

The Version of Record of this Manuscript has been published and is available in TEST (1133-0686) from the 12 July 2018 on <https://link.springer.com/article/10.1007/s11749-018-0597-z>

Objective Bayesian inference with proper scoring rules

F. Giummolè⁺ and V. Mameli⁺

⁺*Dept. of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy*
giummole@unive.it, valentina.mameli@unive.it

E. Ruli* and L. Ventura*

^{*}*Dept. of Statistical Sciences, University of Padova, Italy*
ruli@stat.unipd.it, ventura@stat.unipd.it

July 6, 2018

Abstract

Standard Bayesian analyses can be difficult to perform when the full likelihood, and consequently the full posterior distribution, is too complex or even impossible to specify or if robustness with respect to data or to model misspecifications is required. In these situations, we suggest to resort to a posterior distribution for the parameter of interest based on proper scoring rules. Scoring rules are loss functions designed to measure the quality of a probability distribution for a random variable, given its observed value. Important examples are the Tsallis score and the Hyvärinen score, which allow us to deal with model misspecifications or with complex models. Also the full and the composite likelihoods are both special instances of scoring rules.

The aim of this paper is twofold. Firstly, we discuss the use of scoring rules in the Bayes formula in order to compute a posterior distribution, named SR-posterior distribution, and we derive its asymptotic normality. Secondly, we propose a procedure for building default priors for the unknown parameter of interest, that can be used to update the information provided by the scoring rule in the SR-posterior distribution. In particular, a reference prior is obtained by maximizing the average α -divergence from the SR-posterior distribution. For $0 \leq |\alpha| < 1$, the result is a Jeffreys-type prior that is proportional to the square root of the determinant of the Godambe information matrix associated to the scoring rule. Some examples are discussed.

Keywords: α -divergences, Composite likelihood, Godambe information, M -estimating function, Reference prior, Robustness, Scoring rule.

1 Introduction

In the Bayesian setting, when the full likelihood is too complex and difficult to specify or when robustness with respect to data or to model misspecifications is required, several

authors have proposed the use of surrogate likelihoods in the Bayes formula, in place of the full likelihood. Although this approach is non-orthodox, it is widely used in the recent statistical literature and theoretically motivated in several papers. See, among others, Pauli *et al.* (2011) and Ribatet *et al.* (2012) for the use of composite likelihoods, Lazar (2003), Lin (2006), Schennach (2005), Greco *et al.* (2008), Chang and Mukerjee (2008), Ventura *et al.* (2010), and Yang and He (2012) for the use of empirical and quasi-likelihoods; see also the review by Ventura and Racugno (2016), which addresses the use of pseudo-likelihoods in the Bayesian framework.

To deal with complex models or model misspecifications, useful surrogate likelihoods can be obtained through proper scoring rules. A scoring rule (see, for instance, the recent overviews by Machete, 2013, and Dawid and Musio, 2014, and references therein) is a special kind of loss function designed to measure the quality of a probability distribution for a random variable, given its observed value. It is proper if it encourages honesty in the probability evaluation. Proper scoring rules supply unbiased estimating equations for any statistical model, which can be chosen to increase robustness or for ease of computation. The Brier score (Brier, 1950), the logarithmic score (Good, 1952), the Tsallis score (Tsallis, 1988), and the Hyvärinen score (Hyvärinen, 2005) are widely known. In particular, when using the logarithmic score, the full likelihood and the composite likelihood (Varin *et al.*, 2011) are obtained as special cases of proper scoring rules (see for instance Dawid and Musio, 2014). While frequentist scoring rule inference has been widely discussed (see Dawid *et al.*, 2016, and references therein), Bayesian inference is relatively unexplored. Few exceptions are Dawid and Musio (2015) who focus on Bayesian model selection, Ghosh and Basu (2016) who consider robust Bayes estimation using the density power divergence measure, and Musio *et al.* (2017) who give an illustration of Bayesian inference for directional data through the Hyvärinen score; see also Pauli *et al.* (2011) and Ribatet *et al.* (2012) for the use of composite likelihoods in the Bayes formula.

To perform Bayesian inference, a suitable prior distribution on the parameter of interest must be elicited. In this paper we focus on default priors which are frequently used in Bayesian applications and which are still an active area of research (see, among others, Berger, 2006, Berger *et al.*, 2009, 2012, Ghosh, 2011, Walker, 2016, Leisen *et al.*, 2017). The two most common objective priors are the Jeffreys prior (Jeffreys, 1961), which uses the information about the parameter contained in the Fisher information, and the reference prior (Bernardo, 1979), which is based on the maximization of a distance in information between the prior and the posterior.

The goal of this paper is twofold. The first aim is to discuss the use of scoring rules in order to compute a posterior distribution, called here the SR-posterior distribution. In particular, we suggest a SR-posterior distribution obtained by extending to the general scoring rule setting the curvature adjustment of the composite likelihood proposed by Chandler and Bate (2007) and Ribatet *et al.* (2012). The calibration is necessary in order to show that the SR-posterior distribution is, up to order $O_p(n^{-1/2})$, normally distributed, with the same asymptotic variance of the scoring rule estimator. Indeed, a calibrated scoring rule is needed to reach the right asymptotic variance in the normal approximation, as well as a correct

shape of the posterior distribution.

The second aim is to propose the elicitation of a default prior for the unknown parameter of interest. In particular, we focus on reference priors as pioneered by Bernardo (1979); for a review see Bernardo (2005) and Ghosh (2011). Our purpose is to construct reference priors obtained by maximizing α -divergences from the SR-posterior distribution. The α -divergences are a well known class of discrepancy functions which include as a special case the Kullback-Leibler divergence. We show that, for $0 \leq |\alpha| < 1$, the maximizer is a Jeffreys-type prior that is proportional to the square root of the determinant of the Godambe information matrix.

The paper unfolds as follows. Section 2 gives background on scoring rules. Section 3 discusses the properties of the SR-posterior distribution. Section 4 presents the construction of reference priors based on α -divergences using SR-posteriors. Examples are presented in Section 5 both in the context of complex models and in the context of robustness theory. Further suggestions and comments can be found in the conclusions.

2 Scoring rules

A scoring rule is a loss function designed to measure the quality of a given probability distribution Q for a random variable X , in view of the result x of X ; see Dawid (1986). The function $S(x; Q)$ takes values in \mathbb{R} and its expected value under P will be denoted by $S(P; Q)$. The scoring rule S is called proper relative to the class of distributions \mathcal{P} if the following inequality is satisfied for all $P, Q \in \mathcal{P}$:

$$S(P; Q) \geq S(P; P). \quad (1)$$

It is strictly proper relative to \mathcal{P} if equation (1) is satisfied with equality if and only if $Q = P$. Note that in the following we identify a distribution Q by its probability density q with respect some measure μ ; so the two notations $S(x; q)$ and $S(x; Q)$ are indistinguishable.

An important example of proper scoring rules is the logarithmic score, which is defined as $S(x; Q) = -\log q(x)$ (Good, 1952) and corresponds to minus the log-likelihood function. Other widely used proper scoring rules are the Tsallis and the Hyvärinen scores. The Tsallis score (Tsallis, 1988, Basu *et al.*, 1998) is given by

$$S(x; Q) = (\gamma - 1) \int q(y)^\gamma d\mu(y) - \gamma q(x)^{\gamma-1}, \quad \gamma > 1. \quad (2)$$

The Tsallis score gives in general robust procedures (Basu *et al.*, 1998, Dawid *et al.*, 2016), and γ is a tuning constant that governs the trade-off between efficiency and robustness. The Hyvärinen score in its original formulation (Hyvärinen, 2005) with $x \in \mathbb{R}^d$ is defined as

$$S(x, Q) = \Delta \log q(x) + \frac{1}{2} \|\nabla \log q(x)\|^2, \quad (3)$$

where $\|\cdot\|$ is the standard norm, ∇ denotes the gradient vector and Δ the Laplacian operator.

2.1 Inference based on scoring rules

Let $x = (x_1, \dots, x_n)$ be the realization of a random sample of size n from a distribution with density function $p(x|\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$. Moreover, let $L(\theta) = L(\theta; x)$ and $\ell(\theta) = \log L(\theta; x)$ be the likelihood and the log-likelihood functions, respectively.

The validity of inference about θ using scoring rules can be justified by invoking the general theory of unbiased M -estimating functions. Indeed, inference based on proper scoring rules is a special kind of M -estimation (see, e.g., Dawid *et al.*, 2016, and references therein). The class of M -estimators is broad and includes a variety of well-known estimators. For example it includes the maximum likelihood estimator (MLE), the maximum composite likelihood estimator (see e.g. Varin *et al.*, 2011), and robust estimators (see e.g. Hampel *et al.*, 1986, Huber and Ronchetti, 2009), among many others.

