	0.05	0.04	0.02	5.51E-06
2.1	0.66	0.58	0.51	0.19	0.08	0.05	0.04	0.02	5.43E-06
2.5	0.27	0.26	0.23	0.12	0.06	0.04	0.03	0.02	4.46E-06
3	0.17	0.16	0.15	0.09	0.04	0.03	0.02	0.01	3.60E-06
5	0.07	0.07	0.06	0.04	0.02	0.02	0.01	0.01	2.34E-06
10	0.03	0.03	0.03	0.02	0.01	0.01	0.01	0.00	1.11E-06
20	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	5.59E-07
50	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00E-07

Table 1: Mean of the prior distribution over $\Theta = (0, \infty)$ for different combinations of $u(0)$ and c.

	$u(0)$								
c	0.01	0.05	0.1	0.5	1	1.31	1.5	2	10
2	0.78	0.99	0.54	0.10	0.03	0.02	0.01	0.00	2.15E-08
2.01	0.88	0.63	0.45	0.10	0.03	0.02	0.01	0.00	2.14E-08
2.05	0.45	0.39	0.32	0.09	0.03	0.01	0.01	0.00	2.10E-08
2.1	0.33	0.29	0.25	0.08	0.03	0.01	0.01	0.00	2.05E-08
2.5	0.12	0.11	0.10	0.04	0.02	0.01	0.01	0.00	1.54E-08
3	0.07	0.06	0.06	0.03	0.01	0.01	0.00	0.00	1.14E-08
5	0.02	0.02	0.02	0.01	0.00	0.00	0.00	0.00	5.83E-09
10	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	2.03E-09
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.69E-10
50	0.00	0.00	0.00	0.00	0.00	7.77E-05	5.87E-05	2.78E-05	2.07E-10

Table 2: Variance of the prior distribution over $\Theta = (0, \infty)$ for different combinations of $u(0)$ and c.

	$u(0)$								
c	0.01	0.05	0.1	0.5	1	1.31	1.5	2	10
2	10.50	3.44	1.59	0.17	0.04	0.02	0.01	0.01	2.16E-08
2.01	2.93	1.94	1.24	0.16	0.04	0.02	0.01	0.00	2.15E-08
2.05	1.21	0.97	0.75	0.14	0.04	0.02	0.01	0.00	2.10E-08
2.1	0.77	0.65	0.53	0.13	0.03	0.02	0.01	0.00	2.05E-08
2.5	0.21	0.19	0.17	0.06	0.02	0.01	0.01	0.00	1.55E-08
3	0.11	0.10	0.09	0.04	0.01	0.01	0.01	0.00	1.14E-08
5	0.03	0.03	0.02	0.01	0.01	0.00	0.00	0.00	5.84E-09
10	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	2.03E-09
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.70E-10
50	0.00	0.00	0.00	0.00	0.00	8.07E-05	6.08E-05	2.85E-05	2.07E-10

Table 3: Variance of the prior distribution over $\Theta = (-\infty, \infty)$ for different combinations of $u(0)$ and c.

line with the tail behaviour discussed above, where the larger the value chosen for $u(0)$, the faster the prior drops to 0.

For the case $\Theta = (-\infty, \infty)$ we have chosen to center the prior ate zero; although other values are easily attainable yielding similar results. Again, the prior variance (Table 3) decreases as we set the initial conditions to increasingly large values.

A final application of the proposed prior is for a Poisson regression model. The example is presented in Appendix C of the Supplementary Material, where we have worked with simulated data as well as real data. One objective of the simulation study is to show the robustness of the prior when nuisance covariates are added in the regression model; in fact, it can be seen that there is no noticeable impact on the size of the posterior credible intervals.

7 Discussion

In this paper we have introduced a new class of objective priors derived from scoring rules. A remarkable aspect is that we have been able to show that the same result can be achieved via the rigour of calculus of variations, by finding objective priors which solve the Euler–Lagrange equation for finding extremum to integrals of the type

$$\int L(\theta, p, p')\, \mathrm{d}\theta.$$

If we can establish suitable choices of $L(\theta, p, p')$ which can be motivated and satisfy conditions for the existence of extremum, then new classes of objective prior can be sought. The case we have considered, which we can consider as a first step, is to use a combination of two well known measures of information in a prior density function; i.e.

$$L(\theta, p, p') = \frac{1}{2}\frac{p'(\theta)^2}{p(\theta)} + p(\theta)\, \log p(\theta).$$

The objective priors here defined have two desirable properties. The first is that they are somewhat detached from the choice of the sampling distribution and are dependent on the parameter space only. In other words, the information required to derive the prior is limited to the range of values that the quantity of interest can take.

The second property is that the prior can be proper. Besides the advantage of not having to check properness of the posterior, it allows to exploit the prior in scenarios where improper objective priors have been challenging. For example, as illustrated in Section 6.1, the proposed prior is used to estimate the means of a mixture of normal densities with three components. Another potential application, discussed in Section 6.2, is in model selection. In particular, the objective prior may be used to represent minimal information on the parameters that are not common to two models. In fact, the Bayes factor used to compare two models is, in general, sensitive to the proportionality constant of improper priors. While for common parameters the constant will cancel out, this is not the case if the parameter is either at the numerator or at the denominator of the ratio only. Hence, the necessity of having a proper prior assigned to this kind of parameters.

The simulation study, aimed to compare the frequentist performances of the proposed prior with the ones of the Jeffreys prior, has shown no appreciable differences, with the exception of a slightly larger Mean Squared Error (MSE) for the proposed prior; the last result is expected as it is a consequence of the smaller information used to define the proposed prior in comparison with any model based objective prior.

Future and ongoing work involves using the Fisher information alone; i.e. to minimize $\int (p')^2/p \mathrm{d}p$ subject to p having certain constraints; for example, a zero mean or a specified variance or being log-concave. The mathematical results here would be able to provide explicit solutions including a class of non-local prior distributions (Johnson and Rossell, 2010).

Supplementary Material

On a Class of Objective Priors from Scoring Rules. Supplementary Material (DOI: 10.1214/19-BA1187SUPP; .pdf). Supplement to "On a Class of Objective Priors from Scoring Rules". The supplementary material contains the Appendixes A, B, C and D of the paper.

References

Berger, J. O. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, **1**, 1–17. MR2221271. doi: https://doi.org/10.1214/06-BA115.

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. 1345, 1346, 1347

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. 1347

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. 1347

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. 1347, 1360

Berger, J. O., Pericchi, L. R., and Varshavsky, J. (1998). "Bayes factors and marginal distributions in invariant situations." *Sankhya*, **60**, 109–122. MR1718789. 1360

Berger, J. O. and Strawderman, W. (1993). "Choice of hierarchical priors: admissibility in estimation of normal means." *Technical report*, 93-34C, Purdue University, Dept. of Statistics. MR1401831. doi: https://doi.org/10.1214/aos/1032526950. 1346

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470316870.ch1. 1347, 1353

Bobkov, S. G., Gozlan, N., Roberto, C. and Samson, P. M. (2014). "Bounds on the deficit in the logarithmic Sobolev inequality." *Journal of Functional Analysis*, **267**, 4110–4138. MR3269872. doi: https://doi.org/10.1016/j.jfa.2014.09.016. 1352

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley. MR0418321. 1347

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. 1360

Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018). "Prior distributions for objective Bayesian analysis." *Bayesian Analysis*, **13**, 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 1347

Dawid, A. P. (1983). "Invariant prior distributions." In *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, New York: John Wiley. 1347

Dawid, A. P. and Musio, M. (2015). "Bayesian Model Selection based on Proper Scoring Rules (with discussion)." *Bayesian Analysis*, **10**, 479–521. MR3420890. doi: https://doi.org/10.1214/15-BA942. 1360

Dawid, A. P., Musio, M. and Columbu, S. (2017). "A Note on Bayesian Model Selection for Discrete Data Using Proper Scoring Rules." *Statistics and Probability Letters*, **129**, 101–106. MR3688521. doi: https://doi.org/10.1016/j.spl.2017.05.010. 1360

Dey, D. K., Gelfand, A. E. and Peng, F. (1993). "Overdispersed generalized linear models." *Technical report*, University of Connecticut, Dept. of Statistics. 1346

Fonseca, T. C. O. , Ferreira, M. A. R. and Migon, H. S. (2008). "Objective Bayesian analysis for the Student-T regression model." *Biometrika*, **95**, 325–333. MR2521587. doi: https://doi.org/10.1093/biomet/asn001. 1346

Giummolè, F., Mameli, V., Ruli, E. and Ventura, L. (2019). "Objective Bayesian inference with proper scoring rules." *TEST*, **28**, 728–755. MR3992136. doi: https://doi.org/10.1007/s11749-018-0597-z. 1347

Grazian, C. and Robert, C. P. (2018). "Jeffreys priors for mixture estimation: properties and alternatives." *Computational Statistics and Data Analysis*, **121**, 149–163. MR3759204. doi: https://doi.org/10.1016/j.csda.2017.12.005. 1347, 1359

Gronwall, T. H. (1919). "Note on the derivatives with respect to a parameter of the solutions of a system of differential equations." *Annals of Mathematics*, **20**, 292–296. MR1502565. doi: https://doi.org/10.2307/1967124. 1356

Hartigan, J. A. (1964). "Invariant prior distributions." *Annals of Mathematical Statistics*, **35**, 836–845. MR0161406. doi: https://doi.org/10.1214/aoms/1177703583. 1347

Hyvärinen, A. (2005). "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, **6**, 695–709. MR2249836. 1348, 1349

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian analysis of generalized linear models using Jeffreys's prior." *Journal of the American Statistical Association*, **86**, 981–986. MR1146346. 1346

Kass, R. E. (1990). "Data-translate likelihood and Jeffreys's rule." *Biometrika*, **77**, 107–114. MR1049412. doi: https://doi.org/10.2307/2336053. 1347

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, **91**, 1343–1370. 1346, 1347, 1354

Leisen, F., Villa, C. and Walker, S. G. (2019). "On a Class of Objective Priors from Scoring Rules. Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1187SUPP. 1357

Jaynes, E. T. (1957). "Information theory and statistical mechanics I, II." *Physical Review*, **106**, 620–630; **108**, 171–190. MR0087305. 1347

Jaynes, E. T. (1968). "Prior probabilities." *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, 227–241. 1347

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Society of London*, Ser. A, **186**, 453–461. MR0017504. doi: https://doi.org/10.1098/rspa.1946.0056. 1346, 1347

Jeffreys, H. (1961). *Theory of Probability and Inference*, 3rd ed., Cambridge University Press, London. MR0745621. 1345, 1346, 1347

Johson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society, Series B*, **72**, 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 1365

Natarajan, R. and McCulloch, C. E. (1995). "A note on the existence of the posterior

distribution for a class of mixed models for binomial responses." *Biometrika*, **82**, 639–643. MR1366287. doi: https://doi.org/10.1093/biomet/82.3.639. 1346

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society, Series B*, **57**, 99–138. MR1325379. 1347, 1360

Parry, M., Dawid, A. P. and Lauritzen S. (2012). "Proper local scoring rules." *Annals of Statistics*, **40**, 561–592. MR3014317. doi: https://doi.org/10.1214/12-AOS971. 1346, 1348, 1349

Rissanen, J. (1983). "A universal prior for integers and estimation by minimum description length." *Annals of Statistics*, **11**, 416–431. MR0696056. doi: https://doi.org/10.1214/aos/1176346150. 1347

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.

Rubio, F. J. and Steel, M. F. J. (2018). "Flexible linear mixed models with improper priors for longitudinal and survival data." *Electronic Journal of Statistics*, **12**, 572–598. MR3769189. doi: https://doi.org/10.1214/18-EJS1401. 1346

Rustagi, J. S. (1976). *Variational Methods in Statistics*. Academic Press. MR0402569. 1352

Stone, M. (1972). "Strong Inconsistency from Uniform Priors." *Journal of the American Statistical Association*, **71**, 114–116. MR0415866. 1347

Stone, M. and Dawid, A. (1972). "Un-Bayesian implications of improper Bayes inference in routine statistical problems." *Biometrika*, **59**, 369–375. MR0431449. doi: https://doi.org/10.1093/biomet/59.2.369. 1347

Syversveen, A. R. (1998). "Noninformative Bayesian priors. Interpretation and problems with construction and applications." *Technical Report*. 1347

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017). "Penalising model component complexity: a principled, practical approach to constructing priors." *Statistical Science*, **32**, 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 1347, 1354

Sweeting, T. J. (2008). "On predictive probability matching priors." *IMS Collections: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, eds. B. Clarke and S. Ghosal. **3**, 46–59. MR2459215. doi: https://doi.org/10.1214/074921708000000048. 1347

Sweeting, T. J., Datta, G. S. and Ghosh, M. (2006). "Nonsubjective priors via predictive relative entropy loss." *Annals of Statistics*, **34**, 441–468. MR2275249. doi: https://doi.org/10.1214/009053605000000804. 1347

Titterington, D., Smith, A. and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York. MR0838090. 1347, 1359

Vallejos, C. A. and Steel, M. J. F. (2013). *On posterior propriety for the Student-t linear*

regression model under Jeffreys priors. arxiv:1311.1454. MR2949504. doi: https://doi.org/10.1080/00949655.2011.563239. 1346

Villa, C. and Walker, S. G. (2015). "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. 1347

Welch, B. L. and Peers, H. W. (1963). "On formulae for confidence points based on integrals of weighted likelihoods." *Journal of the Royal Statistical Society, Series B*, **35**, 318–329. MR0173309. 1347

Yang, R. and Chen, M. H. (1995). "Bayesian analysis for random coefficient regression models using noninformative priors." *Journal of Multivariate Analysis*, **55**, 283–311 MR1370405. doi: https://doi.org/10.1006/jmva.1995.1080. 1346

Zellner, A. and Min, C. (1993). "Bayesian analysis model selection and prediction." In *Physics and Probability: Essays in honour of Edwin T. Jaynes*, eds. W. T. Grandy, Jr. and P. W. Milonni, Cambridge, U.K.: Cambridge University Press. MR1992316. doi: https://doi.org/10.1017/CBO9780511790423. 1347

Acknowledgments

We thank the Editor, the Associate Editor, and two Reviewers for insightful comments on a previous version of the paper. We also extend our thanks to Guido Consonni for feedback on an earlier draft of the paper. Fabrizio Leisen was supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no: 630677. The third author is partially supported by NSF grant DMS 1612089.

Invited Discussion

Dimitris Fouskakis*

This is an interesting paper introducing a new class of objective priors, derived from scoring rules. The authors do a nice job and show that the same result can be achieved via the rigour of calculus of variations. The new defined objective priors are: (a) detached from the choice of the sampling distribution and depend only on the parameter space; (b) can be proper. Leisen, Villa & Walker measure the quality of the prior with a proper scoring function and require it to be constant; by this way they are no parts of the parameter space Θ that are "preferred", and thus the prior is objective. Focus is given in the uniqueness and invariance of the defined priors and illustrations include the application of the proposed method in mixture models and model comparison via Bayes factor problems.

My comments below are focused on the model selection problem. The main reason for this choice is that the objective Bayes 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; see for example Consonni et al. (2018). With my discussion below I wish to add further insights to the paper resulting in future work in the area.

Bayes factor derivation In Section 6.2 the authors consider a simple (nested) model comparison problem. The proposed prior is then compared with the intrinsic prior in Consonni et al. (2013). I would love to see some discussion on the derivation of the final Bayes factor, since the marginal likelihood is not analytical available. Is it easy, in general, to use the proposed methodology in more complicated model selection problems?

Variable selection problem One of the most common applications of model comparison is the variable selection problem for normal linear regression, as well as, generalized linear models. A variety of approaches have been proposed, such as the g-prior and its extensions (Liang et al., 2008), the robust prior (Bayarri et al., 2012), the PEP prior (Fouskakis et al., 2015, 2018) the benchmark prior (Ley and Steel, 2012) and many others (Consonni et al., 2018). Most of the priors used for variable selection problems belong to the class of "Confluent Hypergeometric Information Criterion" g-priors introduced by Li and Clyde (2018).

Furthermore, when the number of models under comparison is small, the problem can be replaced with an estimation one that focuses on the probability weight of a given model within a mixture model (Kamary et al., 2018). Using this approach, the methodology described in Section 6.1 can be applied here as well.

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

To conclude with, I believe that it would be great to see an application of the proposed methodology in such problems and compare the performance of the proposed method with that of the "most established" Bayesian variable selection techniques.

Desiderata 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. Predictive matching in particular is viewed as the most crucial aspect for objective model selection priors. Informally it says that, with a minimal sample size, one should not be able to discriminate between two models, so that the Bayes factor should be close to one, for all samples of minimal size.

I would love to see some work on this direction; i.e. check if these criteria are satisfied for the proposed prior. In Bayarri et al. (2012), first criteria meaningful for priors tailored to objective model selection were set out, and then priors satisfying them were derived. The authors could proceed in a similar manner here as well, and select constants resulting to priors that satisfying these criteria.

