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

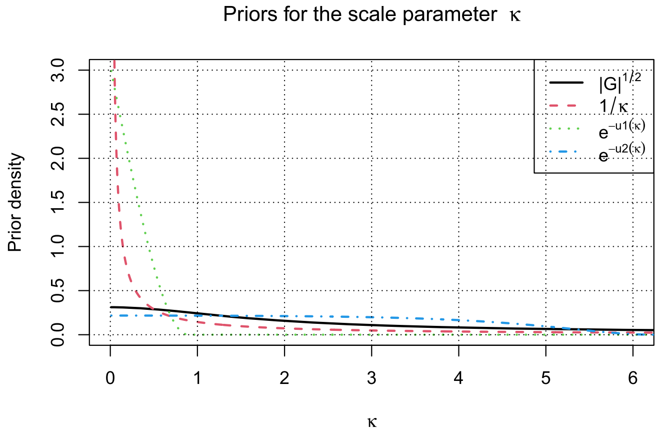

Figure 1: Priors $p(\kappa) \propto 1/\kappa$, $p_G(\kappa) \propto |G(\kappa)|^{1/2}$, $p_{u_1}(\kappa) \propto \exp\{-u_1(\kappa)\}$ and $p_{u_2}(\kappa) \propto \exp\{-u_2(\kappa)\}$.

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

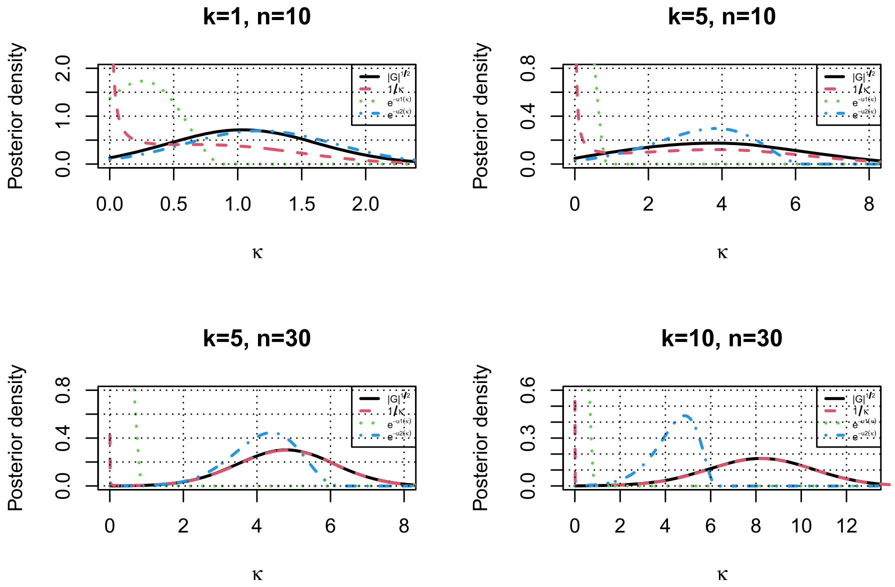

Figure 2: SR-posteriors with different priors, and values of n and κ.

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).