Given a proper scoring rule $S(x; P_\theta) = S(x; \theta)$, let us denote by $S(\theta) = \sum_{i=1}^n S(x_i; \theta)$ the total empirical score. Moreover, let $s(x; \theta)$ be the gradient vector of $S(x; \theta)$ with respect to θ , i.e. $s(x; \theta) = \partial S(x; \theta) / \partial \theta$. Under broad regularity conditions (see Mameli and Ventura, 2015, and references therein), the scoring rule estimator $\tilde{\theta}$ is the solution of the unbiased estimating equation

$$S_\theta(\theta) = \sum_{i=1}^n s(x_i; \theta) = 0,$$

(see Dawid and Lautitzen, 2005, Dawid, 2007, Dawid *et al.*, 2016) and it is asymptotically normal, with mean θ and covariance matrix $V(\theta)/n$, with

$$V(\theta) = K(\theta)^{-1} J(\theta) (K(\theta)^{-1}), \quad (4)$$

where $K(\theta) = E_\theta(\partial s(X; \theta) / \partial \theta^T)$ and $J(\theta) = E_\theta(s(X; \theta) s(X; \theta)^T)$ are the sensitivity and the variability matrices, respectively. The matrix $G(\theta) = V(\theta)^{-1}$ is known as the Godambe information matrix (see Godambe, 1960) and its form is due to the failure of the second Bartlett identity (also called information identity) since, in general, $K(\theta) \neq J(\theta)$. In the special case of the logarithmic score, we have that $G(\theta) = K(\theta) = J(\theta)$ is the Fisher information matrix.

Asymptotic inference for scoring rule is usually based on first-order approximations, but several examples have shown the inaccuracy of these methods; see for instance Dawid *et al.* (2016) and Mameli and Ventura (2015). In order to improve the accuracy of first order methods, refinements have been obtained by resorting to higher-order asymptotic theory, which encompass the classical results for likelihood quantities while allowing for the failing of the information identity; see Mameli and Ventura (2015) and Mameli *et al.* (2017).

3 Posterior distributions based on scoring rules

The use of surrogate likelihoods in the Bayes formula has received great attention in the last decade (see the review by Ventura and Racugno, 2016, and references therein). In particular, the use of the composite likelihood, in the Bayes formula has been considered

for instance by Smith and Stephenson (2009), Pauli *et al.* (2011) and Ribatet *et al.* (2012). Since the composite likelihood does not satisfy the information identity, it is necessary to calibrate it in order to reach the right asymptotic variance in the normal approximation, as well as a correct shape of the posterior distribution. The correct curvature of the posterior distribution based on composite likelihoods can also be reached by using the composite score function in the Approximate Bayesian Computation (ABC) procedure, as discussed in Ruli *et al.* (2016).

While the use of surrogate likelihoods for Bayesian inference has been widely discussed in the recent literature to deal with complex models or robustness, the use of scoring rules in the Bayesian framework has not been deeply explored; exceptions are Dawid and Musio (2015), Ghosh and Basu (2016) and Musio *et al.* (2017).

Paralleling the derivation of posterior distributions based on composite likelihoods, a SR-posterior distribution can be obtained by using a scoring rule instead of the full likelihood in Bayes formula. However, since the scoring rule does not satisfy the information identity, it must be suitably calibrated before deriving the SR-posterior distribution. In particular, here we suggest a SR-posterior distribution obtained by extending to the scoring rule setting the curvature adjustment of the composite likelihood proposed by Chandler and Bate (2007) and Ribatet *et al.* (2012). In the following we study the asymptotic normality of the SR-posterior distribution; this result validates the use of scoring rule in a Bayes formula (see Ventura and Racugno, 2016, Section 3.3).

Let $\pi(\theta)$ be a prior distribution for the parameter θ . The proposed SR-posterior distribution is defined as

$$\pi_{SR}(\theta|x) \propto \pi(\theta) \exp \{-S(\theta^*)\}, \quad (5)$$

with $\theta^* = \theta^*(\theta) = \tilde{\theta} + C(\theta - \tilde{\theta})$, where C is a $d \times d$ fixed matrix such that $C^T K(\theta) C = G(\theta)$. A possible choice of the matrix C is given by $C = M^{-1} M_A$, with $M_A^T M_A = G$ and $M^T M = K$; for details, see Ribatet *et al.* (2012) and references therein. Note that in formula (5), when $S(\theta) = -\sum_{i=1}^n \log p(x_i|\theta)$ is the logarithmic scoring rule, the posterior $\pi_{SR}(\theta|x)$ reduces to the classical posterior distribution based on the likelihood function, i.e. $\pi(\theta|x) \propto \pi(\theta) \exp \{\sum_{i=1}^n \log p(x_i|\theta)\}$.

The choice of a prior distribution $\pi(\theta)$ to be used with a scoring rule involves the same problems typical of the standard Bayesian perspective, namely the elicitation of a proper prior distribution and the fact that default prior distributions can often be improper. The study of the possible impropriety of the SR-posterior has to be considered case-by-case, similarly to what happens for the genuine posterior distribution based on a default prior.

Under the same regularity conditions necessary for the asymptotic results for scoring rule inference (see for instance Mameli and Ventura, 2015) as $n \rightarrow \infty$, it can be shown that the scoring rule posterior distribution (5) is, up to order $O_p(n^{-1/2})$, normally distributed with mean $\tilde{\theta}$ and variance $H(\tilde{\theta})^{-1}/n$, i.e.

$$\pi_{SR}(\theta|x) \sim N_d \left(\tilde{\theta}, \frac{H(\tilde{\theta})^{-1}}{n} \right), \quad (6)$$

where $H(\theta) = C^T(\partial^2 S(\theta)/\partial\theta^2)C/n$. Note that $H(\tilde{\theta})$ converges almost surely to $G(\theta)$ as $n \rightarrow \infty$. This result is stated in the next Theorem 3.1 that gives the expansion of the scoring rule posterior distribution (5) up to third order.

Before stating Theorem 3.1, we introduce the notation which will be used in the rest of the paper. We use indices to denote derivatives of $S(x; \theta)$ and $\pi(\theta)$ with respect to the components of the parameter θ . Therefore, for example, $S_{jkh}(\theta) = \partial^3 S(\theta)/\partial\theta^j\partial\theta^k\partial\theta^h$, $S_{ijkl}(\theta) = \partial^4 S(\theta)/\partial\theta^i\partial\theta^j\partial\theta^k\partial\theta^l$, for $1 \leq i, j, k, l \leq d$, $\pi_i(\theta) = \partial\pi(\theta)/\partial\theta^i$ and $\pi_{ij}(\theta) = \partial^2\pi(\theta)/\partial\theta^i\partial\theta^j$. In the sequel, a tilde over a quantity means evaluation of that quantity in $\tilde{\theta}$. For example, $\tilde{\pi} = \pi(\tilde{\theta})$ and $\tilde{H} = H(\tilde{\theta})$. Further, $\tilde{S}_{jkh} = S_{jkh}(\tilde{\theta})$, $\tilde{S}_{ijkl} = S_{ijkl}(\tilde{\theta})$, $1 \leq i, j, k, l \leq d$, $\tilde{\pi}_i = \pi_i(\tilde{\theta})$ and $\tilde{\pi}_{ij} = \pi_{ij}(\tilde{\theta})$. Let c_{ij} and \tilde{h}^{ij} be the elements of the matrices C and $\tilde{H}^{-1} = H(\tilde{\theta})^{-1}$, respectively. Moreover, we use the Einstein summation convention so that when an index appears twice in an expression, summation on that index is intended.

Theorem 3.1. *Let $w = (w^1, \dots, w^d)^T = n^{1/2}(\theta - \tilde{\theta})$. The SR-posterior distribution for w can be written as*

$$\pi_{SR}(w|x) = \phi_d(w; \tilde{H}^{-1}) [1 + n^{-1/2}A_1(w) + n^{-1}A_2(w)] + O_p(n^{-3/2}), \quad (7)$$

where $\phi_d(w; \tilde{H}^{-1})$ is the density of a d -variate normal distribution with zero mean vector and variance matrix \tilde{H}^{-1} ,

$$A_1(w) = \frac{\tilde{\pi}_i}{\tilde{\pi}} w^i - \frac{1}{6} \frac{\tilde{S}_{ijk}}{n} c_{ir} c_{js} c_{kt} w^r w^s w^t$$

and

$$\begin{aligned} A_2(w) = & \frac{1}{2} \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} (w^i w^j - \tilde{h}^{ij}) \\ & - \frac{1}{6} \frac{\tilde{\pi}_i}{\tilde{\pi}} \frac{\tilde{S}_{jkh}}{n} c_{jr} c_{ks} c_{ht} (w^i w^r w^s w^t - 3\tilde{h}^{ir} \tilde{h}^{st}) \\ & - \frac{1}{24} \frac{\tilde{S}_{ijkl}}{n} c_{ir} c_{js} c_{kt} c_{lu} (w^r w^s w^t w^u - 3\tilde{h}^{rs} \tilde{h}^{tu}) \\ & + \frac{1}{72} \frac{\tilde{S}_{ijk} \tilde{S}_{rst}}{n^2} \left[c_{ia} c_{jb} c_{kc} c_{rd} c_{se} c_{tf} (w^a w^b w^c w^d w^e w^f - 9\tilde{h}^{ab} \tilde{h}^{cd} \tilde{h}^{ef}) \right. \\ & \left. + c_{ia} c_{rb} c_{jc} c_{sd} c_{ke} c_{tf} (w^a w^b w^c w^d w^e w^f - 6\tilde{h}^{ab} \tilde{h}^{cd} \tilde{h}^{ef}) \right]. \end{aligned}$$

The proof is given in the Appendix.