Selection of the constants of integration Returning back to the previous point, how the selection of the constants of integration affects the shape and the theoretical properties of the prior? Selection of different constants, results to local or non-local priors, proper or improper, with light or heavy tails. In Section 6.3 the authors illustrate how the choices of the constants impact the prior, but mostly their work is based on illustrations rather than theoretical properties. In addition, the authors are doing a great job in Section 5, by considering three common parameter spaces in unidimensional case and one in the multidimensional case, and defining a solution; but it still seems to me that the selection of the constants is a bit ad hoc. It would be nice to get a fully automatic approach, depending on the problem that the researcher faces, based on theoretical properties.

Type of prior The group invariance criterion in Bayarri et al. (2012) 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, resulting to the 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. As mentioned above, on a previous point, it seems to me that by changing the values of the constants, the resulting priors could belong to either of these broad categories. If this is the case, would be nice to compare the non-local priors resulting by this approach with the ones that Johnson and Rossell (2010) use, under specific cases.

Model misspecification Does model misspecification affect the selection of the true model, under the proposed prior, or since the prior is detached from the choice of the sampling distribution this problem does not arise? The problem has been explored in various papers recently; see for example Rossell and Rubio (2018), and I suggest exploring this avenue.

Diffuse priors Focusing again on nested model selection problems, it is well known that if proper prior distributions with large variances are used, the resulting Bayes factors can be highly sensitive to the chosen prior variances (Bartlett, 1957). Is this the case when using particular values of the constants?

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. Also in regression settings common default priors are not defined when the number of predictors p is larger than the sample size n. More generally, high-dimensional problems pose new challenges that need be addressed through novel methodologies; such as sparsity and shrinkage. These two "ideas", though distinct, are closely related as we seek 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.

How possible is to apply the proposed methodology in high-dimensional models? How the selection of the constants of integration affects the degree of shrinkage? Would be great if the authors, in a future work, explore this avenue.

Conclusion I would like to thank the authors for a thought-provoking paper on an important issue. The paper is a useful addition to the literature of objective priors. I thank also the Editor for inviting me to contribute to the discussion. All my comments above are basically associated with model selection problems, where the derivation of objective priors is a more challenging task. As a final note, will be also great if the authors create a GitHub or an RPubs in order to share their code, to improve the visibility of the paper, and to allow reproducibility of the results. Even better, as part of a future work, an R package with an ever-increasing set of case studies and automatic tools of choosing the constants for practitioners with limited Statistics training would be great.

References

Bartlett, M. (1957). "Comment on D.V. Lindleys statistical paradox." *Biometrika*, 44: 533–534. MR0086727. doi: https://doi.org/10.1093/biomet/44.1-2.27. 1372

Bayarri, M. J., Berger, J., Forte, A., and Garcia-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1370, 1371

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." *Statist. Sci.*, 28: 398–423. MR3135539. doi: https://doi.org/10.1214/13-STS433. 1370

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). "Prior distributions for objective Bayesian analysis." *Bayesian Analysis*, 13: 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 1370

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. 1370

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). "Power-expected-posterior priors for generalized linear models." *Bayesian Analysis*, 13: 721–748. MR3807864. doi: https://doi.org/10.1214/17-BA1066. 1370

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 1371

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2018). "Testing hypotheses via a mixture estimation model." arXiv:1412.2044. 1370

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. 1370

Li, Y. and Clyde, M. (2018). "Mixtures of g-priors in generalized linear models." *Journal of the American Statistical Association*, 113: 1828–1845. MR3902249. doi: https://doi.org/10.1080/01621459.2018.1469992. 1370

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (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. 1370

Rossell, D. and Rubio, F. J. (2018). "Tractable Bayesian variable selection: beyond normality." *Journal of the American Statistical Association*, 113: 1742–1758. MR3902243. doi: https://doi.org/10.1080/01621459.2017.1371025. 1371

Invited Discussion

Matthew Parry*

It was a pleasure to read this paper and to be asked to provide a discussion piece to go with it. When we wrote our paper on local scoring rules (Parry et al., 2012), we felt it contained a number of deep theoretical ideas but we weren't quite sure when these ideas would see the light of day in applications. For example, the fact that local scoring rules (except for the log score) do not depend on the normalization of the likelihood means that inference should be possible in unnormalized statistical models. Since 2012, work has been carried out on scoring rule inference in general (Mameli and Ventura, 2015; Dawid et al., 2016) and there have been papers on inference in graphical models (Yu et al., 2016, Yu et al., 2018), estimators for directional distributions (Mardia et al., 2016), and on Bayesian model selection with improper priors (Dawid and Musio, 2015). It was very nice therefore to see how Leisen et al. had used local scoring rules to derive a new class of objective priors.

In addition to commending the authors on their very stimulating paper, the intention of my discussion piece is to put some of the material in a slightly more general context. The key idea is the connection between scoring rules and entropy (see Gneiting and Raftery, 2007, and references therein).

1 Generalized uniform distributions and maximum entropy distributions

The objective priors defined in Sections 2 and 3 of the paper satisfy

$$S(\theta, p(\theta)) = \text{constant}, \tag{1.1}$$

and can be thought of as "generalized uniform distributions". This is in analogy with the generalized exponential distributions of Grünwald and Dawid (2004) (see also Dawid, 2007) that satisfy

$$S(\theta, p(\theta)) = \beta_0 + m(\theta) + \sum_i \beta_i t_i(\theta). \tag{1.2}$$

On the other hand, the objective priors outlined in Section 4, are maximum entropy distributions since they minimize the negative entropy functional

$$I(p) = \int L(\theta, p, p')\mathrm{d}\theta. \tag{1.3}$$

As noted in the paper, if, for almost all θ, $L(\theta, p, p')$ is jointly convex in p and p', then the $p(\theta)$ that minimizes $I(p)$ satisfies the Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial L}{\partial p'} - \frac{\partial L}{\partial p} = 0. \tag{1.4}$$

*Dept of Mathematics & Statistics, University of Otago, Dunedin, New Zealand, mparry@maths.otago.ac.nz

It turns out that these two approaches coincide precisely for local scoring rules – and not just the example considered in the paper. The reason for this is that $I(p)$ generates a scoring rule: if, for almost all θ, $L(\theta, p, p')$ is jointly (strictly) convex in p and p', then[1]

$$S(\theta, p(\theta)) = \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial L}{\partial p'} - \frac{\partial L}{\partial p} + \int \left\{ p(\theta)\frac{\partial L}{\partial p} + p'(\theta)\frac{\partial L}{\partial p'} - L \right\} \mathrm{d}\theta \tag{1.5}$$

is a (strictly) proper scoring rule. For non-local scoring rules the integral gives rise to a p-dependent constant; for local scoring rules, the integrand can only ever be proportional to $p(\theta)$ and so the integral is a constant independent of p. This is because for a local scoring rule, $L(\theta, p, p')$ is a weighted sum of 1-homogeneous functions[2] of p and p', and the function $p \log p$. The 1-homogeneous functions give zero contribution to the integrand; the function $p \log p$ generates the log score and gives the contribution proportional to $p(\theta)$.

My real point here is that locality of the scoring rule is not really a requirement for the approach of Sections 2 and 3, or the approach of Section 4. In fact, the paper of Leisen et al. suggests one could actually derive objective priors from any scoring rule (generalized uniform distribution) or its associated entropy (maximum entropy distribution).

2 Invariant distributions

Leisen et al. raise the interesting question as to whether an objective prior should be invariant under transformation of the parameters. This is where being local may actually help. We showed in Parry et al. (2012), that local scoring rules (except for the log score) are invariant under transformation of the data, or, in this case, the parameters. More precisely, given $\bar{\theta} = \bar{\theta}(\theta)$ and $\bar{p}(\bar{\theta}) = p(\theta)|\frac{\partial \bar{\theta}}{\partial \theta}|^{-1}$, then

$$\bar{S}(\bar{\theta}, \bar{p}) = S(\theta, p) \tag{2.1}$$

is a local scoring rule. This invariance is also inherited by the entropy since $I(p) = -\mathbb{E}_{\theta \sim p} S(\theta, p)$.

References

Dawid, A. (2007). "The geometry of proper scoring rules." *Annals of the Institute of Statistical Mathematics*, 59: 77–93. MR2396033. doi: https://doi.org/10.1007/s10463-006-0099-8. 1374

Dawid, A. P. and Musio, M. (2015). "Bayesian model selection based on proper scoring rules." *Bayesian Analysis*, 10(2): 479–499. MR3420890. doi: https://doi.org/10.1214/15-BA942. 1374

[1] This glosses over issues of boundary conditions that are somewhat tricky to specify but, in practice, can usually be checked after the fact; see Parry et al. (2012); Ehm and Gneiting (2012).

[2] A function $f(x, y)$ is 1-homogeneous if $f(\lambda x, \lambda y) = \lambda f(x, y)$, for $\lambda > 0$. As a consequence, $f(x, y) = x f_x(x, y) + y f_y(x, y)$.

Dawid, A. P., Musio, M., and Ventura, L. (2016). "Minimum scoring rule inference." *Scandinavian Journal of Statistics*, 43(1): 123–138. MR3466997. doi: https://doi.org/10.1111/sjos.12168. 1374

Ehm, W. and Gneiting, T. (2012). "Local proper scoring rules of order two." *Annals of Statistics*, 40(1): 609–637. MR3014319. doi: https://doi.org/10.1214/12-AOS973. 1375

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 1374

Grünwald, P. D. and Dawid, A. P. (2004). "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory." *Annals of Statistics*, 32(4): 1367–1433. MR2089128. doi: https://doi.org/10.1214/009053604000000553. 1374

Mameli, V. and Ventura, L. (2015). "Higher-order asymptotics for scoring rules." *Journal of Statistical Planning and Inference*, 165: 13–26. MR3498291. doi: https://doi.org/10.1016/j.jspi.2015.03.005. 1374

Mardia, K. V., Kent, J. T., and Laha, A. K. (2016). "Score matching estimators for directional distributions." arXiv:1604.08470 1374

Parry, M., Dawid, A. P., and Lauritzen, S. (2012). "Proper local scoring rules." *Annals of Statistics*, 40(1): 561–592. MR3014317. doi: https://doi.org/10.1214/12-AOS971. 1374, 1375

Yu, M., Kolar, M., and Gupta, V. (2016). "Statistical inference for pairwise graphical models using score matching." In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, 2829–2837. Curran Associates, Inc. 1374

Yu, S., Drton, M., and Shojaie, A. (2018). "Graphical models for non-negative data using generalized score matching." *Proceedings of Machine Learning Research*, 84: 1781–1790. MR3960930. 1374

Invited Discussion

Guido Consonni[*,‡] and Gonzalo García-Donato[†,‡,§]

1 General aspects

This paper is a welcome addition to the research on objective Bayes (OB) methods. The Authors (A's) summarize their main motivations under two main headings:

1) current methodology for building OB priors is predicated on the statistical model; as models grow in complexity, so does the difficulty in obtaining OB priors. A related issue is model misspecification which would naturally propagate to OB priors;

2) OB priors are often improper. This is seen by the A's as a serious concern, because of a host of difficulties. In particular assessing posterior properness becomes onerous in complex models; also improper priors cannot be used (in general) for model comparison *via* Bayes Factor due to their dependence on arbitrary normalizing constants.

Because of 1 and 2 the A's set out their program as finding OB priors which only depend on the parameter space, thus severing the connection with the model. This is a major departure from more traditional principles to formalize ignorance such as *context invariance* (Dawid, 2006) which states that only the formal structure of the distributional model of the observables should matter when specifying the prior (model-dependent priors).

The content of this discussion is as follows. In Section 1 we examine some structural aspects of the criterion proposed to define an objective prior. In Section 2 we assess a prior employed by the A's in their paper from the point of view of testing. Specifically we consider two scenarios: i) testing normality *versus* Cauchy errors in a problem with several means (a non-nested setup); ii) testing a sharp null hypothesis against an unrestricted alternative on the mean of a normal model with unknown variance (a nested setup).

The broad conclusion we reach is that the method is too general as it stands to provide useful guidelines especially in applied work. We also believe that the structure of the statistical model cannot be easily dispensed with when constructing an objective prior, and that its neglect may generate undesirable features especially in a testing scenario.

[*]Università Cattolica del Sacro Cuore, Milan (Italy), guido.consonni@unicatt.it
[†]Universidad de Castilla-La Mancha (Spain), gonzalo.garciadonato@uclm.es
[‡]Partially supported by UCSC Research track D1 and D.3.2.
[§]Partially supported by Ministerio de Ciencia e Innovación (MCI, Spain) grant PID2019-104790GB-I00 and Junta de Comunidades de Castilla-La Mancha grant SBPLY/17/180501/000491/2.

1.1 Is the constant scoring rule criterion too weak?

The A's proposal starts with a specific proper *scoring rule*, which measures the quality of a quoted probability density p for the parameter $\theta \in \Theta$; next they look for those p's such that the score is *constant* over Θ. The intuition is that if the score were not constant, then some parts of the parameter space would be given preferences over the remaining ones, thus violating noninformativity and objectivity.

It turns out that the solution to the above program is a whole *class* of priors (which always includes the *uniform* prior, irrespective of the nature of the parameter space, e.g. the real line, half the real line, or a bounded interval). The A's see this as an advantage, in that it adds flexibility to their method. They write (Section 5): "... given a particular problem, the objective prior over a given parameter space could be proper or improper; differentiable everywhere or not; convex; log-concave; etc. In other words, a prior can be objective and exhibit desirable features of choice without impinging on subjective components relating to information". While the above quotation encourages optimism, from a pragmatic viewpoint it is troublesome because it requires supplementing the basic criterion, based solely on a scoring rule, with additional specifications. The latter appear a critical aspect of the method, and the illustrations provided in Section 5 (A's paper) are paradigmatic.

For the case $\Theta = (0, 1)$, a possible specification is to set $p(0) = p(1) = 0$ with additionally symmetry leading to a concave function. On the other hand, it is also possible to obtain a distribution which mimics Jeffreys prior, i.e. convex and unbounded in the neighborhood of 0 and 1. Unfortunately no guidelines are available to set up these additional features, especially from an objective viewpoint. Another concern is that the resulting prior seems somewhat sensitive to numerical specifications of the base function $u(\cdot)$ at some specific points; see plot (a) and (b) in Figure 2 (A's paper) where the former holds for $u(1/2) = 1.1$ and the latter for $u(1/2) = 1.14$. It is hard to believe that one might sensibly distinguish between values of a prior density at a single point with such a high level of precision.

As another illustration the A's consider the case $\Theta = (0, \infty)$ where they require a specific shape property (convex and decreasing), the reason being that "this property is common to most objective priors on $(0, \infty)$". What they fail to add however is that this property usually stems from considerations involving the model, as in the Jeffreys' prior. While this is admittedly a mere illustration of their methodology, it reflects the pragmatic problem of specifying sensible additional features which are not model-dependent.

1.2 Does invariance matter?

The method proposed by the A's is not invariant to re-parametrizations. Yet they contend that this is not an issue because, for any given model, one usually works with a specific parameter throughout the analysis. However we remark that often the choice of a parametrization is conventional, as in the example they quote, namely a normal model parametrized either by the variance or the precision. An important point about invariance to re-parametrization is that the prior and posterior predictive of the observ-

ables will be unchanged whatever the parametrization; this desirable feature is lost if a method produces priors which are not probabilistically equivalent.

The A's concede that if invariance is a concern, then one can always transform by change of variable the prior obtained under one parametrization. However this begs the question of where to start from. The whole idea of invariance is that it does not matter which parameter we start with because the analysis will be equivalent in the end. In this context, they consider the natural exponential family with canonical parameter θ. Because of the special status enjoyed by θ, one could apply the method to θ, and then induce the resulting prior onto any other parameter ϕ say. However if θ is a canonical parameter, so is any affine transformation of θ; so their method should be at least invariant to affine transformation. Is this the case? (In the paper the A's state that their method is invariant to location transformations).

2 Priors for testing

How do priors based on scoring rules (SR-priors) behave in a problem involving hypothesis testing? In this section we assess a proper SR-prior on the real line presented by the A's in two testing scenarios.

2.1 Priors for common parameters

Example 1 We consider a measurement-error model applied to k populations with unknown distinct means. For each population n independent replicates are available. Errors can be either normal or Cauchy (t_1). Specifically

$$x_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n,$$

and the testing problem is $H_0 : \varepsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, 1)$ *versus* $H_1 : \varepsilon_{ij} \overset{iid}{\sim} t_1$. Notice that the unknown k-dimensional parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ appears in both competing models, and can be regarded as a common parameter. To compute the Bayes factor of H_1 to H_0, priors are required, and we denote them with $\pi_0(\boldsymbol{\mu})$ (under H_0) and $\pi_1(\boldsymbol{\mu})$ (under H_1). If $\pi^{SR}(\mu_i)$ denotes an objective prior based on the scoring-rule methodology for $\mu_i \in R$, a prior for the vector $\boldsymbol{\mu}$ can be constructed as

$$\pi_0^{SR}(\boldsymbol{\mu}) = \pi_1^{SR}(\boldsymbol{\mu}) = \prod_{i=1}^{k} \pi^{SR}(\mu_i). \tag{2.1}$$

Specifically we take π^{SR} to be the proper prior presented by the A's in Figure 3b of their paper. Notice that it is unimodal with the mode at zero.

To assess both priors, we apply the criterion of predictive matching introduced by Jeffreys (1961), and subsequently revisited by Bayarri et al. (2012). Informally it states that, for samples of minimal size, one should not be able to clearly discriminate between two models –the Bayes factor should be close to one–. In this Example, a minimal sample size is $n = 1$, that is one observation per group. Accordingly we write such minimal

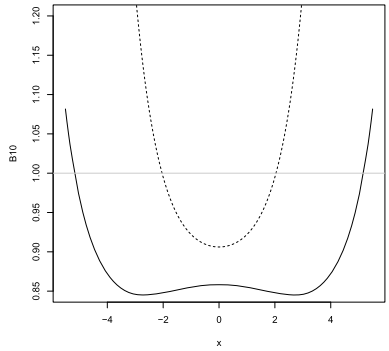


Figure 1: Example 1: Bayes factor $\mathcal{B}^{SR}$ under SR-prior (solid line) and under $N(0,1)$ prior (dashed line).

sample as $\boldsymbol{x}^\star = (x_{11}^\star, x_{21}^\star, \ldots, x_{k1}^\star)$. The Bayes factor associated to $\boldsymbol{x}^\star$ and the priors in (2.1) is

$$B_{10}^{SR}(\boldsymbol{x}^\star) = \prod_{i=1}^{k} \mathcal{B}^{SR}(x_{i1}^\star).$$

A plot of $\mathcal{B}^{SR}(x)$ is presented in Figure 1. For values of $|x| \leq 5$, $\mathcal{B}^{SR}(x)$ ranges in [0.85, 1.05]. This implies that violation of the predictive matching criterion could be large even for a moderate number of populations. For instance, if $k = 4$ and $x_{i1}^\star$'s are small then $B_{10}^{SR}(\boldsymbol{x}^\star) \approx 0.85^4 \approx 0.44$. We experienced computational problems when trying to study the behavior of $\mathcal{B}^{SR}$ when the x_{i1}'s are large, because of the implicit definition of the SR prior. While this remains an open question, our conjecture is that $\mathcal{B}^{SR}(x)$ will grow to infinity as $|x|$ grows given the different nature of the tails of the t_1 and $N(0,1)$, similarly to what happens with other proper priors like the standard normal (see dashed line in Figure 1).

On the other hand, it can be easily seen that the Bayes factor associated with the improper priors $\pi_0(\boldsymbol{\mu}) = \pi_1(\boldsymbol{\mu}) = 1$ is always unitary (independently of $\boldsymbol{x}^\star$), so that they satisfy predictive matching. (Notice that the arbitrary constant inherent in the improper priors cancels out when taking ratios). An additional argument in favor of these priors connects with the family of intrinsic Bayes factors proposed by Berger and Pericchi (1996). It is a remarkable fact that if $\pi_0(\boldsymbol{\mu}) = \pi_1(\boldsymbol{\mu}) = 1$ then *all* intrinsic Bayes factors coincide and are equal to one.

In conclusion the SR-prior used by the A's in Figure 3 of their paper fails to achieve predictive matching between two models involving several common location parameters. On the other hand the standard flat improper priors achieve this goal, and the computational effort is trivial. Of course it could be argued that flat priors on the real line are also SR-priors. The issue however remains: how to decide on this choice which looks optimal in this case? Predictive matching cannot be invoked as a property to be satisfied because it involves the statistical model, thus undermining the premises set out by the A's.

2.2 Nested models

The assignment of an objective prior for a multidimensional parameter is treated only briefly in the paper. However, possibly to alleviate the burden of the computation of the SR-prior, the assumption of independence of the parameter components seems like a natural choice in their approach; see for instance the preamble of Section 5. On the other hand Section 5.2 suggests to take marginal SR-prios as inputs for the construction of the overall prior. In this subsection we discus objective priors for vectors of parameters in the context of testing in view of assessing whether independence of the components is in general a sensible choice.

Consider the classic example of testing the mean of a normal model with unknown variance.

Example 2. Let $x_i \sim N(\mu, \sigma^2)$, independently for $i = 1, \ldots, n$, and consider the testing problem $H_0 : \mu = 0$ *versus* $H_1 : \mu \neq 0$. Objective priors $\pi_0(\sigma)$ and $\pi_1(\mu, \sigma)$ have to be specified.

Since we want to investigate the impact of assuming independence *a priori* between (μ, σ), we take the SR-prior to be $\pi_1^{SR}(\mu, \sigma) = \pi_1^{SR}(\mu)\pi_1(\sigma)$, where for concreteness the SR prior for $\mu \in (-\infty, \infty)$ is the proper prior we employed in the previous subsection. For the marginal priors on the common parameter σ we use for both models the standard objective prior $\pi(\sigma) = \sigma^{-1}$. This is done for simplicity and should not affect the take-home message of this example.

The Bayes factor can be expressed as the univariate integral

$$B_{10}^{SR}(\bar{x}, S) = \int \Big(\frac{\bar{x}^2 + S^2}{(\bar{x} - \mu)^2 + S^2} \Big)^{n/2} \pi^{SR}(\mu) d\mu,$$

(where $\bar{x}$ and S^2 are the sample mean and sample variance). It can be shown that the minimal training sample associated with these priors is $n = 2$. We have represented, in Figure 2 (left panel) B_{10}^{SR} as a function of $\bar{x}$, and for $n = 2$ and a few values of S. Because of symmetry, the plot is presented only for $\bar{x} \geq 0$. We immediately conclude that $\pi_1^{SR}(\mu, \sigma)$ is not predictive matching. Nevertheless, one could argue that this is to be expected because a sample with $n = 2$ might well contain some information. What is worrisome however is the behavior of B_{10}^{SR} exhibited in the left panel of Figure 2: this is not monotone increasing in $\bar{x}$ –for fixed S and n– and this holds as n increases; clearly an undesirable feature.

Now consider a conventional objective prior for this problem

$$\pi_0(\sigma) = \pi_1(\sigma) = \sigma^{-1}, \;\; \pi_1(\mu \mid \sigma) = \text{Cauchy}(\mu \mid 0, \sigma^2). \tag{2.2}$$

This prior was first proposed by Jeffreys (1961) who derived it based on a certain desirable properties. Berger and Pericchi (2001) named it 'conventional' and later Bayarri and García-Donato (2008) showed that it can be obtained applying a general rule based on the Kullback-Leibler divergence. This prior leads to the Bayes factor

$$B_{10}(\bar{x}, S) = \int (1 + nt)^{(n-1)/2} \Big(1 + nt \frac{S^2}{\bar{x}^2 + S^2} \Big)^{-n/2} \text{IGa}(t \mid 0.5, 0.5)\, dt,$$

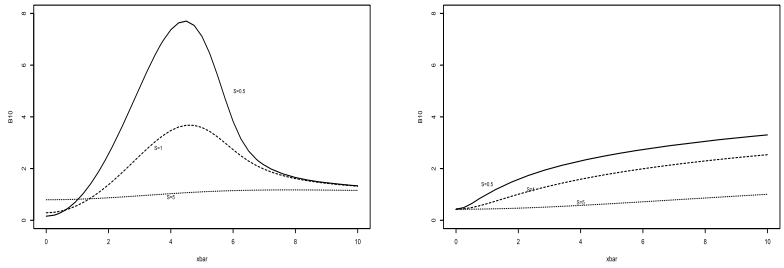


Figure 2: Example 2. B_{10}^{SR} (left) and B_{10} (right) for $n = 2$ as functions of $\bar{x}$ and for a few values of S. (Both Bayes factors are symmetric around zero with respect to $\bar{x}$ and only positive values are represented).

where $IGa(a, b)$ stands for the inverse gamma density with shape a and rate b. The corresponding minimal training sample size is $n = 1$, for which we obtain a unitary Bayes factor (see expression above with $S = 0$), so the conventional prior is exact predictive matching. For comparison purposes we also computed the conventional Bayes factor for $n = 2$, whose plot is reported in the right panel of Figure 2. Its behavior appears quite sensible; in particular as $\bar{x}$ gets larger the evidence in favor of H_1 increases.

The SR-prior we used in Example 2 was constructed by assuming independence between μ and σ, which we surmise might explain the unsatisfactory behavior of B_{10}^{SR} depicted in Figure 2. A crucial feature of the conventional prior leading to its sensible testing performance in this problem is that the prior on μ scales by σ. While this seems to be a sensible option in location-scale problems, prior dependence in other problems has to be carefully assessed. Inevitably however, intrinsic features of the entertained models will come to the fore (see the invariance criterion in Bayarri et al., 2012). It would be interesting to see whether the A's methodology could produce an SR-prior with such adaptive structural property without invoking aspects of the statistical model for its construction.

References

Bayarri, M., Berger, J., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1379, 1382

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*, 70(5): 981–1003. MR2530326. doi: https://doi.org/10.1111/j.1467-9868.2008.00667.x. 1381

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): pp. 109–122. URL http://www.jstor.org/stable/2291387. MR1394065. doi: https://doi.org/10.2307/2291387. 1380

Berger, J. O. and Pericchi, L. R. (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison." In Lahiri, P. (ed.), *Model Selection*, volume 38, pp. 135–207. Institute of Mathematical Statistics. URL http://www.jstor.org/stable/4356165. MR2000750. 1381

Dawid, A. P. (2006). "Invariant Prior Distributions." In *Encyclopedia of Statistical Sciences*. Wiley, New York. 1377

Jeffreys, H. (1961). *The Theory of Probability*. Oxford Univ. Press. MR0187257. 1379, 1381

Invited Discussion

Federica Giummolè* and Laura Ventura†

We congratulate the authors for this interesting contribution to the wide world of objective priors (see Consonni et al., 2018, for a recent review). The authors tackle the problem of providing an objective prior which is model-free and based on the sole knowledge of the parameter space. We think that the main result can be a useful practical tool for objective Bayesian analysis in many applications and can open new ideas about objective priors.

With our discussion, we hope to shed light on some aspects of the proposed approach, which is based on seeking a prior such that a combination of the log-score and of the Hyvärinen scoring rule is constant. In particular, we briefly comment on the following points:

1. extensions of the proposed approach using different scoring rules, and objectiveness and invariance of the proposed prior densities;
2. double use of the Hyvärinen scoring rule, both for the derivation of the prior and to replace the likelihood function in models known up to the normalization constant.

1 Background on proper scoring rules

Consider a random sample $y = (y_1, \ldots, y_n)$ of size n from a parametric model with probability density function $f(y|\theta)$, indexed by a k-dimensional parameter θ. A proper scoring rule (SR) $S(y, f)$ provides a way of judging the quality of a quoted model $f(y|\theta)$ for a random variable Y in the light of its outcome y. The mathematical theory of proper SRs has a wide range of applications in statistics; a review of the general theory, with applications, has been given in Dawid and Musio (2014). SRs are particularly useful when classical likelihood-based methods may be infeasible, for example in models with complex dependency structure, or when robustness with respect to data or to model misspecification is required.

There is a very wide variety of SRs. The most famous is the logarithmic score or *log-score*, which is highly connected with likelihood inference. Proper SRs, different from the log-score, can be used as an alternative to the full likelihood, when the interest is in increasing robustness or simplifying computations. Examples of particular interest include the general *separable Bregman score* (see e.g. Dawid, 2007, eq. 16) given by

$$S(y, f) = -\psi'\{f(y|\theta)\} - \int [\psi\{f(y|\theta)\} - f(y|\theta)\psi'\{f(y|\theta)\}]\, dy, \tag{1}$$

*Ca' Foscari University Venice, Venice, Italy, giummole@unive.it
†University of Padova, Padua, Italy, ventura@stat.unipd.it

where the defining function $\psi : \mathbb{R}^+ \to \mathbb{R}$ is convex and differentiable. Taking, respectively, $\psi(t) = t^2$ and $\psi(t) = t \log t$ the *Brier score* and the *log-score* are obtained. Another important special case of this construction arises when $\psi(t) = t^\gamma$ ($\gamma > 1$). This yields the *Tsallis score* (Tsallis, 1988)

$$S(y, f) = (\gamma - 1) \int f(y|\theta)^\gamma \, dy - \gamma f(y|\theta)^{\gamma - 1}, \quad \gamma > 1, \tag{2}$$

which gives in general robust procedures (see e.g. Dawid et al., 2016), where the parameter γ is a trade-off between efficiency and robustness. The density power divergence d_α of Basu et al. (1998) is just (2), with $\gamma = \alpha + 1$, multiplied by $1/\alpha$.

In the case of a real sample space, the Hyvärinen scoring rule

$$S(y, f) = 2 \frac{\partial^2 \log f(y|\theta)}{\partial y^2} + \left| \frac{\partial \log f(y|\theta)}{\partial y} \right|^2 \tag{3}$$

satisfies the property of homogeneity, which implies that the quoted distribution need only to be known up to the normalization constant (see Ehm and Gneiting, 2012; Parry et al., 2012).

Proper scoring rules can also be extended to the case of a random vector. Let $\{Y_k\}$ be a set of marginal or conditional variables with associated proper scoring rule S_k. A proper scoring rule for the random vector Y is defined as $S(y, f) = \sum_k S_k(y_k, f_k)$, where $X_k \sim f_k$ when $Y \sim f$, and y and y_k are the values assumed by Y and Y_k, respectively. Scoring rules of this form are called *composite scoring rules*; see Dawid and Musio (2014) and Dawid et al. (2016). Note that when each S_k is the log-score, then $S(y, f)$ is a negative composite log-likelihood (see Varin et al., 2011).

2 Priors from the log-score and the Hyvärinen scoring rule

Consider, for simplicity of notation, a scalar parameter θ. The method proposed by Leisen, Villa and Walker considers to seek a prior $p(\theta)$ on $\theta \in \Theta$ such that a combination of the log-score and the Hyvärinen scoring rule is constant, that is

$$S(\theta, p(\theta)) = \text{constant} \quad \forall \theta \in \Theta, \tag{4}$$

where

$$S(\theta, p(\theta)) = -w \log p(\theta) + \frac{p''(\theta)}{p(\theta)} - \frac{1}{2} \left(\frac{p'(\theta)}{p(\theta)} \right)^2.$$

Here w is a weighting factor usually taken equal 1. The resulting objective prior $p_u(\theta)$, that we will call in the following *u-prior*, takes the form $p_u(\theta) \propto \exp\{-u(\theta)\}$, where the function $u(\theta)$ is obtained by solving the differential equation

$$u'(\theta) = \pm \sqrt{c e^{u(\theta)} - 2(1 + u(\theta))}, \tag{5}$$

for some suitable constant c and a specified value of $u(\theta)$ at some point, e.g. $u(0)$. Typically the prior $p_u(\theta)$ is obtained via numerical methods, even in simple cases.

Objectiveness In the practice, the u-prior defines a class of priors, since it depends on the constraints $u(0)$ and c, that have to be suitably fixed. Is this in contrast with an objective Bayes method? Indeed, as shown in the example in the next Section 3, the choice of the constraints $u(0)$ and c may have a great impact on the resulting u-prior and thus on the posterior distribution. These two constraints are in practice two hyper-parameters of the proposed prior and it seems that their choice makes the proposed prior less "objective". Moreover, not only for the parametric space $(0, \infty)$ but also for the case $(-\infty, +\infty)$, for some choices of $u(0)$ and c the u-prior often lies in a limited support, thus being very informative about the unknown parameter. When the sample size n is small or moderate, this may have a great impact on the corresponding posterior.

Changing the scoring rule The idea behind (4) is very appealing and can potentially be applied to different scoring rules, such as the log-score, the Tsallis (2) or the general Bregman scoring rules (1). Unfortunately, as also noticed by Leisen, Villa and Walker for the log-score, in all these cases the resulting prior is constant and thus not very useful in the practice. More interesting priors are possibly obtained by combining different SRs, as in the paper proposal. In particular, it could be interesting to investigate combinations of the Hyvärinen SR, which involves first and second order derivatives of $p(\theta)$, with some SR different from the log-score.

Invariance An important point of discussion about prior distributions, and in particular objective priors, is invariance. Jeffreys' rule to derive a prior distribution for the parameter of a given model is based on an invariance with respect to one-to-one changes in the parametrization. Other common objective priors, such as reference priors, have been shown to be invariant, and the same applies to priors obtained from α-divergences (Giummolè et al., 2019).

Let us focus on invariance with respect to one-to-one changes in the parametrization. Let $\psi(\theta)$ be a reparametrization, with inverse $\theta(\psi)$. Then $p_\psi(\psi) = p_\theta(\theta(\psi))|\theta'(\psi)|$ is the prior for ψ obtained by transforming $p_\theta(\theta)$. If we seek to derive an invariant prior from (4), we have to require that

$$S(\theta, p_\theta(\theta)) = \text{constant} \quad \forall\theta \qquad \Longleftrightarrow \qquad S(\psi, p_\psi(\psi)) = \text{constant} \quad \forall\psi,$$

for every reparametrization $\psi(\theta)$. Fulfillment of this requirement may depend on the particular SR considered. Anyway, it can be easily shown that the previous condition is not satisfied for the most common SRs mentioned above, nor for the mixture of the log-score and the Hyvärinen scoring rule proposed by Leisen, Villa and Walker. Instead, invariance is usually satisfied with respect to the restricted class of linear transformations of the parameter, for which $\theta'(\psi) = \text{constant}$. For this reason, we believe that the proposed method is particularly useful for inference on scale and location models, where the induced family of transformations in the parametric space is that of affine transformations for the location parameter and multiplicative changes for the scale parameter.