In practice, Theorem 3.1 shows that the SR-posterior distribution (5) has the correct curvature, namely the variance of the posterior distribution (6) is asymptotically equivalent to that of the minimum scoring rule estimator. When in particular $S(x; \theta)$ is the logarithmic score, it can be shown that (7) reduces to the expansion given in Datta and Mukerjee (2004) for the classic posterior distribution $\pi(\theta|x) \propto \pi(\theta)L(\theta; x)$.

4 Reference priors obtained from α -divergences

The information on $\pi(\theta|x)$ induced by $\pi(\theta)$ may be measured in terms of a divergence $D(\cdot, \cdot)$ between the prior and the posterior distribution: the higher the divergence, the lower the influence of the prior on the posterior. Let $D_\pi(x) = D(\pi(\theta), \pi(\theta|x))$, and let $p(x)$ be the marginal distribution of X . Minimizing the information in a prior is equivalent to maximize the expected divergence D_π from the corresponding posterior, i.e. the functional

$$T(\pi) = \int_{\mathcal{X}} D_\pi(x)p(x)dx. \quad (8)$$

Here, we focus on the well-known family of α -divergences, defined as

$$D_\pi(x) = \frac{1}{\alpha(1-\alpha)} \int_{\Theta} \left\{ 1 - \left(\frac{\pi(\theta)}{\pi(\theta|x)} \right)^\alpha \right\} \pi(\theta|x)d\theta,$$

which for $\alpha \rightarrow 0$ reduces to the Kullback-Liebler divergence, for $\alpha = 1/2$ corresponds to twice the Hellinger distance and for $\alpha = -1$ is equivalent to the χ^2 -divergence.

In this section, we extend the results of Ghosh *et al.* (2011) and Liu *et al.* (2014) to the context of scoring rules. In particular, in the following theorem we propose reference priors obtained by maximizing the expected α -divergence from the SR-posterior distribution (5).

In the sequel, we denote by g_{ij} and g^{ij} the components of matrices $G(\theta)$ and $G(\theta)^{-1}$ respectively. Moreover, let $E_\theta(S_{ijk}/n) = B_{ijk}^S(\theta) + o(n^{-1/2})$ and let $a^S = (a_1^S, \dots, a_d^S)^T$, with $a_i^S = B_{khl}^S c_{ki} c_{hu} c_{lt} g^{ut}$. Similarly, we define $E_\theta(\ell_{ijk}/n) = B_{ijk}^\ell(\theta) + o(n^{-1/2})$ and $a^\ell = (a_1^\ell, \dots, a_d^\ell)^T$, with $a_i^\ell = B_{ihk}^\ell i^{hk}$. Here, i^{ij} are the components of the inverse of the Fisher information matrix $I = E_\theta(-\partial^2 \ell(\theta)/\partial \theta \partial \theta^T)$. Finally, $\Sigma = (\sigma^{ij})_{ij}$, with $\sigma^{ij} = n \text{Cov}_\theta(\hat{\theta}^i, \hat{\theta}^j)$, where $\hat{\theta}$ is the maximum likelihood estimator for θ . Here, all indices range between 1 and d .

Theorem 4.1. *When $0 \leq |\alpha| < 1$, the prior which asymptotically maximizes the expected α -divergence to the SR-posterior distribution (5) is*

$$\pi_G(\theta) \propto |G(\theta)|^{1/2}. \quad (9)$$

When $\alpha = -1$, the prior which asymptotically maximizes the expected χ^2 -divergence to the SR-posterior distribution (5) is such that

$$\frac{\partial \log \pi_\chi(\theta)}{\partial \theta} = \frac{1}{4} \left[6a^S G^{-1} + 4a^\ell I^{-1} + |G|^{-1} \frac{\partial |G|}{\partial \theta} (5G^{-1} - 4\Sigma) + 2G \nabla \cdot (5G^{-1} - 4\Sigma) \right] \Gamma, \quad (10)$$

where $\nabla \cdot G^{-1} = (\partial g^{1j}/\partial \theta^j, \dots, \partial g^{dj}/\partial \theta^j)^T$, $\nabla \cdot \Sigma = (\partial \sigma^{1j}/\partial \theta^j, \dots, \partial \sigma^{dj}/\partial \theta^j)^T$ and $\Gamma = (G^{-1} - 4I^{-1} + 4\Sigma)^{-1}$.

The proof is given in the Appendix.

It is important noticing that, if we use the full likelihood in the Bayes formula, (9) and (10) reduce to the results obtained in Liu *et al.* (2014) and, for the case of a scalar parameter, in Ghosh *et al.* (2011).

As shown in the Appendix, for $|\alpha| > 1$ and $\alpha = 1$ a maximizer for the expected α -divergence does not exist.

Note that, when dealing with the multidimensional parameter case, the use of the square root of the determinant of the Godambe information matrix may be questionable as in the case of the Jeffreys prior. Following the advices for default priors, one possibility is to assume that the components of θ are a priori independent and to use the one-dimensional reference prior for each of the parameters. An alternative is to consider a sequential scheme as in Berger and Bernardo (1992). On the other hand, it should be noted that the ordering of the parameters is relevant. Unless the practitioner has a specific ordering in mind, different orderings may lead to different reference priors.

For $\alpha = -1$, the preceding expression (10) depends on quantities related to both the scoring rule and the likelihood. In practice this makes the proposed prior useful only when the likelihood function is available, as for instance in the context of robust inference. We also note that the required calculations remarkably simplify in the scalar parametric case.

The proposed prior distribution (9) shares some important properties with the Jeffreys prior. The most relevant is invariance with respect to one-to-one changes in the parametrization. As shown in Liu *et al.* (2014), when the functional (8) is invariant, the maximizer is also invariant. Thus, invariance of the proposed prior distributions follows from the well known invariant properties of α -divergences. Anyway, for $0 \leq |\alpha| < 1$, it can be easily seen that $G(\theta)$ is a second order tensor so that, if $\psi(\theta)$ is a new parametrization, $G_{rs}(\psi) = G_{ij}(\theta(\psi))\theta_r^i(\psi)\theta_s^j(\psi)$, where $\theta_r^i = \partial\theta^i/\partial\psi^r$. Thus it follows that $\pi_G(\psi) = \pi_G(\theta(\psi))|\partial\theta(\psi)/\partial\psi|$.

We would like to stress that this is a general property that holds for every model and every scoring rule. Other properties of the proposed prior distribution (9) in general depend on the scoring rule under consideration. In the following we present two special examples.

4.1 The Tsallis scoring rule

Let X be distributed as a location model with parameter $\mu \in \mathbb{R}$, so that $p(x|\mu) = p_0(x - \mu)$. The Tsallis scoring rule in this case takes the following form

$$\begin{aligned} S(x; \mu) &= (\gamma - 1) \int p_0(x - \mu)^\gamma dx - \gamma p_0(x - \mu)^{\gamma-1} \\ &= \text{const} - \gamma p_0(x - \mu)^{\gamma-1}. \end{aligned}$$

Its derivatives with respect to μ are all functions of $x - \mu$, so that the corresponding expected values, i.e. $K(\mu)$, $J(\mu)$ and $G(\mu)$ are independent of μ . Thus, $\pi_G(\mu) \propto 1$, as the Jeffreys prior.

Also for scale models, $\pi_G(\sigma)$ with the Tsallis score coincides with the Jeffreys prior. Indeed, let X be distributed as a scale model with parameter $\sigma > 0$, so that $p(x|\sigma) =$

$p_0(x/\sigma)/\sigma$. The Tsallis scoring rule in this case reduces to

$$\begin{aligned} S(x; \sigma) &= (\gamma - 1) \int \frac{1}{\sigma^\gamma} p_0 \left(\frac{x}{\sigma} \right)^\gamma dx - \gamma \frac{1}{\sigma^{\gamma-1}} p_0 \left(\frac{x}{\sigma} \right)^{\gamma-1} \\ &= \frac{1}{\sigma^{\gamma-1}} \left[\text{const} - \gamma p_0 \left(\frac{x}{\sigma} \right)^{\gamma-1} \right]. \end{aligned}$$

It is easy to show that the first three derivatives of $S(x; \sigma)$ can be written as

$$S_\sigma = \frac{1}{\sigma^\gamma} f_1 \left(\frac{x}{\sigma} \right), \quad S_{\sigma\sigma} = \frac{1}{\sigma^{\gamma+1}} f_2 \left(\frac{x}{\sigma} \right), \quad S_{\sigma\sigma\sigma} = \frac{1}{\sigma^{\gamma+2}} f_3 \left(\frac{x}{\sigma} \right),$$

where f_1 , f_2 and f_3 are suitable functions. Notice that the expectation of any function of X/σ with respect to $p(x; \sigma)$ does not depend on σ itself. Thus, $J(\sigma) \propto 1/\sigma^{2\gamma}$ and $K(\sigma) \propto 1/\sigma^{\gamma+1}$, so that $G(\sigma) \propto 1/\sigma^2$ and $\pi_G(\sigma) \propto 1/\sigma$.