3 Double scoring rule in the posterior

Standard Bayesian analyses can be unpleasant when robustness with respect to data or to model misspecifications is required or in models with complex dependency structures. To deal with these issues the use of a surrogate likelihood in the Bayes formula has received considerable attention in the last decade (see the review by Ventura and Racugno 2016, and references therein). In particular, Bayesian inference based on scoring rules has been considered in Ghosh and Basu (2016); Bissiri et al. (2019); Giummolè et al. (2019), and Girardi et al. (2020); see also references therein.

Let $S(\theta) = \sum_{i=1}^{n} S(y_i, f)$ be the total score for θ, and let $\tilde{\theta}$ be the scoring rule estimator given by $\arg\min_\theta S(\theta)$. This estimator is asymptotically normal, with mean θ and covariance matrix $V(\theta) = K(\theta)^{-1}J(\theta)(K(\theta)^{-1})^T$, where $K(\theta)$ and $J(\theta)$ are the sensitivity and the variability matrices, respectively. The matrix $G(\theta) = V(\theta)^{-1}$ is known as the Godambe information matrix, and in the case of the log-score, we have that $G(\theta) = K(\theta) = J(\theta)$ is the Fisher information matrix. A *SR-posterior distribution* can be obtained by using the total score $S(\theta)$ instead of the full likelihood in Bayes formula. Let $p(\theta)$ be a prior distribution for the parameter θ. The SR-posterior distribution is defined as

$$p(\theta|y) \propto p(\theta)\, \exp\{-S(\theta^*)\}, \tag{6}$$

with $\theta^* = \theta^*(\theta) = \tilde{\theta} + C(\theta - \tilde{\theta})$, where C is a $d \times d$ fixed matrix (see Giummolè et al., 2019, for details).

The choice of a prior distribution $p(\theta)$ to be used in (6) involves the same problems typical of the standard Bayesian perspective. For objective Bayesian inference, a prior can be chosen such that the expected α-divergence to the SR-posterior distribution is maximized (Giummolè et al., 2019). The α-divergences are a well-known class of discrepancy functions which include as a special case the Kullback-Leibler divergence. For $0 \leq |\alpha| < 1$, a Jeffreys-type prior is derived, that is proportional to the square root of the determinant of the inverse of the asymptotic covariance matrix of $\tilde{\theta}$, i.e. $p_G(\theta) \propto |G(\theta)|^{1/2}$. This G-prior is shown to be invariant with respect to one-to-one changes in the parameterization.

In the next example, we explore the use of the Hyvärinen scoring rule twice in order to derive a posterior distribution: first to construct a model-free prior as suggested by Leisen, Villa and Walker and second to replace the likelihood function when interest is, for instance, in simplifying computations in complex models. Note however that a SR-posterior may be obtained also using different scoring rules than the Hyvärinen. In particular, when using the Tsallis scoring rule a robust SR-posterior can be derived, or the composite log-score can be usefully considered to deal with models with complex dependency structures.

Example: Directional models Inference for directional models is difficult because typically the density function contains an intractable normalization constant, which cannot be explicitly computed in closed form. In this setting and to avoid the issue of the intractable normalising constant, Mardia et al. (2016) propose to use the Hyvärinen

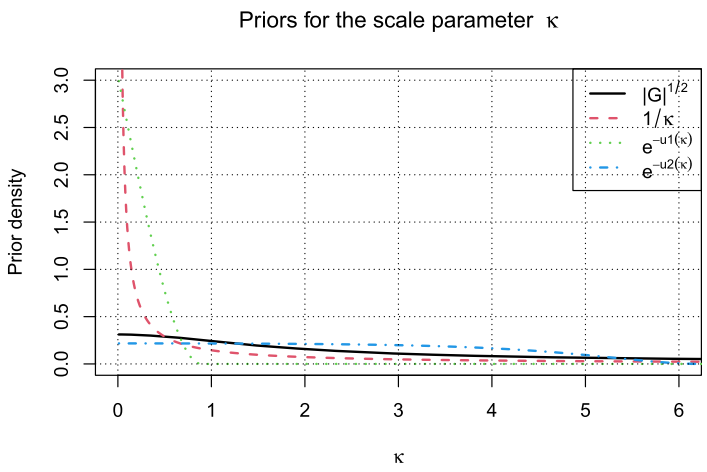


Figure 1: Priors $p(\kappa) \propto 1/\kappa$, $p_G(\kappa) \propto |G(\kappa)|^{1/2}$, $p_{u_1}(\kappa) \propto \exp\{-u_1(\kappa)\}$ and $p_{u_2}(\kappa) \propto \exp\{-u_2(\kappa)\}$.

scoring rule. In particular, let us consider the von Mises-Fisher density, which is a directional distribution defined on the unit sphere $\mathcal{S}_{q-1} \subset \mathbb{R}^q$ given by

$$p(y|\kappa) \propto \exp\{-\kappa \mu^T y\}, \quad y \in \mathcal{S}_{q-1},$$

with $\kappa \in \mathbb{R}^+$ a scalar concentration parameter and μ the mean direction, $||\mu|| = 1$. In this example we consider $q = 2$ and $\mu = (0, 1)$, known.

We discuss three different priors for κ: the classical non-informative prior $p(\kappa) \propto 1/\kappa$, the G-prior $p_G(\kappa) \propto \sqrt{A_1^2(\kappa)/(\kappa\,[2\kappa - 3A_1(\kappa)])}$, with $A_1(\kappa) = I_1(\kappa)/I_0(\kappa)$, where I_0 and I_1 are the modified Bessel functions of order 0 and 1, respectively, and the u-prior $p_u(\kappa)$ defined on the space $(0, +\infty)$, where $u(\kappa)$ is obtained as the solution of (5). In particular we consider two u-priors: 1. u_1-prior with $u(0) = 1.31$ and $c = 2$; 2. u_2-prior with $u(0) = 0.01$ and $c = 2(1 + u(0)) \exp\{-u(0)\}$. Both these u-priors are suggested in Leisen, Villa and Walker (Section 5.1). The four priors are depicted in Figure 1. It can be seen that the G-prior and the u_2-prior are similar on a bounded interval, while the u_1-prior has a very limited support. The choice of the constraints $u(0)$ and c has thus a great impact on the u-prior, and in particular, when fixing $u(0) = 1.31$ and $c = 2$ we obtain a very informative prior on $(0, +\infty)$.

Figure 2 shows the four SR-posteriors for different values of the sample size n and the parameter κ. It can be noted that the SR-posterior obtained with $p(\kappa) \propto 1/\kappa$ may not be proper or puts too much mass at zero. Moreover, the SR-posteriors based on the G-prior and on the u_2-prior are very similar when the true value of the parameter is 1 or 5. The SR-posterior obtained with the u_1-prior appears centred away from the true value of the parameter. This latter prior may be completely misleading when the true value of the parameter is larger than 1. Finally, for $\kappa = 10$, the SR-posterior based on the G-prior still gives sensible results, while both the u-priors fail to give a useful

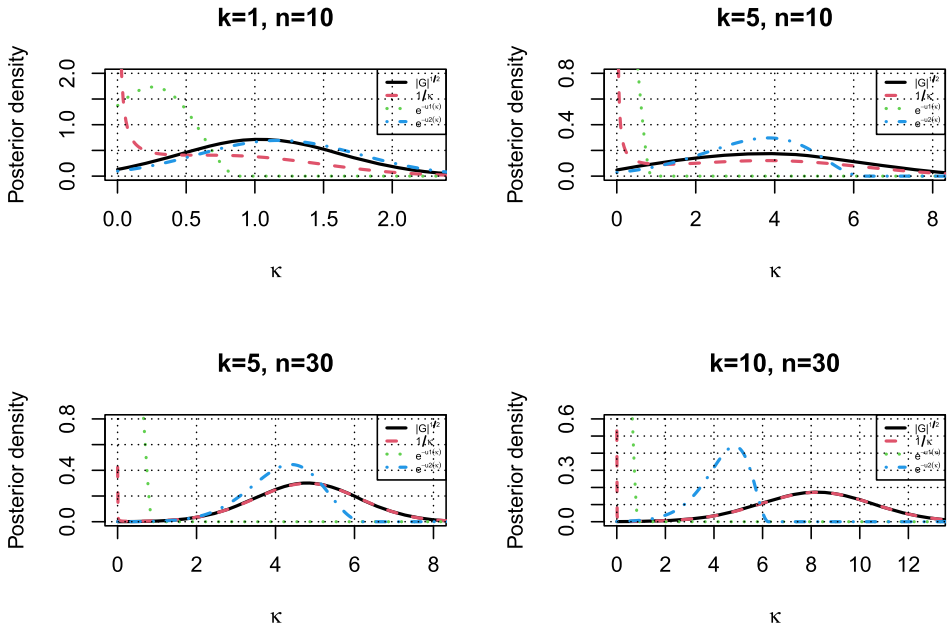


Figure 2: SR-posteriors with different priors, and values of n and κ.

posterior. Indeed, since both the u-priors have a limited support, when the true value of the parameter is big enough (larger than 6) the resulting posteriors are misleading.

References

Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559. 1385

Bissiri, P. G., Walker, S. G. (2019). On general Bayesian inference using loss functions. *Statistics & Probability Letters*, **152**, 89–91. MR3952612. doi: https://doi.org/10.1016/j.spl.2019.04.005. 1387

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, **13**, 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 1384

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics* **59**, 77–93. MR2396033. doi: https://doi.org/10.1007/s10463-006-0099-8. 1384

Dawid, A. P., Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, **72**, 169–183. MR3233147. doi: https://doi.org/10.1007/s40300-014-0039-y. 1384, 1385

Dawid, A. P., Musio, M., Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, **43**, 123–138. MR3466997. doi: https://doi.org/10.1111/sjos.12168. 1385

Ehm, W., Gneiting, T. (2012). Local proper scoring rules of order two. *Annals of Statistics*, **40**, 609–637. MR3014319. doi: https://doi.org/10.1214/12-AOS973. 1385

Ghosh, M., Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, **68**, 413–437. MR3464228. doi: https://doi.org/10.1007/s10463-014-0499-0. 1387

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., Ventura, L. (2020). Robust inference for nonlinear regression models from the Tsallis score: application to COVID-19 contagion in Italy. *Stat*, 9, e309. 1387

Giummolè, F., Mameli, V., Ruli, E., Ventura, L. (2019). Objective Bayesian inference with proper scoring rules. *Test*, **28**, 728–755. MR3992136. doi: https://doi.org/10.1007/s11749-018-0597-z. 1386, 1387

Mardia, K. V., Kent, J. T., Laha, A. K. (2016). Score matching estimators for directional distributions. arXiv:1604.08470v1. 1387

Parry, M., Dawid, A. P., Lauritzen, S. L. (2012). Proper local scoring rules. *Annals of Statistics*, **40**, 561–592. MR3014317. doi: https://doi.org/10.1214/12-AOS971. 1385

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487. MR0968597. doi: https://doi.org/10.1007/BF01016429. 1385

Varin, C., Reid, N., Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42. MR2796852. 1385

Ventura, L., Racugno, W. (2016). Pseudo-likelihoods for Bayesian inference. In: *Topics on methodological and applied statistical inference*, Studies in Theoretical and Applied Statistics, 205–220. Springer, Berlin. 1387

Acknowledgments

This work was supported by a grant from the University of Padova (BIRD197903).

Contributed Discussion

Alexander Y. Shestopaloff[*] and Silvia Liverani[†]

We thank the authors for an interesting and thought-provoking paper. In our discussion, we would like to point out several points that may be of interest to expand upon.

Sensitivity of inferences to prior specification The choice of c, $u(0)$ and w remains up to the user and is based on the desired properties of the prior. We wonder how one can further study this aspect of prior specification. An example of achieving additional flexibility for c and $u(0)$ can be through putting a prior distribution on any or all of c, $u(0)$ and w.

Another study of possible interest, related to the above point, would be to better understand how robust model inferences are to erroneous specification of c, $u(0)$ and w. That is, looking at the variation of functions of the posterior depending on changes in these parameters.

Invariance Consider a prior on the variance or the precision of a normal distribution. The authors claim that, whichever parameterisation is used, the corresponding objective prior is adequate for the purpose to which it has been assigned. However, we are not aware of a rationale in the literature for priors to be defined for the precision, apart from mathematical convenience (due to conjugacy or conditional conjugacy) in some cases. Therefore, we believe that invariance is an important aspect that merits study. We can use the Table 1 and Table 2 of the paper to choose a prior for the variance with a certain desired prior mean and prior variance. The prior for the precision would then be implied to have a certain mean and variance as well.

This means that in principle, we can derive what values of c and $u(0)$ we can use to ensure an approximate consistency between the variance and precision priors, in terms of matching at least the first 2 moments. This would allow for approximate invariance under a $1-1$ transformation of the parameter θ.

Approximate Bayesian Computation (ABC) Given that the authors introduce a mechanism for sampling from the introduced class of prior densities, it would be interesting to see how these objective priors can be used in the context of ABC. Specifically, when using these priors, what would the effect be on performance compared to priors that do not satisfy the constant score criterion.

[*]School of Mathematical Sciences, Queen Mary University of London, a.shestopaloff@qmul.ac.uk
[†]School of Mathematical Sciences, Queen Mary University of London, s.liverani@qmul.ac.uk

Contributed Discussion

Matteo Iacopini[*,§], Francesco Ravazzolo[†], and Luca Rossini[‡]

We have greatly appreciated the work by Leisen, Villa and Walker, who have proposed a novel method for constructing objective prior distributions which circumvents the need to specify a statistical model and relies solely on a specific proper scoring rule.

Scoring rules are of tantamount importance in practical statistical analysis, since they encode the preferences of the stakeholder (e.g., a policymaker). In fact, since scoring rules provide a simple mean to assess the performance of a set of statistical models with respect to a specific user-defined goal, they have been widely used as a tool for evaluation, comparison and ranking of competing statistical models.

Recently, Iacopini et al. (2020) proposed the asymmetric continuous probabilistic score (ACPS), a new proper scoring rule for evaluating density forecasts according to asymmetric preferences of the stakeholder. Being the ACPS a proper scoring rule, the methodology developed by the Authors can be directly applied to derive a class of objective prior distributions. Moreover, since ACPS has one free parameter specifying the type and degree of asymmetric preferences, the ensuing class of priors would inherit this degree of freedom. Intuitively, its role would be analogous to that of the free constants, c and $u(0)$, in Section 3 of the discussed Article.

To obtain the class of objective priors stemming from the ACPS proper scoring rule, first recall its definition. Let $P : \mathcal{D} \to [0,1]$ be a cumulative distribution function, with $\mathcal{D} \subseteq \mathbb{R}$, let $y \in \mathcal{D}$, and denote with $c \in (0,1)$ the asymmetry parameter. Then

$$\begin{aligned} ACPS(P,y;c) &= \int_{-\infty}^{y} (c^2 - P(u)^2) f(P,c)\,\mathrm{d}u + \int_{y}^{+\infty} ((1-c)^2 - (1-P(u))^2) f(P,c)\,\mathrm{d}u \\ &= \int_{\mathbb{R}} \Big[(c^2 - P(u)^2)\mathbb{I}(u<y) + ((1-c)^2 - (1-P(u))^2)\mathbb{I}(u>y) \Big] f(P,c)\,\mathrm{d}u, \end{aligned}$$

where $f(P,c) = \mathbb{I}(P(u) > c)/(1-c)^2 + \mathbb{I}(P(u) \leq c)/c^2$. Therefore, solving for P the equation $ACPS(P,y;c) = k$, with $k \in \mathbb{R}$, would be equivalent to solve the minimization problem

$$\min_{P} \left| \int_{\mathbb{R}} L(u,P,P')\,\mathrm{d}u - k \right|, \tag{1}$$

where $L(u,P,P') = \big[(c^2 - P(u)^2)\mathbb{I}(u<y) + ((1-c)^2 - (1-P(u))^2)\mathbb{I}(u>y)\big] f(P,c)$. This minimization problem can be reconciled to a standard calculus of variations problem, as follows. First, one can obtain an approximation by substituting $\mathbb{R}$ with a bounded

[*]Vrije Universiteit Amsterdam, The Netherlands, m.iacopini@vu.nl
[†]Free University of Bozen, Italy, francesco.ravazzolo@unibz.it
[‡]Queen Mary University, United Kingdom, l.rossini@qmul.ac.uk
[§]Corresponding author.

region $(a, b) \subset \mathbb{R}$. Second, the absolute value can be removed by a suitable choice of the constant k. Since the equality $ACPS(P, y; c) = k$ can be satisfied for any arbitrary choice of $k \in \mathbb{R}$, and $ACPS(P, y; c) < +\infty$, then there exists $k^* \in \mathbb{R}$ such that $ACPS(P, y; c) - k^* > 0$. Therefore, by choosing $k = k^*$ one gets

$$ACPS(P, y; c) = k^*, \quad \text{where} \quad |ACPS(P, y; c) - k^*| = ACPS(P, y; c) - k^* > 0.$$

Putting all together one gets the following minimization problem

$$\min_P \Big\{ ACPS(P, y; c) - k^* \Big\} \equiv \min_P \Big\{ \int_a^b L(u, P, P')\, \mathrm{d}u - k^* \Big\}. \tag{2}$$

Finally, since k^* is constant, the optimum is $P^* = k^* + P^{**}$, where P^{**} is the solution of the problem

$$\min_P \int_a^b L(u, P, P')\, \mathrm{d}u,$$

for which the corresponding Euler-Lagrange equation is

$$\frac{\partial L}{\partial P} = 0.$$

Therefore, solving the minimization problem (2) yields a class of cumulative distribution functions, P^*, which depends on two free parameters: (i) the asymmetry parameter of the ACPS, c, and (ii) the constant of integration, $\bar{c}$.