4.2 Hyvärinen scoring rule

Let X be a non-degenerate random variable belonging to the one-parameter natural exponential family

$$p(x|\theta) = \exp \{ \theta x - k(\theta) + a(x) \}, \quad x \in \mathbb{R}. \quad (11)$$

Let denote by $a'(x_i) = \partial a(x_i)/\partial x_i$ and $a''(x_i) = \partial^2 a(x_i)/\partial x_i^2$. The Hyvärinen total empirical score in this case reduces to

$$S(\theta) = - \left\{ 2 \sum_{i=1}^n a''(x_i) + \sum_{i=1}^n [\theta + a'(x_i)]^2 \right\}. \quad (12)$$

The Hyvärinen score estimator for this family is

$$\tilde{\theta} = - \frac{\sum_{i=1}^n a'(X_i)}{n},$$

which can be computed without knowledge of the normalizing constant $k(\theta)$; see Barndorff-Nielsen (1976), Hyvärinen (2007) and Parry *et al.* (2012). The Hyvärinen score estimator is an unbiased estimator for the parameter θ and its variance coincides with the inverse of the Godambe information $G(\theta)^{-1} \propto \text{Var}(a'(X))$; see Mamei and Ventura (2015). Thus, $\pi_G(\theta) \propto \text{Var}(a'(X))^{-1/2}$, while the Jeffreys prior is $\pi_J(\theta) \propto (\partial^2 k(\theta)/\partial \theta^2)^{1/2}$ which requires the knowledge of $k(\theta)$. In models in which $k(\theta)$ is difficult to evaluate, the Jeffreys prior is not available.

5 Examples

In many applications, 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. In this section, we consider the analysis of these statistical issues from a scoring rule viewpoint. In particular, we illustrate the behaviour of the SR-posterior distribution and of the corresponding reference prior by means of three examples. In these examples different non-informative priors are also considered. MCMC methods based on data augmentation, such as the Gibbs sampler, seem difficult to adapt in this setting. The reason is that, in general, the proposed prior and the associated scoring rule may not be conjugate and, the full conditionals may not correspond to known densities. Therefore, except for Example 5.1 in which different posteriors are computed by numerical integration, the posterior distributions are computed via the classical Metropolis-Hastings algorithm using a multivariate normal proposal. The R code for fully reproducing the examples of this section can be found in the Electronic Supplementary Material.

5.1 Hyvärinen scoring rule for the von Mises-Fisher distribution

The first example deals with the study of directional models (Mardia and Jupp, 2000). These models arise naturally in many areas of scientific research, such as meteorology for wind direction, biology for animal movement, in the investigation of biological processes and many others fields. 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 scoring rule. An important feature of this score is related to the homogeneity property for which the distribution needs to be known up to the normalization constant, thus avoiding the calculation of the density function.

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

$$p(z|\kappa) \propto \exp(-\kappa\mu_0^T z), \quad z \in \mathcal{S}_{q-1},$$

with $\kappa \in \mathbb{R}^+$ scalar concentration parameter and $\mu_0 = (\cos \theta_0, \sin \theta_0)$, with $\theta_0 \in \mathbb{R}$ known. When $q = 2$ and the data are represented in polar coordinates $(z_{h1}, z_{h2}) = (\cos \theta_h, \sin \theta_h)$, $h = 1, \dots, n$, the Hyvärinen estimator for κ can be derived in closed form and it is given by

$$\tilde{\kappa} = \frac{2\sqrt{\bar{R}^2(1 + \bar{R}_2^2) + 2(\bar{C}^2 - \bar{S}^2)\bar{C}_2 + 4\bar{C}\bar{S}\bar{S}_2}}{(1 - \bar{R}_2^2)}, \quad (13)$$

where $\bar{C} = \frac{1}{n} \sum_{h=1}^n \cos(\theta_h)$, $\bar{S} = \frac{1}{n} \sum_{h=1}^n \sin(\theta_h)$, $\bar{C}_2 = \frac{1}{n} \sum_{h=1}^n \cos(2\theta_h)$, $\bar{S}_2 = \frac{1}{n} \sum_{h=1}^n \sin(2\theta_h)$ and $\bar{R}_2 = \sqrt{\bar{C}_2^2 + \bar{S}_2^2}$.

For $\theta_0 = 0$, Mardia *et al.* (2016) show that the asymptotic variance of $\tilde{\kappa}$ is

$$\frac{V(\kappa)}{n} = \frac{\kappa [2\kappa - 3A_1(\kappa)]}{nA_1^2(\kappa)},$$

with $A_1(\kappa) = I_1(\kappa)/I_0(\kappa)$, where I_1 and I_0 are the modified Bessel functions of order 0 and 1, respectively. The reference prior (9) is thus

$$\pi_G(\kappa) \propto \sqrt{\frac{A_1^2(\kappa)}{\kappa [2\kappa - 3A_1(\kappa)]}}.$$

Figure 1 shows the plot of the reference prior $\pi_G(\kappa)$ and, for comparison, it also presents the classical non-informative prior $\pi(\kappa) \propto 1/\kappa$ for the scale parameter. It can be seen from the plot that $\pi_G(\kappa)$ puts finite mass at 0, leading to more appropriate inference when the true parameter values are close to the boundary of the parametric space. Actually, using properties of first order Bessel functions, it can be proved that $\lim_{\kappa \rightarrow 0^+} \pi_G(\kappa) = 2^{-1/2}$.

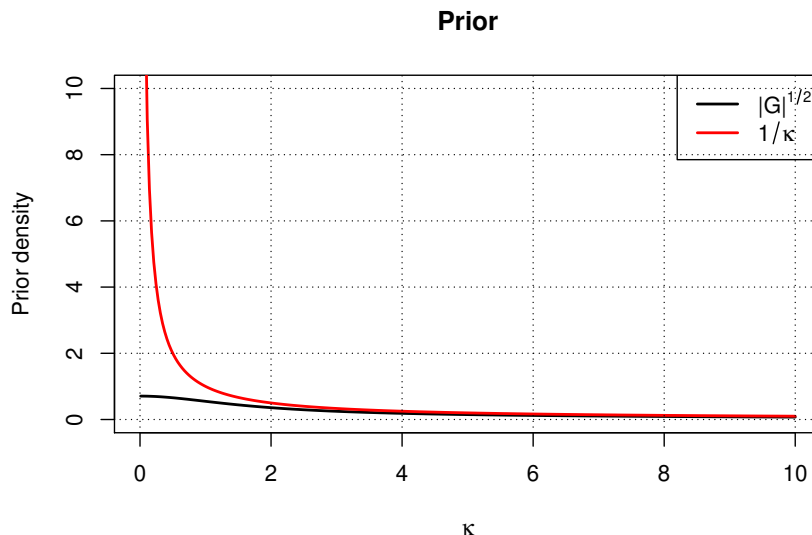


Figure 1: Priors $\pi_G(\kappa) \propto |G(\kappa)|^{1/2}$ and $\pi(\kappa) \propto 1/\kappa$.

In order to illustrate that the calibration of the Hyvärinen scoring rule is necessary to obtain a posterior distribution (5) with the right curvature, consider a sample of size $n = 50$ from the von Mises-Fisher distribution with $(\kappa, \mu_0) = (3, 0)$. Hereafter, we take the location parameter μ_0 to be fixed and equal to zero, and consider the problem of estimating κ . Figure 2 compares the full posterior (black line) based on the likelihood function of the von Mises-Fisher model, with the calibrated (red) and non calibrated (green) SR-posteriors based on the Hyvärinen scoring rule. Here, the non-informative prior $\pi(\kappa) \propto 1/\kappa$ is used. The vertical line corresponds to the Hyvärinen scoring rule estimate of κ , i.e. $\tilde{\kappa} = 2.36$. It can be noted that the non calibrated SR-posterior is too narrow compared with the genuine posterior, while the calibrated SR-posterior shows a curvature similar to the genuine posterior.

Genuine, cal. and no cal (default prior)

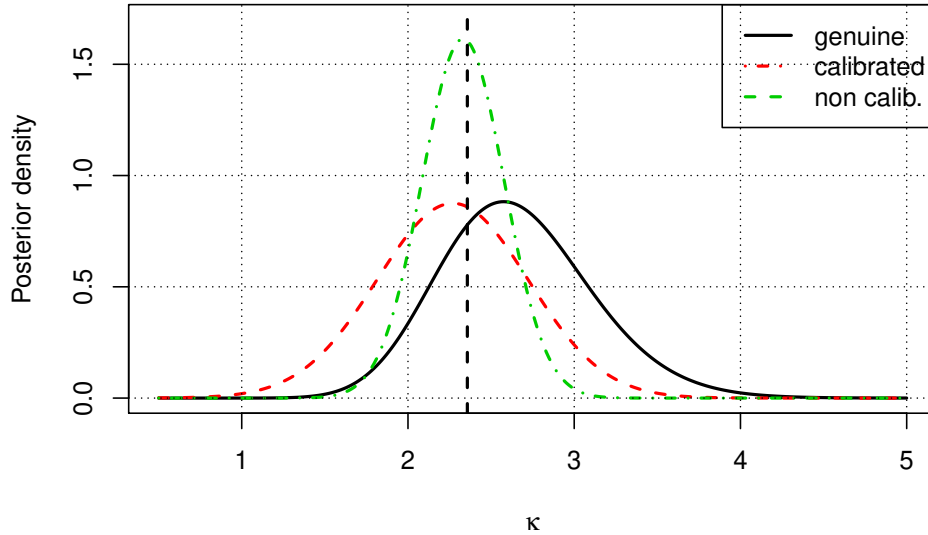


Figure 2: Comparison of calibrated SR-posterior vs non calibrated and proper posteriors.