References

Iacopini, M., Ravazzolo, F., and Rossini, L. (2020). "Proper scoring rules for evaluating asymmetry in density forecasting." *CAMP BI Working Paper 06/2020*. 1392

Contributed Discussion

F. J. Rubio* and M. F. J. Steel†

We congratulate the authors on a very interesting and thought-provoking paper. We welcome the exploration of novel ideas for producing priors using formal rules. Our comments relate to: (i) the principle for constructing the proposed priors and the role of the parameterisation; (ii) the use of the proposed priors for hypothesis testing.

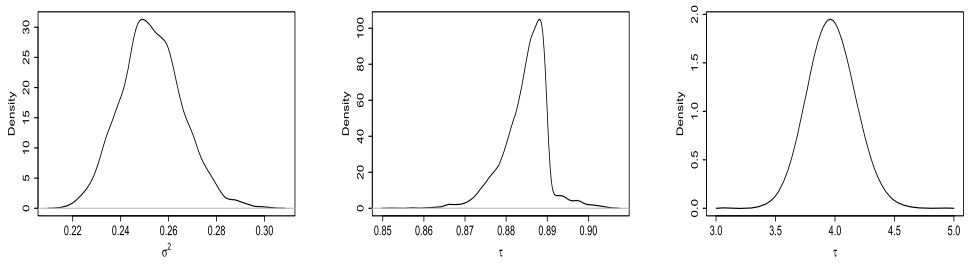
On the principle

In order to make the proposed principle operational, it is necessary to select the values w, c and $u(0)$. The authors provide some guidelines on how these values could be selected, but the precise mapping between the choice of these values and the properties of the resulting prior distribution still seems to be a partly open question.

The choice of these values is crucial for determining the properties of the prior. For instance, for the prior with support on $(0, \infty)$ presented in the paper, the solution for $u(\theta)$ grows very rapidly after a threshold that depends on the values of w, c and $u(0)$. This implies that the prior density decreases at a very fast rate after this point, seemingly super-exponentially fast, making it virtually zero beyond such a threshold. Thus, the prior is close to a truncated prior and then it is clear that things like (arbitrary) choices of scaling or parameterisation will have a substantial effect on the outcome. For example, consider a simple scenario where we wish to estimate the variance $\sigma^2 = 0.25$ in the model $X_i \sim N(0, \sigma^2)$, $i = 1, \ldots, 750$, using the proposed prior (with the choices recommended in the paper, see their Figure 3(a)). Figure 1(a) shows the posterior distribution obtained with a Markov Chain Monte Carlo (MCMC) sample of size 5,000. If we specify, say, a gamma$(0.01, 0.01)$ prior on σ^2 we get a similar result with the posterior concentrated around the true value. In contrast, suppose that now we are interested in estimating the precision $\tau = 1/\sigma^2 = 4$ using the same prior but now on τ. In order to implement the proposed prior for τ, we need to solve for $u(\tau)$ in a region where it grows at a fast rate, so we have discretised the prior on a grid of values of τ and obtained a numerical solution in Mathematica on that grid (R leads to numerical problems). Figure 1(b) shows the corresponding posterior, which is concentrated far from the true value of the parameter as the vanishingly thin tail of the prior dominates the posterior distribution. In contrast, Figure 1(c) shows the posterior distribution of τ obtained with a gamma$(0.01, 0.01)$ prior on τ, which is a gamma$(375.01, (1/2)\sum_{i=1}^{750} x_i^2 + 0.01)$, now concentrated around the true value of the parameter. In view of the sample size, any other prior with a tail that is not decreasing (much) faster than exponential would give a similar reasonable answer.

*Department of Mathematics, King's College London. London, WC2R 2LS, UK, javier.rubio_alvarez@kcl.ac.uk

†Department of Statistics, University of Warwick. Coventry, CV4 7AL, UK, m.steel@warwick.ac.uk



(a) (b) (c)

Figure 1: $X_i \sim N(0, \sigma^2)$. Posterior distribution of: (a) σ^2 using the proposed prior on σ^2; (b) $\tau = \sigma^{-2}$ using the proposed prior on τ; (c) τ using a gamma$(0.01, 0.01)$ prior on τ.

The aforementioned points seem to be effects induced by the addition of the Hyvärinen scoring rule, but the gain in terms of the frequentist and Bayesian performance remains unclear. The adoption of this particular rule induces the extremely thin tails which is, for many practical purposes, akin to a prior truncation and this flies in the face of the usual "folklore" that objective priors should rather err on the side of heavy tails to avoid excluding potentially important parts of the parameter space. Heavy-tailed priors tend to induce better frequentist performance of the posterior distribution, which is a desirable property for objective priors.

In our view, the interpretation of the parameters plays a crucial role in the specification of a prior distribution. Although the initial formulation of the proposed prior is not connected with the model, the need for specifying additional properties (such as shape) leaves a choice to the user. Specifying the additional ingredients in the proposed rule to produce priors for these parameters would require some analysis of their role, thus re-establishing the connection with the model at a more subjective level. The effect of the role of the parameters may be more noticeable for parameters taking values on the same parameter space but with completely different roles, such as a scale parameter and the degrees of freedom of the Student-t distribution.

Inference versus testing

For practically important situations such as mixture models or Bayesian variable selection it is typically difficult to use automatic principles. It often seems more useful to adopt a bespoke prior framework, which is not borne out of a principle, but reflects the question one wants to address. For example, non-local priors (Johnson and Rossell, 2010) aim to control the convergence rate of the Bayes factor. We would be interested to see how we could achieve such properties using the proposed principle (as suggested in Section 7), but we expect that the priors for testing presented in the paper could have a large impact on the results due to their local nature and extremely light tails.

References

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(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 1395

Contributed Discussion

Zhou Lan*

The authors proposed a novel objective prior from scoring rules. For spatial statisticians, a practical example which the priors from scoring rules can be applied to, is the prior specifications of the parameters of the Matern correlation function, under the Gaussian process model. The Matern correlation function is a commonly used correlation function capturing spatial dependence. It is expressed as

$$\mathcal{K}(d|\rho,\nu) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{d}{\rho}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{d}{\rho}\right),$$

where d is the Euclidean distance, and the function of d depends on two parameters: range parameter $\rho \in (0,\infty)$ and smooth parameter $\nu \in (0,\infty)$. K_ν is the modified Bessel function of the second kind and Γ is the gamma function. The current objective priors (Berger et al., 2001) i.e., Jeffery and reference priors may be complicated. The "weakly"-informative priors are actually implemented in many practical spatial papers. Because of its convenience, the most common one is to give uniform prior such that $\nu \sim \mathcal{U}(a_\nu, b_\nu)$ and $\rho \sim \mathcal{U}(a_\rho, b_\rho)$ (e.g., Reich et al., 2010), where $\mathcal{U}(a,b)$ denotes a uniform distribution ranging from a to b, or to use log-normal distribution such that $\log\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu^2)$ and $\log\rho \sim \mathcal{N}(\mu_\rho, \sigma_\rho^2)$ (e.g., Boehm et al., 2013), where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ denotes a normal distribution with mean (vector) $\boldsymbol{\mu}$ and variance (matrix) $\boldsymbol{S}$. Our discussion aims to use numerical studies to evaluate performances of the proposed prior, in comparison to the uniform prior and the log-normal prior.

Here is the specification of the Gaussian process model. Let $Y(\boldsymbol{s}) \in \mathbb{R}$ be the spatial response at location $\boldsymbol{s} \in \mathcal{D}$, where $\boldsymbol{s}$ is the coordinate of a location and $\mathcal{D}$ is a collection of locations. The model is specified as $[Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)]^T \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{X}_{n\times(p+1)}$ is the design matrix (an intercept and p covariates), $\boldsymbol{\beta}_{(p+1)\times 1}$ is the coefficients, and σ^2 is the marginal variance for each variable. $\boldsymbol{\Sigma}$ is the correlation matrix constructed by Matern such that the correlation between locations $\boldsymbol{s}$ and $\boldsymbol{s}'$ is $\mathcal{K}(||\boldsymbol{s}-\boldsymbol{s}'||\ |\rho,\nu)$. The Bayesian analysis of this model requires the prior specifications of the unknown parameters. We can assign conjugate priors to the coefficients $\boldsymbol{\beta}$ and the variance σ^2: $\boldsymbol{\beta}$ follows a normal distribution with mean $\mathbf{0}$ and covariance matrix $100\boldsymbol{I}$; σ^2 follows a gamma distribution with shape parameter 0.01 and rate parameter 0.01. The data are generated as follows. The data is generated on a 2-dimensional map ($\boldsymbol{s}_i = [x_i, y_i] \in \mathcal{D} \subseteq \mathbb{R}^2$). For each replication, we generated $n = 200$ data points. Their coordinates from uniform distributions ranging from 0 to 1, denoted as $[x_i, y_i] \sim \mathcal{U}(0,1) \times \mathcal{U}(0,1)$. Each data point has four covariates ($X_{ij},\ j = 1:4,\ i = 1:200$) and each was generated from $X_{ij} \sim \mathcal{N}(0,1)$. Thus, the design matrix is $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^T$ and $\boldsymbol{X}_i = [1, X_{i1}, X_{i2}, X_{i3}, X_{i4}]^T$. The coefficients vector is $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^T = [1, 10, 0, 1, 2]^T$. The variance is $\sigma^2 = 0.5$. We give

*Center for Outcomes Research and Evaluation, Yale School of Medicine, zhou.lan@yale.edu

$\{\rho = 0.1, \nu = 0.5\}$ to construct the spatial correlation matrix, where the MSE is averages over all replications and the coverages are the proportions that the 95% posterior covers the true value of a parameter.

We assign the proposed priors to the two unknown parameters (ρ, ν) with $c = 2$ and $u(\theta) = 10^{-5}$. Given Figure 7 of the paper, we give $u(0) = 10^{-5}$ because thus the prior $p(\theta)$ has a gentle decay toward 0. In comparison, the uniform prior is set as $\nu \sim \mathcal{U}(0, 1000)$ and $\rho \sim \mathcal{U}(0, 1000)$, and the log-normal prior is set as $\log \nu \sim \mathcal{N}(-1, 1)$ and $\log \rho \sim \mathcal{N}(0, 100)$. For each replication, 8000 Markov chain Monte Carlo (MCMC) samples are collected after 2000 burn-ins to calculate the mean square error (MSE) of the posterior means and the 95% posterior coverage. 80 simulation replications are generated, and the MCMC chains of the three methods are checked to pass Heidelberger and Welch's convergence diagnostic in each replication. The results are summarized in Table 1. From the results, we can find that the proposed prior performs the best.

Parameter	Proposed		Uniform		LogNormal	
	MSE	Coverage	MSE	Coverage	MSE	Coverage
ρ	0.50	0.95	0.77	0.96	0.86	0.96
ν	0.017	0.91	0.019	0.90	0.095	0.95

Table 1: Simulation results: the MSE is averages over all replications and the coverages are the proportions that the 95% posterior covers the true value of a parameter.

Although there are many concerns on the uniform prior, the uniform prior is still the most popular prior. The uniform prior is simple and "safe". Unlike other priors (e.g., LogNormal) that you have to "guess" the best support for this prior, the uniform prior can be simply set as a very large interval. This numerical study is a preliminary result that may initiate some changes in Bayesian spatial statistics. Besides the efforts of objective priors (Berger et al., 2001), people also put some efforts to construct good priors for spatial models (e.g., Fuglstad et al., 2019). A comprehensive review and some numerical comparisons are encouraged to be given. Furthermore, there are two practical issues which may prohibit the widely use of the proposed prior: (1) how to set up C and $u(0)$ is vague to some practitioners; (2) the computation of the prior is using Taylor's expansion is inconvenient to some practitioners, and a package or some codes are encouraged to be provided. Once the proposed prior can be implemented as easily as the uniform prior, there is no doubt that people will switch to the proposed objective prior.

References

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). "Objective Bayesian analysis of spatially correlated data." *Journal of the American Statistical Association*, 96(456): 1361–1374. MR1946582. doi: https://doi.org/10.1198/016214501753382282. 1397, 1398

Boehm, L., Reich, B. J., and Bandyopadhyay, D. (2013). "Bridging conditional and marginal inference for spatially referenced binary data." *Biometrics*, 69(2): 545–554. MR3071073. doi: https://doi.org/10.1111/biom.12027. 1397

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). "Constructing priors that penalize the complexity of Gaussian random fields." *Journal of the American Statistical Association*, 114(525): 445–452. MR3941267. doi: https://doi.org/10.1080/01621459.2017.1415907. 1398

Reich, B. J., Fuentes, M., Herring, A. H., and Evenson, K. R. (2010). "Bayesian variable selection for multivariate spatially varying coefficient regression." *Biometrics*, 66(3): 772–782. MR2758213. doi: https://doi.org/10.1111/j.1541-0420.2009.01333.x. 1397

Contributed Discussion

F. Llorente*, L. Martino†, and D. Delgado-Gómez‡

In (Leisen et al., 2018), the authors introduce a novel approach for building objective prior densities $p(\theta)$ based on proper score rules. Here we discuss some issues regarding the fact that $p(\theta)$ can only be evaluated up to a normalizing constant, and the application for model selection purposes. We also provide an additional comparison (and the corresponding code) with the intrinsic Bayes factor approach, extending the numerical experiment in Appendix E of the Supplementary Material, provided by (Leisen et al., 2018).

Introduction

We would like to congratulate the authors in (Leisen et al., 2018) for this contribution on devising objective prior densities, as we believe it will further encourage the use of the Bayesian formalism when there is little, or any, *a priori* knowledge of the parameter of interest. Only the knowledge of support domain Θ is required for the construction, hence they are "more" objective than other popular model-based approaches. To derive the objective priors, in the form of $p(\theta) \propto e^{-u(\theta)}$, the authors rely on proper score rules, which defines a second order differential equation for $p(\theta)$. Interestingly, the authors also show that the same differential equation can be obtained by minimizing the sum of entropy and Fisher information. The solution is not a unique prior density but a class of prior distributions, where some parameters can be set by the practitioner. For instance, the resulting priors can be forced to be proper. This is very useful in contexts where other objective priors are problematic, e.g., model selection via Bayes factors.

Analysis of the proposed approach

Although we believe the class of objective priors that result from this work are really useful, below we remark some critical points and suggest some possible solutions. The proposed prior densities $p(\theta) \propto e^{-u(\theta)}$ cannot be analytically obtained, but they are implicitly determined by an equation relating $u(\theta)$ and its derivative $u'(\theta)$. The authors use a Taylor expansion to evaluate $u(\theta+\epsilon)$ provided that $u(\theta)$ is known. Thus, the evaluation of $p \propto e^{-u}$ is only approximate (even if additional terms in the Taylor expansion can be included in order to increase the precision). Furthermore, the normalizing constants of the priors $p(\theta) \propto e^{-u(\theta)}$ are also unknown. These normalizing constants are not needed in the Markov chain Monte Carlo (MCMC) scheme given in Appendix A of the Supplementary Material, but it is required for the computation of Bayes factors. To solve this issue, one possibility is to modify this MCMC algorithm in order to sample from $p(\theta)$ and then use *reverse importance sampling* (RIS) (Llorente et al., 2020, Section 3). Namely,

*Department of Statistics, Universidad Carlos III de Madrid, Spain, felloren@est-econ.uc3m.es
†Dept. of Signal Processing, Universidad Rey Juan Carlos, Spain
‡Department of Statistics, Universidad Carlos III de Madrid, Spain

we first obtain a sample $\theta^{(1)}, \ldots, \theta^{(T)}$ from the prior $p(\theta)$ using a modification of the MCMC algorithm from Appendix A, where the acceptance probability is replaced by

$$\alpha = \min\left\{1, e^{-[u(\theta') - u(\theta^{(t)})]} \frac{q(\theta^{(t)}|\theta')}{q(\theta'|\theta^{(t)})}\right\}. \tag{1}$$

Then, using an auxiliary density $\varphi(\theta)$ defined on Θ, we compute

$$\widehat{c} = \left(\frac{1}{T}\sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{e^{-u(\theta^{(t)})}}\right)^{-1}, \tag{2}$$

which is a consistent estimator of $\int_\Theta e^{-u(\theta)} d\theta$. The auxiliary density $\varphi(\theta)$ should have lighter tails than the prior $p(\theta)$ to avoid the infinite variance problem (as often occurs with the harmonic mean estimator). See the theoretical example in Section 7.1 of (Llorente et al., 2020). Since the evaluation of $p(\theta)$ is approximate, note also that the resulting MCMC algorithms fall within the noisy MCMC framework (Alquier et al., 2016; Andrieu and Roberts, 2009).