To compare the behaviour of the SR-posterior distributions, based on the two priors $\pi_G(k) \propto |G|^{1/2}$ and $\pi(\kappa) \propto 1/\kappa$, two scenarios have been considered and the SR-posteriors with the two priors have been computed. The SR-posteriors are given in Figure 3 for different values of the parameters n and κ . In the first scenario, the true value of κ approaches the boundary and/or the sample size is small. In particular, we consider samples of size $n = 10, 30, 50$ with $\kappa = 1$. It can be noted that for small samples sizes or for the true value of κ near zero, the SR-posterior with $\pi(\kappa) \propto 1/\kappa$ may not be proper or puts too much mass at zero, i.e. it has a vertical asymptote at zero. On the other hand, the SR-posterior with the reference prior $\pi_G(\kappa)$ is proper and unimodal.

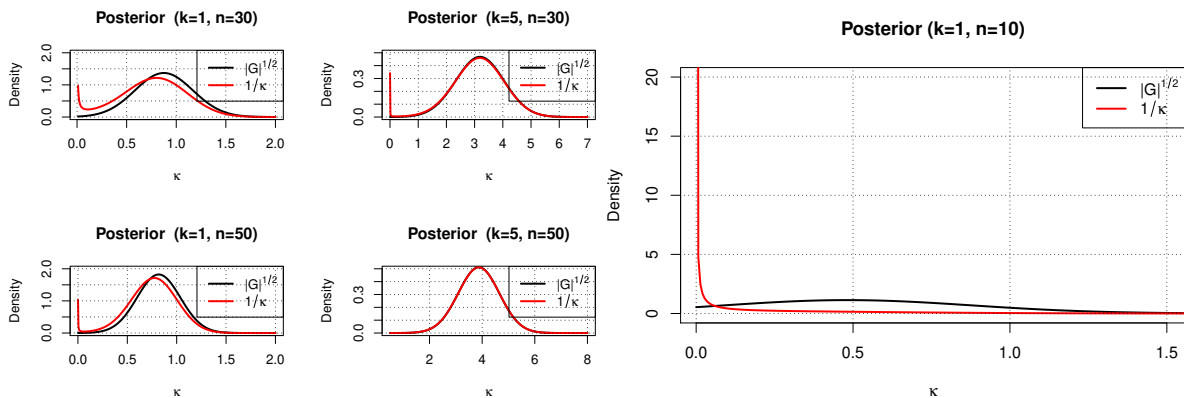


Figure 3: SR-posteriors with the $\pi_G(\kappa) \propto |G(\kappa)|^{1/2}$ prior and with the $\pi(\kappa) \propto 1/\kappa$ prior.

In the second scenario, the true value of κ is away from the boundary and the sample size

is moderate or large. We consider samples of size $n = 30, 50$, with $\kappa = 5$. The SR-posteriors in these cases look very similar, as it can be noted from Figure 3. Indeed, for high values of κ the two SR-posteriors are indistinguishable.

5.2 Pairwise likelihood for the multivariate equi-correlated normal model

Let X be a q -dimensional random vector with mean μ and covariance matrix Σ , with $\Sigma_{rr} = \sigma^2$ and $\Sigma_{rs} = \rho\sigma^2$ for $r \neq s$, with $r, s = 1, \dots, q$ and $\rho \in (-1/(q-1), 1)$. The pairwise log-likelihood for $\theta = (\mu, \sigma^2, \rho)$ is (see Pace *et al.*, 2011)

$$S_p(\theta) = -\frac{nq(q-1)}{2} \log \sigma^2 - \frac{nq(q-1)}{4} \log(1-\rho^2) - \frac{q-1+\rho}{2\sigma^2(1-\rho^2)} SSW + \\ - \frac{q(q-1)SSB + nq(q-1)(\bar{y} - \mu)^2}{2\sigma^2(1+\rho)},$$

where $SSW = \sum_{i=1}^n \sum_{r=1}^q (x_{ir} - \bar{x}_i)^2$, $SSB = \sum_{i=1}^n (\bar{x}_i - \bar{x})^2$, $\bar{x}_i = \sum_{r=1}^q x_{ir}/q$ and $\bar{x} = \sum_{i=1}^n \sum_{r=1}^q x_{ir}/(nq)$.

The reference prior (9) is proportional to the square root of the determinant of the Godambe information matrix whose components are given in Pace *et al.* (2011).

Consider, first, the simplest situation with μ and σ^2 known. The SR-posterior distribution based on the pairwise likelihood with the uniform prior and the proposed reference prior is computed for different values of ρ and q (see Figure 4). We notice that the two priors give very similar results when the value of the parameter ρ is away from zero.

Consider now the multi-parametric case. We compare the SR-posteriors based on the pairwise likelihood with the reference prior and with a non-informative prior derived by using two alternative parametrizations:

- (a) $\theta = (\mu, \sigma, \rho)$, for which we assume the non-informative prior $\pi(\theta) \propto 1/\sigma$.
- (b) $\xi = (\mu, \tau, \kappa)$, with $\tau = \log(\sigma)$ and $\kappa = \text{logit}(\rho)$, for which we assume the flat prior $\pi(\xi) \propto 1$.

The need for calibration of the composite likelihood has been studied by Ribatet *et al.* (2012); see also Pauli *et al.* (2011). The calibrated SR-posterior distribution has been computed with the parametrization ξ by using formula (6), in order to avoid the constraints on $\theta \in \Theta$.

As a first example, consider $n = 10$, $q = 10$ and a sample generated under the equi-correlated normal model with $\theta = (0, 1, 0.5)$. The SR-posterior distributions with the reference prior and the two non-informative priors in cases (a) and (b) are shown in Figure 5. The SR-posterior distributions, described by means of histograms, show no appreciable differences.

As a second example, consider the more problematic scenario with $n = 10$, $q = 4$ and data generated under the equi-correlated normal model with $\theta = (0, 0.5, 0.1)$. The SR-posteriors with the reference prior and the two non-informative priors are shown in Figure 6.

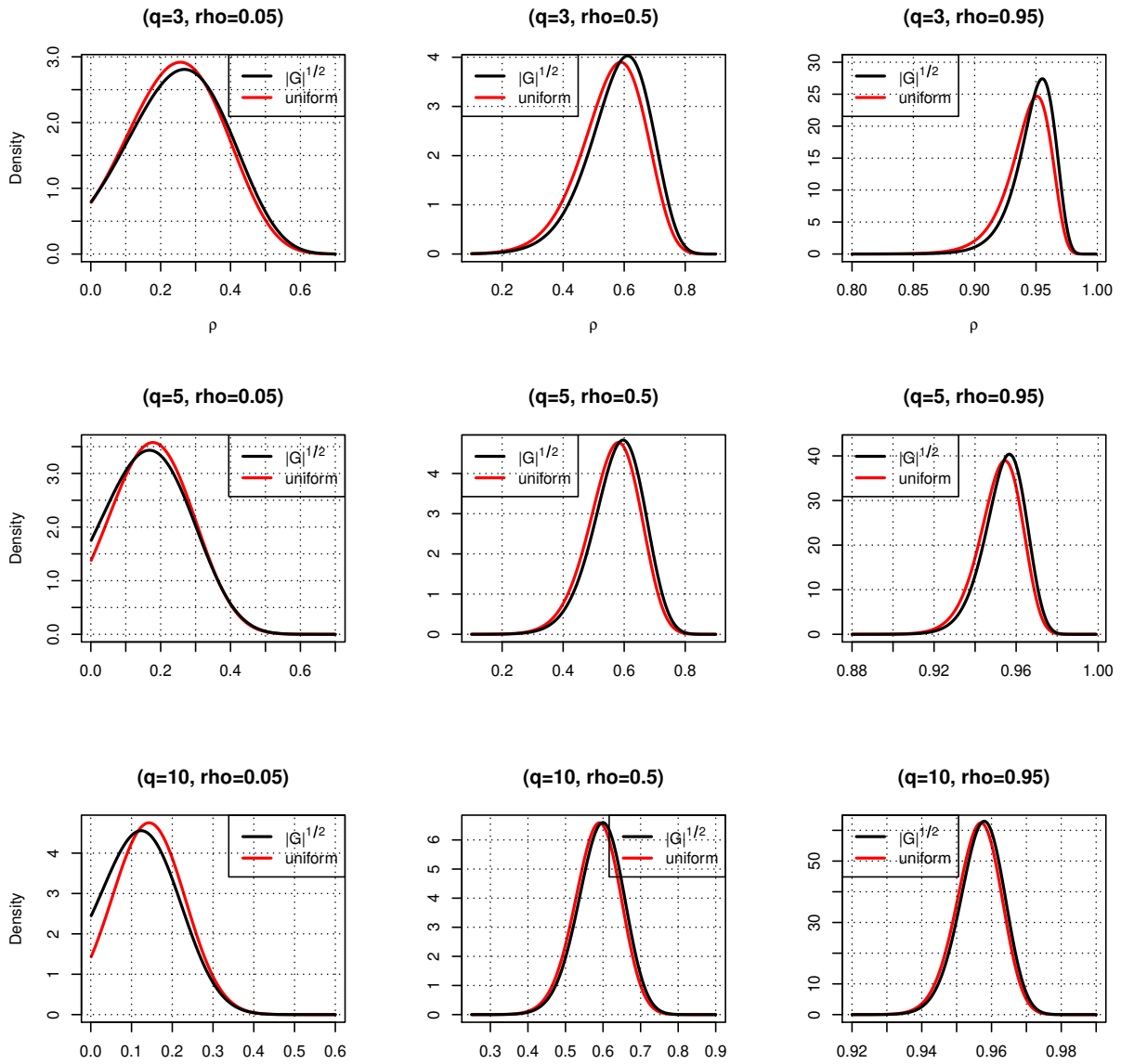


Figure 4: SR-posteriors with the $\pi_G(\rho)$ and with the uniform prior in $(0,1)$ for different values of ρ and q where $n = 10$, $\sigma^2 = 1$ and $\mu = 0$.

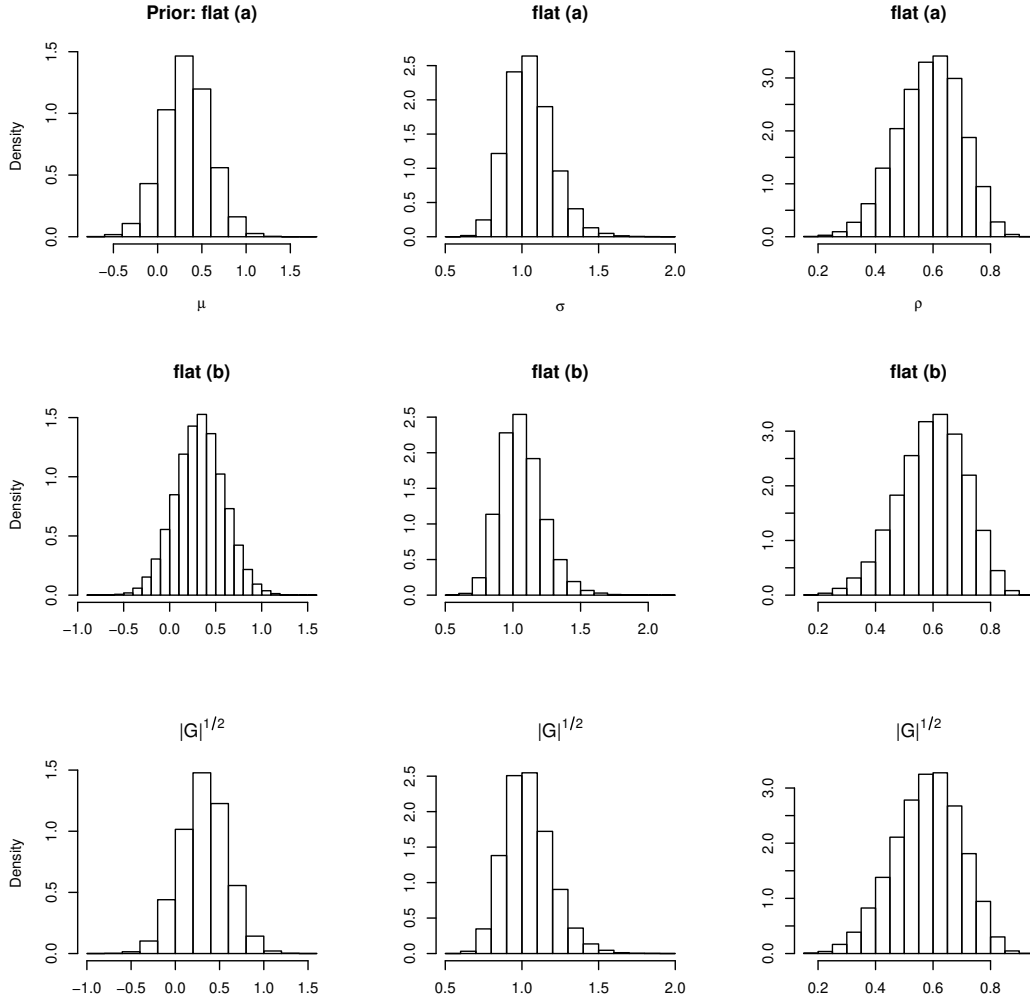


Figure 5: SR-posterior with the reference prior $\pi_G(\theta)$ compared with the two alternative flat priors (a) and (b). The data are generated from the equi-correlated model with $n = 10$, $q = 10$, $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0.5$

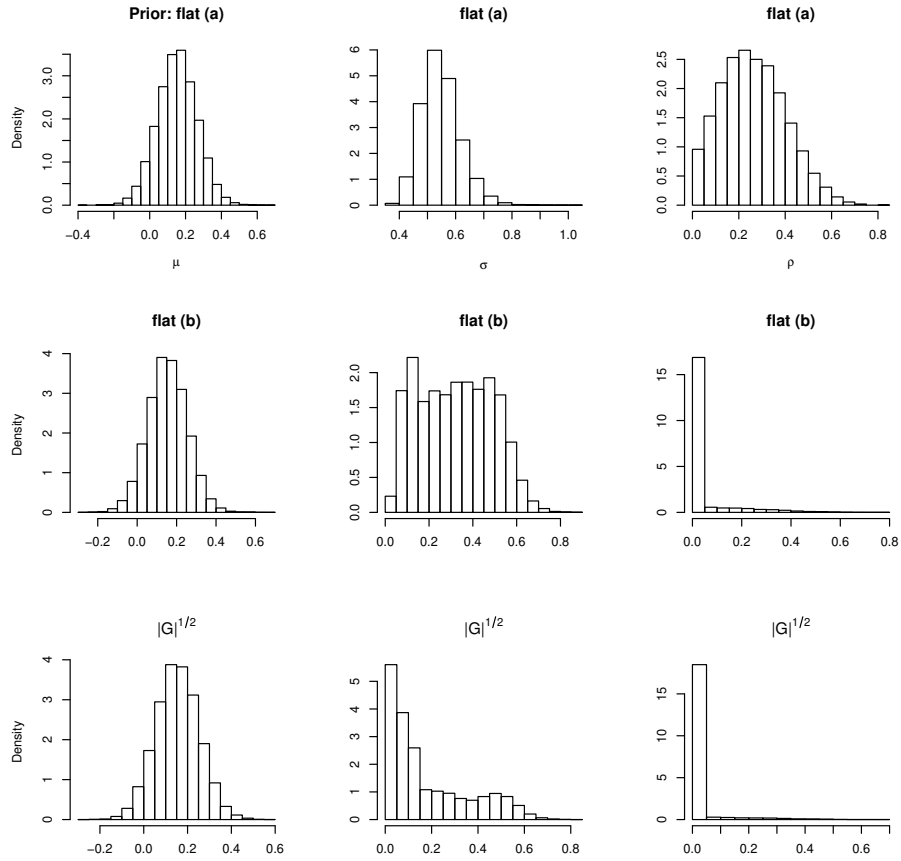


Figure 6: SR-posterior with the reference prior $\pi_G(\theta)$ (third row) compared with the two alternative flat priors (a) (first row) and (b) (second row). The data are generated from the equi-correlated model with $n = 10$, $q = 4$ $\mu = 0$, $\sigma^2 = 0.5$ and $\rho = 0.1$

It can be noted that a flat prior on θ may not necessarily be flat also on ξ , at least not for σ and ρ . Indeed, the uniform prior on ρ implies a standard logistic prior on κ , which is quite informative. This explains the “regularised” behaviour of the marginal posteriors of ρ and σ in case (a). On the other hand, the proposed reference prior behaves quite similarly to the uniform improper prior $\pi(\xi) \propto 1$.

5.3 Tsallis scoring rule for linear regression models

The third example is related to robustness procedures, which are more stable and reliable than their classical likelihood-based counterparts under misspecification. As it is well known, the ordinary Bayes estimator based on the classical based likelihood posterior density lacks robustness. Gosh and Basu (2016) studied the robustness properties of the posterior distribution based on the Tsallis scoring rule. There is a trade-off between robustness and efficiency in the Tsallis scoring rule: small values of γ in (2) entails to a small loss in efficiency but a reduced robustness; while large values of γ are associated to great robustness but low efficiency. Here, we consider a linear regression model to demonstrate the robustness performance of the Tsallis scoring rule and of the derived prior.

Consider the linear regression model

$$y = X\beta + \sigma\epsilon,$$

where X is a fixed $n \times p$ matrix, $\beta \in \mathbb{R}^p$ ($p \geq 1$) is the vector of unknown regression coefficients, $\sigma > 0$ is a scale parameter, and ϵ is an n -dimensional vector of random errors from a standard normal distribution.