Comparison with partial and intrinsic Bayes factors

Consider the example from Appendix E of the Supplementary Material. The goal is to compare two models by means of the Bayes factor, namely $M_1 = \{f_1(x|\theta) = \theta^x e^{-\theta}/x!,\ p_1(\theta)\}$ and $M_2 = \{f_2(x|\phi) = \phi(1-\phi)^x,\ p_2(\phi)\}$, when independent data $\mathbf{x} = (x_1, \ldots, x_n)$ are generated from each of them. The authors used a uniform prior $p_2(\phi) = 1,\ \phi \in (0,1)$, and their proposed prior for $p_1(\theta)$. Hence, the Bayes factor BF_{12} is well defined since $p_1(\theta)$ is proper. Note that the unknown normalizing constant of $p_1(\theta)$ needs to be approximated in order to compute BF_{12}. Tables 7 and 8 in the Supplementary Material give the results obtained for $n = 30$ and $n = 100$, showing the good performance in the detection of true model using the proposed objective prior $p_1(\theta)$.

Here, we aim to replicate Tables 7 and 8 using another uniform prior $p_{1,u}(\theta) \propto 1,\ \theta \in (0, \infty)$ for model M_1 instead, i.e. an improper prior. In this situation, the Bayes factor is not well-defined due to the arbitrary constant in $p_{1,u}(\theta)$. Hence, we need to resort to *partial* Bayes factors (O'Hagan, 1995, Sect. 2), where we compute the posterior of a single observation x_i (training set) under prior $p_{1,u}(\theta)$, i.e., $p_1(\theta|x_i) \propto f_1(x_i|\theta)p_{1,u}(\theta)$, and use $p_1(\theta|x_i)$ now as a proper prior in the computation of BF_{12}. In order to avoid the dependence on the training sample, we use the idea in (Berger and Pericchi, 1996), called *intrinsic* Bayes factors, that consists in averaging over all possible training samples, resulting in the following intrinsic Bayes factor

$$IBF_{12} = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_0^\infty f_1(\mathbf{x}_{-i}|\theta)p_1(\theta|x_i)d\theta}{\int_0^1 f_2(\mathbf{x}|\phi)d\phi} = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_0^\infty f_1(\mathbf{x}|\theta)d\theta / \int_0^\infty f_1(x_i|\theta)d\theta}{\int_0^1 f_2(\mathbf{x}|\phi)d\phi}. \tag{3}$$

Note that the cost of computing IBF_{12} increases with n. We compute IBF_{12} in 100 different runs for several true values of θ and ϕ, and we show the results in Table 1 and Table 2 for $n = 30$ and $n = 100$, respectively.[1] The results clearly show that the

[1]The Matlab code is available at http://www.lucamartino.altervista.org/Code_Llorente_CommentLeisen.m.

uniform prior on θ allows for correctly selecting M_1 when it is indeed the true model, with very few exceptions as also obtained in the paper. However, when M_2 is the true model, the use of such improper prior makes probable selecting M_1 as the most likely model, as proves the 43 exceptions obtained when $\phi = 0.8$ and $n = 100$, that is, almost half of the time we would wrongly select M_1 over M_2. This also shows the benefit of the objective priors proposed in (Leisen et al., 2018).

True model = M_1				True model = M_2			
θ	min	max	Exceptions	ϕ	min	max	Exceptions
5	6.28×10^{3}	3.95×10^{11}	0	0.5	5.45×10^{-9}	884.25	30
2	0.55	7.40×10^{6}	1	0.2	1.61×10^{-26}	9.76	2
				0.8	0.004	10.51	66

Table 1: Model comparison for $n = 30$. Minimum and maximum IBF_{12} under true model M_1 (Poisson) and M_2 (Geometric) for 100 simulations.

True model = M_1				True model = M_2			
θ	min	max	Exceptions	ϕ	min	max	Exceptions
5	2.38×10^{11}	4.52×10^{29}	0	0.5	1.98×10^{-13}	500.52	4
2	2.22×10^{3}	2.60×10^{14}	0	0.2	2.02×10^{-72}	3.34×10^{-18}	0
				0.8	0.003	6.69	43

Table 2: Model comparison for $n = 100$. Minimum and maximum IBF_{12} under true model M_1 (Poisson) and M_2 (Geometric) for 100 simulations.

References

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels." *Statistics and Computing*, 26(1-2): 29–47. MR3439357. doi: https://doi.org/10.1007/s11222-014-9521-x. 1401

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *The Annals of Statistics*, 37(2): 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 1401

Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91(433): 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 1401

Leisen, F., Villa, C., and Walker, S. G. (2018). "On a Class of Objective Priors from Scoring Rules." *Bayesian Analysis*. 1400, 1402

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2020). "Marginal likelihood computation for model selection and hypothesis testing: an extensive review." *arXiv preprint* arXiv:2005.08334. 1400, 1401

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 99–118. MR1325379. 1401

Contributed Discussion

Francisco Louzada[*], Eduardo Ramos[†], and Pedro L. Ramos[‡]

We congratulate the authors for their stimulating and excellent work on proposing a different procedure to obtain objective priors. The current methods usually depend on difficult metrics that may not be obtained when the model has a more complex structure. On the other hand, the Leisen, Villa and Walker (LVW) proposal does not depend on the selected model and can be constructed based solely on the parametric space.

In this discussion, we show how to compute the prior presented in the LVW paper on an specific scenario, where the parametric space of the prior is $(0,\infty)^k$, for $k \geq 1$. More specifically, we provide (i) the steps to reduce the computation for the k-dimensional initial value problem to the unidimensional case, under certain constraints on the used constants, (ii) a method for obtaining a closed formula for approximating the solution for the unidimensional case, and (iii) an illustration of our procedure to the Gamma distribution.

Following the LVW paper, we analyse the prior $p : (0,\infty)^k \subset \mathbb{R}^k \to \mathbb{R}$ given by $p = e^{-u}$, where $u : (0,\infty)^k \subset \mathbb{R}^k \to \mathbb{R}$ is the solution, for $j = 1, \cdots, k$, of

$$\frac{\partial u}{\partial \theta_j} = s_j \sqrt{ce^u - 2w(1+u)/k},\ u(0,\cdots,0) = u_0 > 0, \tag{1}$$

where $c \in \mathbb{R}$, $w \in \mathbb{R}$ and $s_j \in \{-1, 1\}$ for $1 \leq j \leq k$. We consider only the cases where $s_j = 1$ for all j, since $s_j = -1$ implies in u being decreasing in θ_j which, together with $p = e^{-u}$, implies in p being improper. Moreover we suppose $c \geq 2w/k > 0$ in order to have $ce^u - 2w(1+u)/k > 0$, for all $u > 0$.

Firstly, we prove that the solution of (1), for dimension $k \geq 1$, can be reduced to the solution for the unidimensional case with any initial condition v_0, as long as $0 < v_0 \leq u_0$. Indeed if $v : (0,\infty) \to \mathbb{R}$ is the solution of

$$v'(\theta) = \sqrt{ce^{v(\theta)} - 2(w^*)(1+v(\theta))},\ v(0) = v_0 > 0, \tag{2}$$

where $w^* = w/k$ and $0 < v_0 \leq u_0$, since we are supposing $c \geq 2(w^*) > 0$, it follows that $ce^v - 2(w^*)(1+v) \geq cv^2/2 \geq cv_0^2/2$ for all $v > 0$. Thus, $v'(\theta) \geq \sqrt{cv_0^2/2} > 0$ for all θ, which implies that v is strictly increasing with $\lim_{\theta\to\infty} v(\theta) = \infty$. Therefore, given $u_0 \geq v_0$, there exists a unique $\theta(u_0)$ such that $v(\theta(u_0)) = u_0$, and thus a direct computation shows that $u(\theta_1, \cdots, \theta_k) = v(\theta_1 + \cdots + \theta_k + \theta(u_0))$ is the solution of the

[*]Institute of Mathematical Science and Computing, University of São Paulo, São Carlos, SP, Brazil, louzada@icmc.usp.br

[†]Institute of Mathematical Science and Computing, University of São Paulo, São Carlos, SP, Brazil, eduardoramos@usp.br

[‡]Institute of Mathematical Science and Computing, University of São Paulo, São Carlos, SP, Brazil, pedrolramos@usp.br

equation (1). In practice, since (2) is an ordinary differential equation, the value $\theta(u_0)$, such that $v(\theta(u_0)) = u_0$, can be computed by numerically solving (2).

To construct an explicit formula for $v(\theta)$ for $c \geq 2w^*$ and $v_0 > 0$, since $v(\theta)$ has a fast growing rate, we propose approximating $\log(v(\theta))$ via a polynomial by the following method: using the least squares method, we find $a = (a_1, \cdots, a_m)$ minimizing the function $G : \mathbb{R}^m \to \mathbb{R}$ given by

$$G(a) = \sum_{i=1}^{h} (g_a(\theta_{\{i\}}) - \log(v(\theta_{\{i\}})))^2 \text{ for all } a \in \mathbb{R}^m,$$

where $g_a(\theta) = a_1 + a_2\theta + \cdots + a_m\theta^{m-1}$ and $\{\theta_{\{1\}}, \cdots, \theta_{\{h\}}\}$, is a large set of values in $[0, \theta_{\max}]$, for a given $\theta_{\max} > 0$. Therefore, due to the least squares formula, a is obtained by solving the linear equation $(A^TA)a^T = A^Tb^T$, with $A = (\theta_{\{i\}}^j) \in M_{h\times m}(\mathbb{R})$ and $b = (\log(v(\theta_{\{1\}})), \cdots, \log(v(\theta_{\{h\}}))) \in \mathbb{R}^h$, where the values $v(\theta_{\{j\}})$ are computed by solving numerically the differential equation (2).

Thus, if h and m are large enough, $g_a(\theta)$ should provide a good approximation for $\log(v(\theta))$ in $[0, \theta_\infty]$, and from $p = \exp(-\exp(\log(v)))$, we conclude that $q(\theta) = \exp(-\exp(g_a(\theta)))$ should provide a good approximation for $p(\theta)$ in $[0, \theta_\infty]$. Moreover, as long as the dominating coefficient of $g_a(\theta)$ is positive, both $q(\theta)$ and $p(\theta)$ should decay fast to 0 as $\theta \to \infty$, and thus we expect the approximation to be accurate in $[0, \infty)$ as well.

For instance for $c = 2$, $w^* = 1$ and $u_0 = 0.01$, we consider the values $\theta_{\{j\}} = 0.01 \cdot j$ for $0 \leq j \leq 600$, $\theta_\infty = 6$, $m = 5$ and, following the above method, we obtain

$$q(\theta) = \exp(-\exp(-4.5461 + 0.7322\,\theta + 0.2734\,\theta^2 - 0.0935\,\theta^3 + 0.0105\,\theta^4)), \quad (3)$$

as an approximation for $p(\theta) = \exp(-v(\theta))$, where v satisfies (2) with $v_0 = u_0 = 0.01$ (see Figure 1). Moreover, numerically, we can obtain $v(2.28) \approx 0.1$ and $v(4.63) \approx 1.31$ and thus from the above discussion (2) $q(\theta + 2.28)$ provides a good approximation for

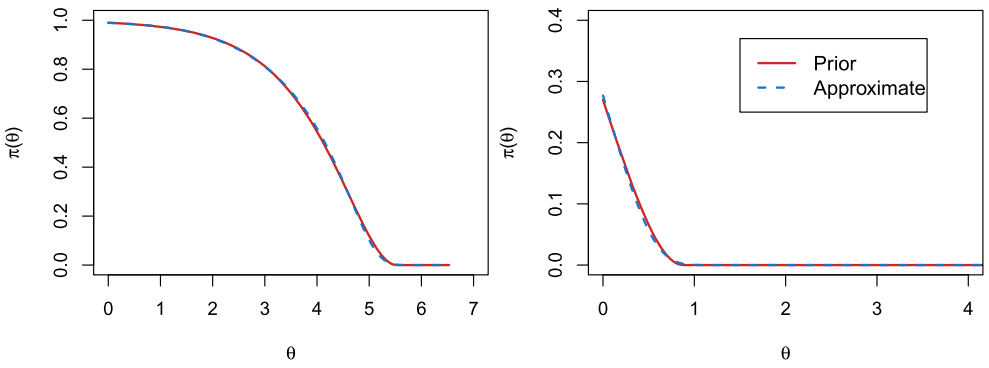


Figure 1: Priors $p(\theta)$ and our proposed approximation $q(\theta)$ for $u_0 = 0.01$ (left panel) and $u_0 = 1.31$ (right panel).

the prior p under the condition $u_0 = 0.1$ and $q(\theta + 4.63)$ provides a good approximation for the prior p under the condition $u_0 = 1.31$ (see Figure 1).

Now we consider an application in the Gamma distribution with shape and scale parameters. From (3), an approximation of the proposed prior (without the normalized constant) for $u_0 = 0.01$ can be written as

$$\pi(\theta_1 + \theta_2) \propto q(\theta_1 + \theta_2) = \exp(-\exp(-4.5461 + 0.7322(\theta_1 + \theta_2) + 0.2734(\theta_1 + \theta_2)^2 - 0.0935(\theta_1 + \theta_2)^3 + 0.0105(\theta_1 + \theta_2)^4)). \tag{4}$$

Likewise, from the above discussion, we can obtain the priors assuming other values in $u_0 = 1.31$, e.g., we have $\pi(\theta_1, \theta_2) \propto q(\theta_1 + \theta_2 + 2.28)$ for $u_0 = 0.1$ and $\pi(\theta_1, \theta_2) \propto q(\theta_1 + \theta_2 + 4.63)$ for $u_0 = 1.31$.

In Figure 2, we compared the proposed prior (4) to the Jeffreys prior when the model follows a Gamma distribution. The Metropolis-Hastings algorithm was constructed for both models using the same approach. The comparison was made for $n = 10, 15, \ldots, 120$, assuming 5,000 different random samples for each n. The bias is presented to compare the obtained posterior means. Although the aim is not to select a specific prior, we can observe that both return satisfactory results in terms of bias, decreasing as the sample samples increase.

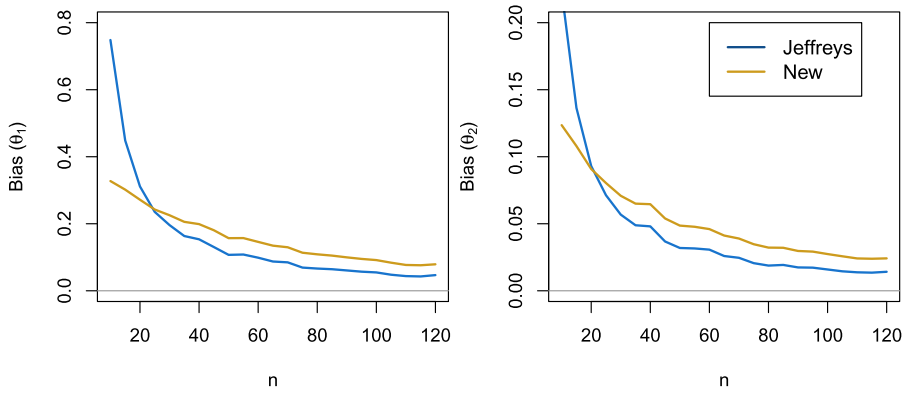


Figure 2: Average Bias for the estimates of the Gamma distributions with $\theta_1 = 2$ and $\theta_2 = 0.5$ under different samples sizes.

Overall, we observe that the LVW approach to construct new objective priors can be useful when the standard methods cannot be applied due to the model's complexity. This note proposed a useful polynomial approximation that can be easily implemented on the standard software/packages such as OPENBUGS, JAGS, and Stan. The results for our simulated scenarios were satisfactory when compared with Jeffreys prior. We observed that the value of u_0 impacts the posterior estimates, and sensitivity analysis plays an essential role during the prior selection. As a possible extension of this note, it may be considered the same approach for other parametric spaces.

Contributed Discussion

Rob Trangucci*, Derek Hansen†, and Yang Chen‡

The authors propose a method for using proper local scoring rules of order two to generate novel objective priors that, depending on the modeler's choices, can be constrained to be proper distributions. We wholeheartedly agree that the method's ability to generate proper priors is one of its main strengths. In addition to the scenarios explored by the authors, a proper objective prior is useful as a baseline prior for methods that measure the information content of informative priors applied to model classes for which objective priors do not exist or posterior propriety is not guaranteed Reimherr et al. (2020); Jones et al. (2020).

Figure 7 and Tables 1 and 2 illustrate the flexibility of the prior within the broad class of proper priors over $\Theta = (0, \infty)$. While the focus on the parameter space alone does, in some sense, indicate the prior is more "objective" compared to other objective prior procedures, the parameters ultimately do interact with the likelihood function (Gelman et al., 2017). Tables 1 and 2 also indicate that for many values of c and $u(0)$ the priors can be quite informative depending on the likelihood.

Jones et al. (2020) gives a measurement of just how much information a prior distribution adds to a posterior distribution for a given dataset, namely, the observed prior effective sample size (OPESS). OPESS measures how many incremental observations the prior represents in a posterior when compared to a separate posterior employing a noninformative prior.

We can use the method from Jones et al. (2020) to compare the OPESS for two objective priors over $\Theta = (0, \infty)$: prior π_1 with $c = 2, u(0) = 0.1$, and prior π_2 with $c = 2, u(0) = 0.65$. The strength of the method is that it is a proper Bayesian posterior estimand, so one can use the distribution over OPESS just as one would the posterior over the parameters. We will use it to compute the probability that the OPESS under π_2 is greater than the OPESS under π_1 for a given dataset, and the probability that the OPESS for π_2 is greater than the number of observed data points.

In keeping with Appendix B, suppose we observe 20 samples with a mean of 1.35 from a Poisson(θ) distribution. That the objective prior with $u(0) = 0.65$ is informative is readily apparent from a comparison of posteriors in the first column of Figure 1; the right tail of the posterior does not extend past 1.5, while the right tail of the posterior for π_1 extends past 2.5. The probability that OPESS under π_2 is greater than the OPESS under π_1 is 1, and the probability that OPESS is greater than 20 for π_2 is 0.96. The distribution of OPESS is shown in the second column of Figure 1. If we take the Kass and Wasserman view of objectivity (Kass and Wasserman, 1996), as the authors seem

*Department of Statistics, University of Michigan, Ann Arbor, MI, trangucc@umich.edu
†Department of Statistics, University of Michigan, Ann Arbor, MI
‡Department of Statistics and Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI

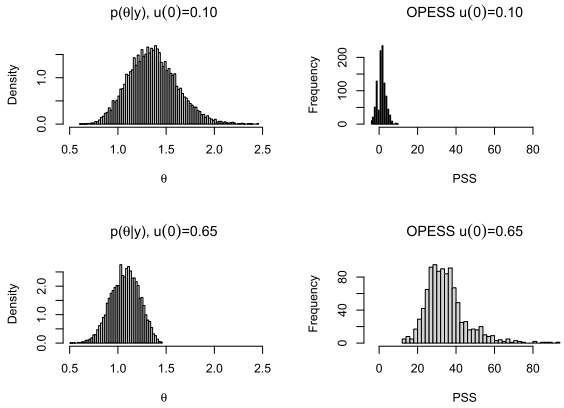


Figure 1: Comparison of posteriors and observed prior effective sample size under objective priors with $c = 2$ and $u(0) = 0.1$ in the top row and $u(0) = 0.65$ in the bottom row. The draws from the posterior are generated with CmdStanR (Gabry and Češnovar, 2020; Carpenter et al., 2017). The code that reproduces the figure is at `https://github.com/rtrangucci/pss-objective-prior`.

to, that an objective prior is a prior generated via a set of rules, how should the modeler choose a value for $u(0)$?

It is not immediately clear how a prior that yields a constant proper local scoring rule of order two should be interpreted. The connection between the solution to equation 4 and the p which minimizes $I_E(p) + \frac{1}{2} I_F(p)$ is interesting, and could provide more intuition about the properties of these priors. Our interpretation is that minimizing the functional:

$$\int_\Theta \left(\log p(\theta) + \frac{1}{2} \left(\frac{\partial \log p(\theta)}{\partial \theta} \right)^2 \right) p(\theta) d\theta$$

is akin to maximizing the entropy of $p(\theta)$ while penalizing the square of the L_2 norm of the score. When viewed through this lens, we can make the connection to smoothing spline regression, where a loss function is minimized with respect to an unknown function of covariates while penalizing the roughness of the function by penalizing its derivatives (Keener, 2011; Friedman et al., 2001). The solution to the above minimization problem makes sense in terms of generating an objective prior: We maximize the entropy of the prior but favor smoother distributions. We look forward to more research into this class of objective priors.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "*Stan*: A Probabilistic Programming Language." *Journal of Statistical Software*, 76. 1407

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics. Springer, New York. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 1407

Gabry, J. and Češnovar, R. (2020). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org. 1407

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy*, 19: 555. 1406

Jones, D. E., Trangucci, R. N., and Chen, Y. (2020). "Quantifying Observed Prior Impact." *arXiv:2001.10664*. 1406

Kass, R. E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, 91: 1343–1370. 1406

Keener, R. W. (2011). *Theoretical statistics: Topics for a core course*. Springer. MR2683126. doi: https://doi.org/10.1007/978-0-387-93839-4. 1407

Reimherr, M., Meng, X.-L., and Nicolae, D. L. (2020). "Prior Sample Size Extensions for Assessing Prior Informativeness and Prior–Likelihood Discordance." *arXiv:1406.5958*. 1406

Contributed Discussion

Brunero Liseo*

Let me first congratulate with the Authors for this highly thought-provoking paper.

The selection of a really objective prior for a statistical model is an impossible task, since any (especially proper) prior would bring some sort of information. Today, there is a broad agreement that an objective prior should be taken as a "reference" or "benchmark" point, and evidence from the data should be measured – in some way – in terms of deviation from this reference point.

The main novelty in the paper, in my opinion, is represented by the fact that the search of the "objective" prior is disconnected by the statistical model and it is only based on the parameter space. This choice is supported, in the Introduction of the paper, in terms of "objectivity", since the specific choice of a model is considered as subjective. The obvious consequence of this choice is the loss of invariance, which seems to me the real point to discuss.

According to the de Finetti's representation theorem, at least in an exchangeable setting, the parameters in the model are merely "Greek letters" (Lindley, 1990) which allow us to represent, in a convenient way, our probability statements on the observables. In the (parametric!) statistical practice, one selects a model and the mutual "distances" among the probability distributions which are members of that model are rarely represented by the Euclidean distance. In a location model, this can be safely assumed, in more complex situation, it cannot. Here I consider one of my favourite examples, which often operates as a counterexample, namely the simple standard skew-normal model, that is

$$p(x|\lambda) = 2\varphi(x)\Phi(\lambda x), \qquad x \in \mathbb{R}; \lambda \in \mathbb{R},$$

with $\varphi(\cdot)$ and $\Phi(\cdot)$ being the pdf and the cdf of a standard Normal distribution, respectively. For large values of λ, $p(x|\lambda)$ is almost undistinguishable from $p(x|\lambda+\theta)$ in the sense that the ratio of the two densities, i.e. $\Phi(\lambda x)/\Phi((\lambda+\theta)x)$, is practically constant for any value of x. The use of the same prior for such a shape parameter and for a location parameter might be highly misleading. Also, the skew-normal model has another "natural" parameter space, the set $[-1, 1]$, which is obtained from the transformation

$$\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$$

None of the parameters, either λ or δ, can be considered natural here: what is the objective prior suggested by the Authors in the two cases? Do they produce the same inference, in a predictive sense?

From a more general perspective regarding invariance, my impression with the proposed method is that the price of buying objectivity "and" properness is not negligible, since:

*Sapienza Università di Roma, brunero.liseo@uniroma1.it

- you have to select a parametrization: but this is exactly what the Jeffreys' method wants to avoid in order to find a sort of conventional prior.
- the method finds a class of priors, and additional subjective choices for c and $u(\cdot)$ are often necessary.

Another issue, which might deserve attention, is that a Bayesian analysis with reference or Jeffreys' improper priors is often the best way to understand and use classical results. If you judge priors from their frequentist behaviour, then it is hard to beat them, at last in regular models.

Finally, the Authors say in the Introduction that "*models are by and large misspecified and consequently model based priors are propagating this misspecification*". In this regard, I like to take a nonparametric perspective: when we choose a parametric model, we confine ourselves on a negligible subset of all possible cdf's on the observables. The search of an objective prior is actually the search of a conditional (on the model!) objective prior: that makes, in my opinion, the link between model and prior indissoluble.

References

Lindley, D. V. (1990). The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics, *Statistical Science*, Volume 5, Number 1 (1990), 44–65. MR1054857. doi: https://doi.org/10.1214/ss/1177012253. 1409

Contributed Discussion

Ian H. Jermyn[*] and Karthik Bharath[†]

The authors propose a nice procedure for prior construction using maximum entropy under a constraint on the local 'non-uniformity' of the density (and potentially further constraints such as concavity or monotonicity). Use of the resulting priors in mixture models, model selection criteria, and their (possible) propriety, make them attractive.

We are concerned, however, by the lack of 'invariance' (more accurately, equivariance under diffeomorphisms of parameter space) of the approach. This condition is essential for a well-defined method, and hence for one that claims to be 'objective'. One can see this from several points of view.

1. Mathematically, the lack of equivariance implies that the method is not well-defined as it stands, since different parameterizations will lead to different probability measures on the parameter space, and therefore different posteriors and inferences. Only if a procedure for selecting a distinguished parameterization (or, more precisely, as discussed below, an underlying measure) is specified does the method become well-defined. Such a procedure is not specified in the paper.

2. Example: If the prior density only depends on the parameter space, then the prior density on $\mathbb{R}_+ \times \mathbb{R}_+$ will have the same functional form whether we parameterize a Gamma likelihood with 'shape' and 'scale' or 'shape' and 'rate'. This one functional form corresponds to two different prior probability measures, leading to different posteriors. Which should we use? In a similar but physically motivated example: if two researchers choose to parameterize a model using temperature T or inverse temperature $\beta = 1/T$, both common choices, the parameter space will be $\mathbb{R}_+$ in both cases leading to densities of the same functional form, and therefore different prior probability measures.

3. The lack of equivariance is equivalent to the lack of a well-defined underlying measure against which to define a density. It is well known (and obvious) that the expression for the entropy, I_E, is not well-defined unless the logarithm contains a ratio of p to another reference density m, both of which are defined with respect to an (arbitrary) measure $\mathrm{d}x$, meaning that $m(x)\,\mathrm{d}x$ is the underlying measure with respect to which the density is defined. Similarly, I_F is not well-defined unless p is replaced by p/m.

4. The paper attempts to avoid this issue by invoking Lebesgue measure in Definition 1. We note, however, that Lebesgue measure is itself *not* a well-defined quantity until the additive algebraic (as opposed to topological) structure of $\mathbb{R}$ (or, alternatively, the action of the translation group on $\mathbb{R}$) is defined. This structure is rarely present *a priori*, yet without it, the underlying measure is arbitrary.

[*]Department of Mathematical Sciences, Durham University, UK
[†]School of Mathematical Sciences, University of Nottingham, UK

5. Finally, in a *reductio ad absurdum*, we note that lack of equivariance is the only argument against using the uniform density in a particular parameterization to represent ignorance. If equivariance is no longer a requirement, then this choice is once again on the table. Indeed, the method predicts this: it is easy to see that the *global* minimizer of $I(p)$ (minimizing over boundary conditions as well as densities) is simply the uniform density.

The proposed method, maximum entropy under a constraint on the local 'non-uniformity' of the density, together with further constraints, in the form of boundary conditions or otherwise, might be described as 'objective', in Jaynes' sense (when discussing maximum entropy more generally) of depending on objectively-defined constraints only, were it not for the arbitrariness introduced by the lack of a well-defined underlying measure (equivalently, lack of equivariance to diffeomorphisms of the parameter space). This makes the proposed method dependent on arbitrary choices, which is the opposite of 'objective'. However, procedures for defining such measures do already exist: in particular, group invariance and the use of the likelihood, as in Jeffreys' prior, provide solutions.

The use of the likelihood for this purpose should not be scorned. If we know nothing about a parameter *a priori*, from whence does its connection to reality, its meaning, arise? This can only come from the likelihood connecting the parameter to current data; this is all that remains to define, for example, the difference between temperature and inverse temperature. In this situation, it is not only unsurprising, it is inevitable, that any prior will depend on the model.

The introduction of such a model-dependent measure as the underlying measure for the definition of the entropy and non-uniformity terms in the proposed method would result in a well-defined method with great utility.

Contributed Discussion

D. Stephen Coad* and Hugo Maruri-Aguilar*

Abstract. This discussion contains some comments on the paper by Professors Leisen, Villa and Walker. These concern the class of objective priors known as probability matching priors.

Keywords: differential equation, Jeffreys' prior, probability matching prior.

We commend the authors on their proposal to generate new types of objective prior distributions, tailored according to specific inference needs. It is briefly mentioned in Section 1 that one important class of objective priors are the probability matching priors. The connection between these and those that are proposed in the present work is not entirely clear.

In Datta et al. (2000) and Sweeting (2008), Jeffreys' prior is obtained by solving a differential equation. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with common density $f(x;\theta)$, where $\theta = (\theta_1, \ldots, \theta_k)'$. Further, let $I_F^{ij}(\theta)$ denote element (i,j) of the inverse of the Fisher information matrix per observation. Now define

$$\mu_j(\theta,\alpha) = \int_{q(\theta,\alpha)}^{\infty} \frac{\partial}{\partial\theta_j} f(x;\theta)dx,$$

where $q(\theta,\alpha)$ satisfies

$$\int_{q(\theta,\alpha)}^{\infty} f(x;\theta)dx = \alpha$$

for $0 < \alpha < 1$. Then, using the summation convention, Theorem 1 in Datta et al. (2000) states that, if a prior $\pi(\theta)$ satisfying

$$\frac{\partial}{\partial\theta_i}\{I_F^{ij}(\theta)\mu_j(\theta,\alpha)\pi(\theta)\} = 0$$

exists for every α, then it must be Jeffreys' prior. Such a prior $\pi(\theta)$ is termed a *level-α predictive probability matching prior*.

It would be interesting to know whether an objective prior $p(\theta)$ that satisfies (2) can be seen as also satisfying the above differential equation. If this is the case, then this would provide an important link between the proposed class of objective priors and Jeffreys' prior. Note that it was shown in Datta et al. (2000) that the latter only satisfies this differential equation in the case $k = 1$. Hence, an objective multiparameter prior $p(\theta) \propto \exp\{-u(\theta)\}$ satisfying (4) would not, in general, correspond to Jeffreys' prior.

*School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK, d.s.coad@qmul.ac.uk; h.maruri-aguilar@qmul.ac.uk

References

Datta, G. S., Mukherjee, R., Ghosh, M., and Sweeting, T. J. (2000). "Bayesian prediction with approximate frequentist validity." *The Annals of Statistics*, 28(5): 1114–1126. MR1805790. doi: https://doi.org/10.1214/aos/1015957400. 1413

Sweeting, T. J. (2008). "On predictive probability matching priors." In Clarke, B. and Ghosal, S. (eds.), *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanka K. Ghosh*, 3: 46–59. Beachwood, OH: Institute of Mathematical Statistics. MR2459215. doi: https://doi.org/10.1214/074921708000000048. 1413

Contributed Discussion

Theodore Kypraios*

I would like to thank the authors for a thought-provoking paper on constructing objective priors. I would like to highlight the following two (unrelated) points:

1. **[Sampling]** The authors discuss how one can draw samples from the posterior distribution of the parameter θ given the observed data using Markov Chain Monte Carlo methods. For example, in the supplementary material they employ a Metropolis-Hastings algorithm which requires evaluating the prior density at both the current and the candidate value of the chain which is done numerically.

   What if one is interested in sampling from the posterior using a (naive) Approximate Bayesian Computation algorithm or perhaps a Sequential Monte Carlo algorithm both of which will require to sample from the density of the proposed objective prior? Can the authors discuss how one can sample from such prior distributions?

2. **[Epidemic Modelling]** In a standard (homogeneously mixing) Susceptible-Infective-Removed (SIR) stochastic epidemic model during its infectious period, an infective individual has infectious contacts with each susceptible individual at times given by the points of a Poisson process of rate β/N with the further assumption that these Poisson processes are mutually independent, and where denotes N denotes the size of the population.

   The basic reproduction number, R_0 is a quantity of tremendous importance not only within the epidemic modelling community but also for policy makers as the ongoing COVID-19 outbreak has profoundly demonstrated. In the standard SIR model $R_0 = \beta E[T_I]$, where T_I denotes the infectious period distribution.

   A common choice for the infectious period is the *Exponential* distribution which is either parameterised in terms of its rate, say γ, or its mean $\frac{1}{\gamma}$.

   (a) Due to the lack of the invariance property when constructing the proposed objective priors wouldn't be the case that the implied prior distribution for R_0 will be different depending on which reparameterisation is used?

   (b) A typical choice in the literature for uninformative priors for both the infection (β) and removal (γ) rate is to use an Exponential distribution with very low rates (e.g. 10^{-6}). It has been shown (Clancy and O'Neill, 2008) that such a choice leads to an implied prior distribution of R_0 with prior mean $E[R_0] > 1$ which is not ideal since often researchers will report the posterior of R_0 which depending on how much information is available in the data may be influenced by the prior.

*School of Mathematical Sciences, University of Nottingham, United Kingdom, theodore.kypraios@nottingham.ac.uk

Can the authors elaborate on how their approach can be implemented in this case? R_0 is parameter that is positive but at the same time depending on the model that is used, it will be defined accordingly. For example, if $T_I \sim Gamma(u, v)$ with $E[T_I] = \frac{u}{v}$ then $R_0 = \frac{\beta u}{v}$. In other words, will the ratio of two objective prior distributions also lead to an objective distribution?