Let $\theta = (\beta, \sigma^2)$. The Tsallis total empirical score in this case assumes the following form (Ghosh and Basu, 2013)

$$S_T(\theta) = \frac{\gamma}{(2\pi\sigma^2)^{\frac{\gamma-1}{2}}} \sum_{i=1}^n e^{-\frac{(\gamma-1)}{2\sigma^2}(y_i - x_i^T\beta)^2} - \frac{n(\gamma-1)}{\sqrt{\gamma}(2\pi\sigma^2)^{(\gamma-1)/2}}.$$

The asymptotic variance of the Tsallis estimator of θ is given in Ghosh and Basu (2013). In particular, the asymptotic distribution of $(X^T X)^{1/2}(\tilde{\beta} - \beta)$ is a $N_p(0, v_\gamma^\beta I_p)$, with $v_\gamma^\beta = \sigma^2 \left(1 + \frac{(\gamma-1)^2}{2\gamma-1}\right)^{3/2}$, while $\sqrt{n}(\tilde{\sigma}^2 - \sigma^2)$ follows a normal distribution with mean 0 and variance v_γ^e , where

$$v_\gamma^e = \frac{4\sigma^4}{(2 + (\gamma-1)^2)^2} \left(2(1 + 2(\gamma-1)^2) \left(1 + \frac{(\gamma-1)^2}{2\gamma-1}\right)^{5/2} - (\gamma-1)^2\gamma^2 \right).$$

Since the asymptotic distributions of $\tilde{\beta}$ and $\tilde{\sigma}^2$ are independent, the proposed reference prior is proportional to the square root of the determinant of the Godambe information, i.e.

$$\pi_G(\theta) \propto (v_\gamma^\beta v_\gamma^e)^{-1/2}. \quad (14)$$

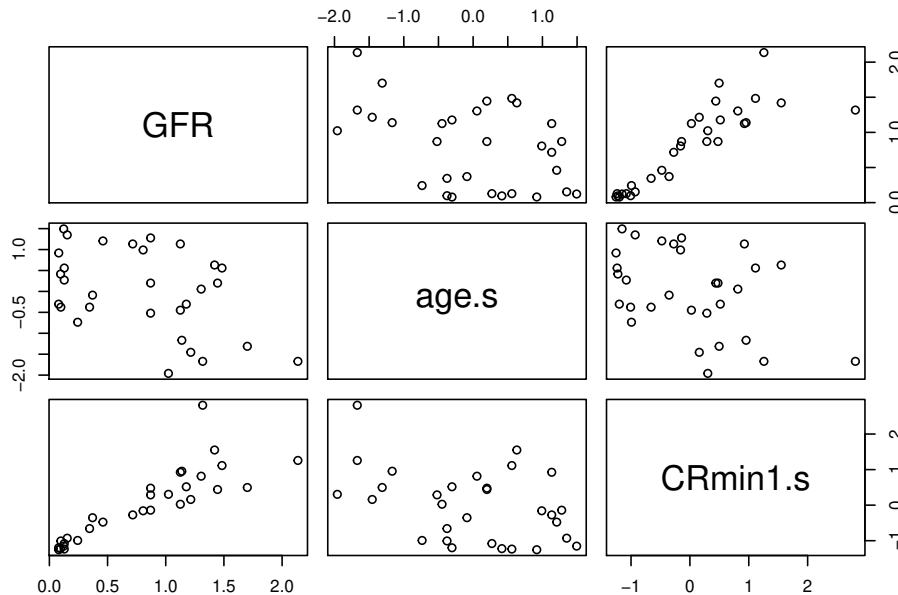


Figure 7: Scatterplot diagram of GFR data.

Therefore, a prior for β is $\pi_G(\beta) \propto (v_\gamma^\beta)^{-1/2}$, while a prior for σ^2 is $\pi_G(\sigma^2) \propto (v_\gamma^e)^{-1/2}$.

In order to illustrate the behaviour of the Tsallis scoring rule in the context of the linear regression model, we consider the GFR dataset (Heritier *et al.*, 2009), which contains measurements of the glomerular filtration rate (GFR) and serum creatinine (CR) on $n = 30$ subjects. The GFR is the volume of fluid filtered from the renal glomerular capillaries into the Bowmans capsule per unit of time (typically in millilitres per minute) and, clinically, it is often used to determine renal function. Its estimation, when not measured, is of clinical importance and several techniques are used for that purpose. One of them is based on CR, an endogenous molecule, synthesized in the body, which is freely filtered by the glomerulus (but also secreted by the renal tubules in very small amounts). Several models have been proposed in the literature to explain the logarithm of GFR as a function of CR. Here, following Heritier *et al.* (2009), we consider a model for GFR based on CR^{-1} and AGE, i.e.

$$\text{GFR} = \beta_0 + \beta_1 \frac{1}{\text{CR}} + \beta_2 \text{AGE} + \varepsilon.$$

Both covariates are scaled to have mean zero and unit variance, whereas the response is scaled to have unit variance. The data are illustrated in Figure 7: note that there are some observations which look like outliers. The need and consequence to resort to robust procedures to deal with these data have been highlighted in several papers; see, among others, Heritier *et al.* (2009) and Farcomeni and Ventura (2012).

For the parameter $\theta = (\beta, \log \sigma)$ we assume both the usual non-informative flat prior $\pi(\theta) \propto 1$ and the proposed reference prior (14) in the SR-posterior distribution, while the classical posterior distribution is based on the non-informative flat prior. Figure 8 gives the violin plots of the marginal SR-posterior distributions based on the Tsallis score, with

$\gamma = 1.25$, and of the classical posterior distribution. From Figure 8 it can be noted that the proposed reference prior behaves similarly to the non-informative prior. Moreover, the classical marginal posterior distribution shows in general heavier tails and, in particular, the robust and the classical posterior distributions give different inferences on β_2 and σ .

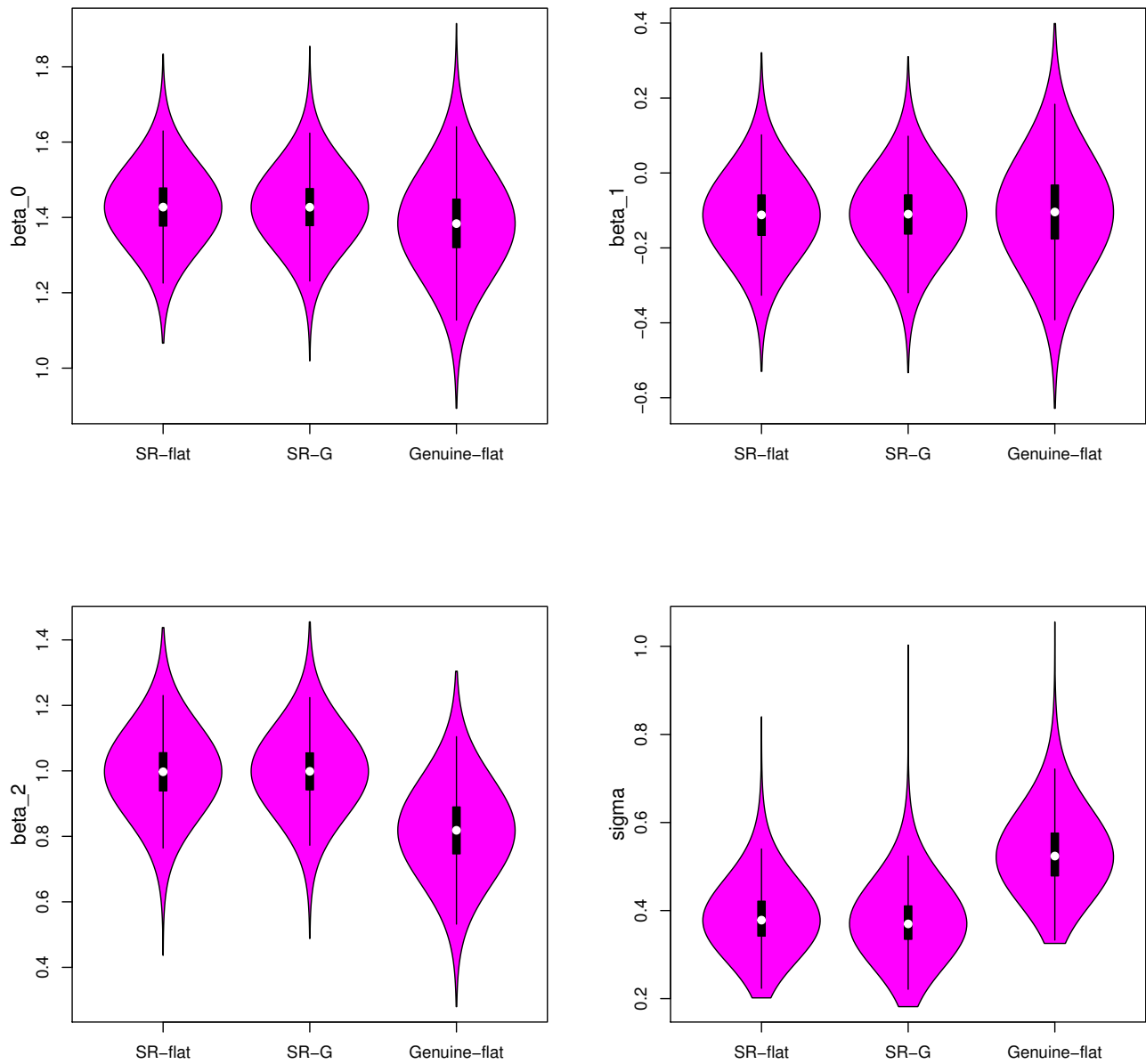


Figure 8: SR-posteriors based on the Tsallis score with $\gamma = 1.25$ (with the flat or the reference prior) vs the genuine posterior for the GFR dataset.