References

Clancy, D. and O'Neill, P. D. (2008). "Bayesian estimation of the basic reproduction number in stochastic epidemic models." *Bayesian Analysis*, 3(4): 737–757. MR2469798. doi: https://doi.org/10.1214/08-BA328. 1415

Rejoinder

Fabrizio Leisen[*], Cristiano Villa[†], and Stephen G. Walker[‡]

We are very grateful to the many discussions we have received which contain insightful comments and stimulating ideas. Objective Bayes is a cornerstone of the Bayesian framework; numerous problems and scenarios require the specification of default priors. While there are established approaches that have successfully contributed to the development of this research area, the growing complexity of real problems requires serious reconsideration as to how default priors can be chosen. We advocate a switch in focus in order to achieve practical implementations of the Bayesian framework. In fact, the methodology we propose in the paper challenges the *status quo* by re-defining the notion of objectivity. Clearly, this new view, as the many contributions we have received indicate, opens stimulating debates on the philosophical and practical aspects of the proposed class of priors. This paper, has therefore to be considered as the beginning of a novel research stream within objective Bayes. This is supported by the fact that we have received several suggestions of its implementation and extensions.

1 Invariance

A number of discussants mention invariance and comment on the lack of it, including **Consonni and García-Donato**, **Liseo**, **Jermyn and Bharath**. The reason is due to the lack of invariance of traditional information criterion. For example, for the Shannon information $I(p) = \int p \log p$, it is quite possible for $I(p_1) > I(p_2)$ and yet following a transformation $p \to \widetilde{p}$, the reverse inequality arises; i.e. $I(\widetilde{p}_1) < I(\widetilde{p}_2)$. Nevertheless, the Shannon information is used *a lot* despite this negative property.

Our argument in the paper is that the priors under consideration are all adequate for what it is they are chosen to do. Even within a single parameterization, the notion of a chosen prior is more about adequacy rather than any notion of uniqueness.

Nevertheless, to question a Bayesian procedure for lack of invariance is of fundamental interest and could arise even in simple and well used routines, such as a lack of invariance for Bayes estimates. Let us consider a very specific example.

Suppose $\pi(\theta) = e^{-\theta}$ is the prior for parameter θ for the model $f(x \mid \theta) = \theta\, e^{-x\theta}$. The posterior based on a sample $x_{1:n}$ is

$$\pi(\theta \mid x_{1:n}) = \text{Gamma}\,(1 + n, 1 + n\bar{x})\,.$$

[*]School of Mathematical Sciences, University of Nottingham, UK, Fabrizio.Leisen@nottingham.ac.uk

[†]School of Mathematics, Statistics and Physics, Newcastle University, UK, cristiano.villa@newcastle.ac.uk

[‡]Department of Mathematics, University of Texas at Austin, USA, s.g.walker@math.utexas.edu

The Bayes point estimate, with respect to square loss, is

$$\widehat{\theta} = \frac{1+n}{1+n\bar{x}}.$$

An alternative parameterization is $\phi = 1/\theta$. The prior for ϕ, based on the transform of densities, is $\pi(\phi) = \phi^{-2}e^{-1/\phi}$ which is IG(1, 1); i.e. inverse gamma. The posterior is

$$\pi(\phi \mid x_{1:n}) = \mathrm{IG}\,(1+n, 1+n\bar{x})$$

and the Bayes estimate now with square error loss is

$$\widehat{\phi} = \frac{1+n\bar{x}}{n}.$$

So, depending on the parameterization, we have two different answers. Since, note that $\widehat{\theta} \neq 1/\widehat{\phi}$.

Note that the above issue does not disappear even when a prior with invariance properties, such as Jeffreys, is employed. In fact, for the exponential distribution, the Jeffreys prior is $\pi^J(\theta) \propto \theta^{-1}$. If the parameterization is $\phi = \theta^{-1}$, then Jeffreys is $\pi^J(\phi) \propto \phi^{-1}$. In the first case we have

$$\pi^J(\theta|x_{1:n}) = \mathrm{Gamma}(n, n\bar{x}),$$

yielding the estimator $\hat{\theta} = 1/\bar{x}$. For the second parameterization, we have

$$\pi^J(\phi|x_{1:n}) = \mathrm{IG}(n, n\bar{x}),$$

with estimate $\hat{\phi} = n\bar{x}/(n-1)$. Therefore, even in the case of a noninformative prior, $\widehat{\theta} \neq 1/\widehat{\phi}$. It is quite clear that despite the two answers, there is really no concern from a practical perspective. Both answers are good and satisfactory. Hence, to criticize a Bayesian procedure for lack of invariance and to "prohibit" any such procedure, would need at first to take a long hard look at the Bayes estimates and potentially a much wider class of Bayesian practices.

2 Prior based on the parameter space only

A key idea in our approach is to derive the class of priors on the basis of the knowledge of the parameter space only. This aspect has been discussed by **Consonni and García-Donato**, **Liseo**, **Jermyn and Bharath** and **Rubio and Steel**.

To clarify on this aspect, we argue that the Bayesian prior is given by the probability measure Π on a suitable space of density functions and constructed from statistical model $f(\cdot|\theta)$, $\theta \in \Theta$, and probability density $p(\theta)$ on Θ, via

$$\Pi(f \in A) = \int_{\{\theta:\, f(\cdot|\theta)\in A\}} p(\theta)\, d\theta,$$

for all measurable sets A. Here, the random density function f is generated by taking $\theta \sim p$ and then setting $f(\cdot) = f(\cdot|\theta)$.

This is a more direct interpretation of the prior and avoids the inconvenient separation between Bayesian parametric and nonparametric methods. Indeed, it is useful to visualize Bayesian inference as the generation of a random density and associated functions; even if one is only considering random normal density functions. From this perspective, the two components that are traditionally known as the likelihood, $f(\cdot|\theta)$, and the prior, $p(\theta)$, are mere tools used for constructing Π and, perhaps, the reasons for this nomenclature are more historical than accurate.

Hence, the prior Π can never be assigned completely on objective grounds, as the function $f(\cdot|\theta)$ is the data model. However, the function $p(\theta)$ can be derived through some objective method. This said, it is puzzling to understand why $p(\theta)$ and $f(\cdot|\theta)$ need to feed off each other in any way; other than the θ in both needs to be sitting in the same parameter space Θ.

Consequently, in the paper we investigated the possibility of defining objective prior distributions that are not model dependent and based on the sole knowledge of the parameter space Θ. The $p(\theta)$ using only Θ loses the connection with the data model part of Π and hence could be argued as a consequence to be *more objective*. In fact, data models are by and large misspecified and, consequently, model based priors are propagating this misspecification. So, while a model based prior reinforces the connection between the misspecified model and the prior itself, a prior that depends on the parameter space loses only the connection.

We would like to stress the fact that we are not claiming that a prior can be structured without having any connection with the reality and, therefore, with the aim of the statistical analysis. A decision-maker (or experimenter) would naturally have knowledge of the problem she wants to tackle and, of course, this will be reflected in the whole modelling, including the choice of features the prior should have. Our point is that the method that constructs the prior does not involve the functional form of the model (i.e. the likelihood function).

3 Choice of c and $u(0)$

A few discussants have highlighted the necessity of clearer guidelines on how to set the initial condition $u(0)$ and the constant c in the differential equation at the basis of the class of objective priors. We appreciate that **Giummolè and Ventura** performed a comparison for a specific directional model when the two settings proposed in the paper are employed. In the specific, there is a direct comparison with the classical noninformative prior for this type of model and the G-prior. We agree that the proposed settings lead to a prior with light tails, rendering it not suitable for specific cases; this was also mentioned by **Rubio and Steel**.

In fact, the suggestions given in the paper are based on initial illustrations. We agree that further research has to be carried out so to achieve a prior that has, in some circumstances, more appealing properties. For example, it will be appropriate to

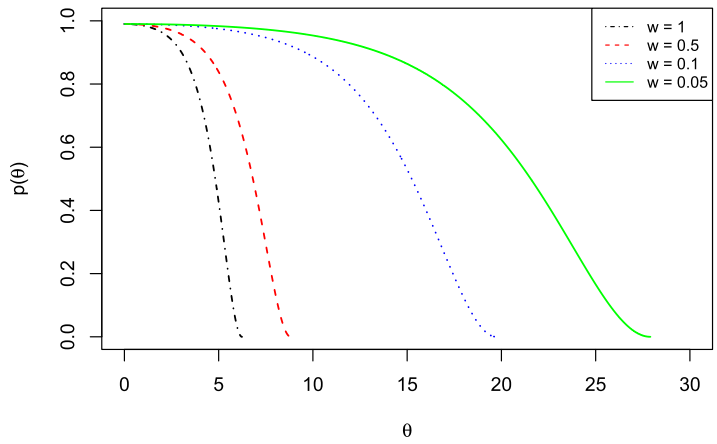


Figure 1: Plot of the positive half of the prior on $(-\infty, \infty)$ for different values of w.

identify a set within the class of priors with heavy tails. An additional idea, would be to consider different values of w (as pointed out by **Shestopaloff and Liverani**); in fact, if we look at Figure 1 in the paper, it is clear that the weighting parameter w can play a role in defining the shape of the tails of the distribution. Motivated by this example, we investigated further the tail behaviour of the prior with respect to the weighting parameter w. As an illustration, consider the case $(-\infty, \infty)$. Figure 1 shows the positive half of the prior with different values of w. It can be appreciated that tail of the prior gets heavier as w decreases.

Choosing different scoring rules could be another way to achieve heavy tails. On this regard, future work will be devoted to an extensive study of applying scoring rules we did not consider in the paper. For example, as suggested by **Giummolè and Ventura**, the Tsallis scoring rule may represent an appealing candidate to derive more robust priors.

As mentioned at the beginning of this rejoinder, the paper sets a new framework to derive objective priors and, therefore, there are still aspects must be explored, such as the calibration of c and $u(0)$, which was not formally and extensively studied in the paper.

An important remark about the choice of c and $u(0)$, is that they in appearance undermine the "objectivity" of the class of priors, as mentioned by **Liseo** and **Fouskakis**. However, the idea here is to obtain prior distributions that have a flavour of objectivity, given by the setting $S(\theta, p) =$ constant; but at the same time we would like to have prior distributions that can be actually implemented in reality. For example, as pointed out by **Zhou**, in spatial statistics it could be not straightforward to implement a reference or Jeffreys prior. Furthermore, it is well known that for some mixture models, both Jeffreys and reference prior are not suitable as they lead to improper posteriors (Grazian and Robert, 2018).

Finally, **Trangucci, Hansen and Chen** illustrate how the setting of c and $u(0)$ can be evaluated, when the obtained prior is proper, by measuring the informational content in the prior itself. In particular, they suggest the use of the observed prior effective sample size (OPESS) method, introduced in (Jones et al., 2020). We believe that this is an interesting idea that surely deserve further investigation.

4 Model selection

This area of research has been touched on by many discussants, including **Fouskakis**, **Consonni and García-Donato**, **Llorente, Martino and Delgado** and **Rubio and Steel**. Besides laying out some points of discussion, they have also contributed to potential directions for future research.

There is a fundamental point we would like to make about the implementation of the objective class of priors proposed in the paper. As mentioned above, the fact that it is a class of priors, implies that one should not apply the resulting priors blindly. Obviously, the decision-maker (or experimenter) has an idea on the problems she would like to solve and therefore the aim of her statistical analysis. If we consider Example 1 in **Consonni and García-Donato**, the discussants point out that the flat prior on the parameters of the model satisfies the predictive matching (Bayarri et al., 2012). Whilst, the proper solution outlined in our proposal does not. However, as we discuss in the paper, the flat prior is included in the class of objective priors proposed in the paper. In other words, the flat prior should not be considered as an alternative to our prior, given that the former is a special case of the latter.

For the Example 2 in **Consonni and García-Donato**, we agree that the conventional prior represents a better choice. The reason, as pointed out by the discussants, is the assumption of prior independence for the two parameters. In fact, the paper focusses mainly on the uni-dimensional case, and multi-dimensional parameter spaces are briefly touched (see Section 5.2). We thank the discussants as indeed future work should be developed on the multi-dimensional case and it will be very interesting to follow the suggestion of employing our methodology to obtain a prior that holds the adaptive structural property.

An interesting argument from **Fouskakis** concerns the evaluation of the Bayes Factor when the marginal likelihood is not analytically available. We would like to point out that in our examples we did not use analytical expressions for the Bayes Factor. The marginal likelihoods have been computationally evaluated. To this respect we welcome the contribution of **Llorente, Martino and Delgado** who propose a more efficient way to evaluate the Bayes Factor when our prior is employed.

In their discussions, **Fouskakis** and **Rubio and Steel**, mention the idea of non-local priors as a robust alternative to our approach. We would like to point out that non-local moment priors (Johnson and Rossell, 2010) can be obtained as a special case of our prior, as discussed in Walker (2020).

Another important application of Bayesian analysis for model selection is in variable selection problems. This is of particular interest when one deals with high-dimensional

models where sparsity is necessary. We illustrate the use of the proposed method for variable selection in a Poisson regression model (refer to the Supplementary Material). Although we do not compare our prior with available alternatives, it is clear that its performance is satisfactory. However, we welcome the suggestion and future research will be devoted to investigate this aspect. In particular, it would be important to explore the shrinking properties of the proposed prior when dealing with high-dimensional models. In fact, let us consider the prior in Figure 3(a) defined over the space $(0, \infty)$, and symmetrise it to extend it to $(-\infty, \infty)$. One can then choose appropriately c and $u(0)$ to have a prior that is narrowing around 0 to achieve a desired shrinking effect.

5 Other issues

In this section we will comment some interesting suggestions raised by the discussants.

Kypraios and **Shestopaloff and Liverani** raised the question if the proposed priors can be used in an approximate Bayesian computational setting and, in particular, if it is possible to sample from the prior. Regarding sampling from the prior; it can be done using a Metropolis–Hastings algorithm with the same Taylor expansion approach illustrated in the paper. However, we believe that this aspect could be improved in some specific cases by considering other methods, such as rejection sampling or, if p is log-concave, adaptive rejection sampling.

Giummolè and Ventura and **Iacopini, Ravazzolo and Rossini** highlight the use of different scoring rules in our approach. Indeed, as **Parry** points out, our paper sets a general approach which is not confined to specific scoring rules. However, as we warn in the paper, in order to avoid a trivial solution, such as the flat prior, it is important to include a scoring rule which includes first and second derivatives. Once this is ensured, the log-score can be replaced by other local scoring rules. This point is stressed by **Giummolè and Ventura**. It is interesting to see how **Iacopini, Ravazzolo and Rossini** propose to use a different scoring rule in an econometric framework.

The discussion of **Louzada, Ramos and Ramos** is interesting. We appreciate that they explored how to solve the differential equation in dimension k. There are two interesting points in their discussion, (1) They propose how to reduce the problem to a one-dimensional problem, and (2) They propose to use standard numerical analysis techniques to provide a polynomial approximation of the solution. We welcome this contribution as an additional tool to make our method more applicable.

Coad and Maruri-Aguillar wonder if our priors can coincide with Jeffreys' priors. We believe not in general, save possibly for priors on location parameters.

We welcome the discussion of **Zhou**. First, there is a discussion on priors used in spatial statistics. Second, the results displayed shows that, in a specific example, our approach is an improvement on the usual uniform prior approach. Furthermore, contrary to our priors, the author highlights the difficulty of implementing reference and Jeffreys' priors in this framework. We agree with **Zhou** and **Fouskakis** that a package will help to popularize the method.

Regarding the variational interpretation of the prior, **Trangucci, Hansen and Chen** propose an interesting connection with smoothing spline regression, where a loss function is minimized with respect to an unknown function of covariates while penalizing the roughness of the function (by penalizing its derivatives). We agree with their conclusion which suggests that our approach makes sense in terms of generating an objective prior because it maximizes the entropy of the prior but favor smoother distributions.

6 A final remark

The main contribution of the paper is to rejuvenate the idea of minimal information in a prior distribution, subject to some constraints. In particular, we aimed at going beyond Shannon's entropy and the log-score. From this perspective, it is clear that the notion of invariance is unnecessary, if not deleterious, as we have shown in Section 1. If one insists in looking for invariant procedures, then many aspects of the Bayesian framework will be undermined. To further see this, an interesting parallelism can be seen with classical statistics. In fact, while Maximum Likelihood Estimators (MLE) have invariance properties, it is well known that confidence intervals do not. But, regardless of this, classical procedures are widely used without any effort in rectifying, what we believe, is a "manufactured" problem.

References

Bayarri, M. J., Berger, J. O., Forte, A. and Garcia-Donato, G. (2009). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, **40**, 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1421

Grazian, C. and Robert, C. P. (2018). Jeffreys priors for mixture estimation: properties and alternatives. *Computational Statistics and Data Analysis*, **121**, 149–163. MR3759204. doi: https://doi.org/10.1016/j.csda.2017.12.005. 1420

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, **72**, 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 1421

Jones, D. E., Trangucci, R. N. and Chen, Y. (2020). Quantifying Observed Prior Impact. *arXiv*. arXiv:2001.10664. 1421

Walker, S. G. (2020). On a property of a non-local moment prior. To appear in *Communications in Statistics – Theory and Methods*. doi: https://doi.org/10.1080/03610926.2020.1804590. 1421