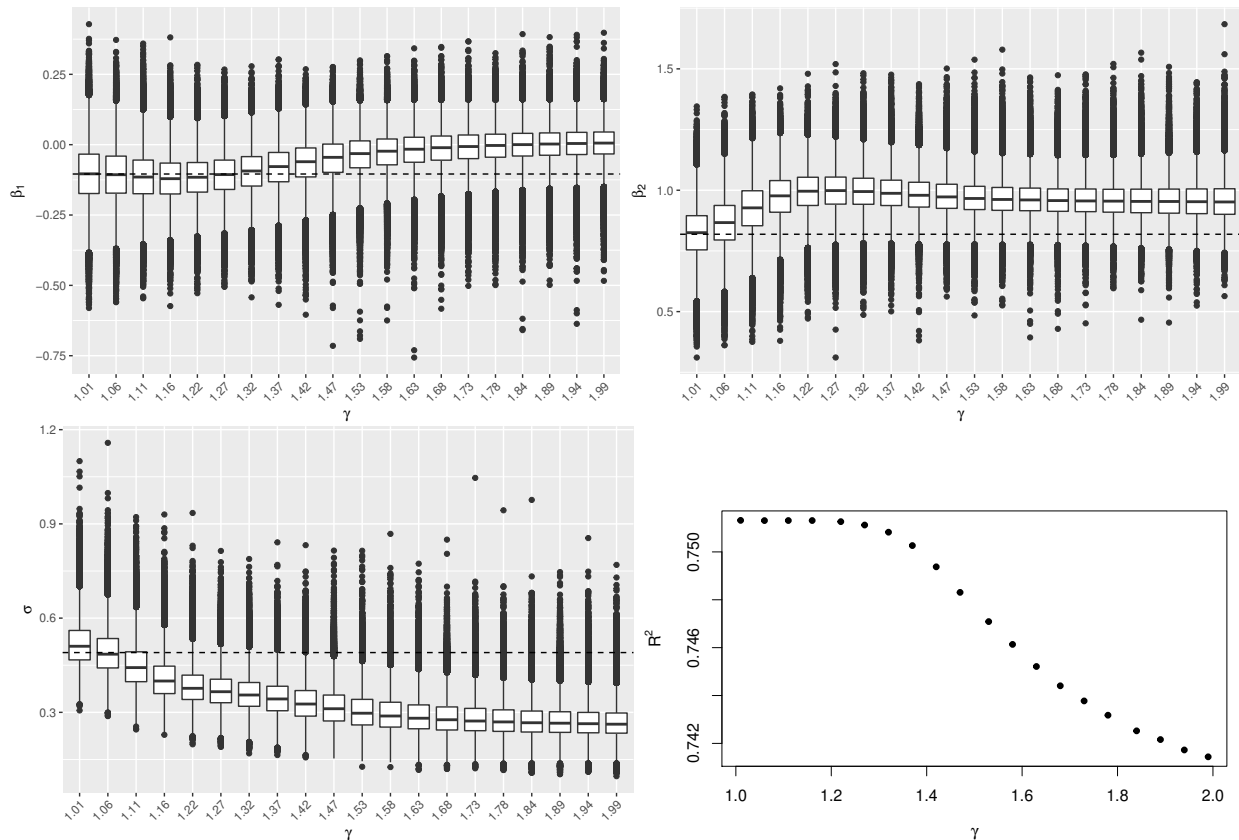


Figure 9: Marginal SR-posteriors based on the Tsallis score, for β_1, β_2 and σ with the reference prior, and values of R^2 for varying values of γ .

Figure 9 gives useful monitoring plots in robustness studies. In particular, these plots illustrate the SR-posterior distributions based on the Tsallis score with reference prior as a function of γ . This approach (see, e.g., Riani *et al.*, 2014) provides tools for gaining knowledge and better understanding of the properties of robust procedures. The horizontal lines in Figure 9 correspond to the posterior mode of the genuine posterior based on the flat prior. The value $\gamma = 1$ corresponds to complete overlap of the two methods. As far as γ increases, robustness is achieved although at the price of losing efficiency at the central model; the value $\gamma = 1.25$ gives approximately 0.95% efficiency under the normal distribution.

As a final remark, note that the Tsallis scoring rule provides a cogent framework for dealing with robust procedures in the Bayesian framework.

6 Conclusions

In this paper, we discuss the use of scoring rules in order to compute a posterior distribution, useful to deal with complex models or if robustness with respect to data or to model misspecification is required. Indeed, scoring rules provide a flexible and robust way of com-

binning data-driven information with prior distributions, either subjective or non-informative.

One should devise a scoring rule that properly captures the structure of the data, otherwise the resulting posterior inferences are not reliable. In this respect, the result of Section 3 shows that a suitably calibrated SR-posterior distribution is, up to first order, normally distributed with the same asymptotic variance of the scoring rule estimator.

When dealing with default priors, in Section 4 reference priors for a vector parameter based on maximising α -divergences are discussed in the framework of scoring rules. We show that, for $0 \leq |\alpha| < 1$, the result is a Jeffreys-type prior that is proportional to the square root of the determinant of the Godambe information matrix.

A possible direction of further research is to extend the proposed methodology considering the class of monotone and regular divergences which is a broad family of divergences asymptotically equivalent to α -divergences; see Corcuera and Giummolè (1998).

Appendix

Proof of Theorem 3.1

Proof. The posterior (5) can equivalently be written as

$$\pi_{SR}(\theta|x) = \frac{\pi(\theta) \exp \{-S(\theta^*) + S(\tilde{\theta})\}}{\int_{\Theta} \pi(\theta) \exp \{-S(\theta^*) + S(\tilde{\theta})\} d\theta},$$

with $\theta^* = \tilde{\theta} + C(\theta - \tilde{\theta})$ and C fixed such that $C^T K(\theta) C = G(\theta)$.

Let $w = (w^1, \dots, w^d)^T = n^{1/2}(\theta - \tilde{\theta})$. Then, $\theta = \tilde{\theta} + n^{-1/2}w$ and $\theta^* = \tilde{\theta} + n^{-1/2}Cw$. A posterior for w is

$$\pi_{SR}(w|x) = \frac{b(w, x)}{\int b(w, x) dw},$$

with

$$b(w, x) = \pi(\tilde{\theta} + n^{-1/2}w) \exp \{-S(\tilde{\theta} + n^{-1/2}Cw) + S(\tilde{\theta})\}.$$

Now

$$\pi(\tilde{\theta} + n^{-1/2}w) = \tilde{\pi} \left(1 + n^{-1/2}R_1(w) + \frac{1}{2}n^{-1}R_2(w) \right) + O_p(n^{-3/2}),$$

with

$$R_1(w) = \frac{\tilde{\pi}_i}{\tilde{\pi}} w^i \quad \text{and} \quad R_2(w) = \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} w^{ij},$$

where $w^{ij\dots} = w^i w^j \dots$ is a product of components of w . Moreover,

$$\begin{aligned}
-S(\tilde{\theta} + n^{-1/2}Cw) + S(\tilde{\theta}) &= -n^{-1} \left(\frac{1}{2}(Cw)^{ij} \tilde{S}_{ij} + \frac{1}{6}n^{-1/2}(Cw)^{ijk} \tilde{S}_{ijk} \right. \\
&\quad \left. + \frac{1}{24}n^{-1}(Cw)^{ijkh} \tilde{S}_{ijkh} \right) + O_p(n^{-3/2}) \\
&= -\frac{1}{2}n^{-1}w^T \left(C^T \frac{\partial^2 \tilde{S}}{\partial \theta \partial \theta^T} C \right) w - \frac{1}{6}n^{-1/2}R_3(w) \\
&\quad - \frac{1}{24}n^{-1}R_4(w) + O_p(n^{-3/2}) \\
&= -\frac{1}{2}w^T \tilde{H}w - \frac{1}{6}n^{-1/2}R_3(w) - \frac{1}{24}n^{-1}R_4(w) + O_p(n^{-3/2}), \quad (15)
\end{aligned}$$

where

$$R_3(w) = n^{-1}(Cw)^{ijk} \tilde{S}_{ijk} \quad \text{and} \quad R_4(w) = n^{-1}(Cw)^{ijkh} \tilde{S}_{ijkh}.$$

The numerator in Bayes formula can thus be written as

$$\begin{aligned}
b(w, x) &= \tilde{\pi} \left(1 + n^{-1/2}R_1(w) + \frac{1}{2}n^{-1}R_2(w) \right) \\
&\quad \exp \left\{ -\frac{1}{2}w^T \tilde{H}w \right\} \left(1 - \frac{1}{6}n^{-1/2}R_3(w) - \frac{1}{24}n^{-1}R_4(w) + \frac{1}{72}n^{-1}R_3(w)^2 \right) + O_p(n^{-3/2}) \\
&= \tilde{\pi} \exp \left\{ -\frac{1}{2}w^T \tilde{H}w \right\} \left[1 + n^{-1/2} \left(R_1(w) - \frac{1}{6}R_3(w) \right) \right. \\
&\quad \left. + n^{-1} \left(\frac{1}{2}R_2(w) - \frac{1}{6}R_1(w)R_3(w) - \frac{1}{24}R_4(w) + \frac{1}{72}R_3(w)^2 \right) \right] + O_p(n^{-3/2}). \quad (16)
\end{aligned}$$

The denominator can be approximated using the moments of the d -variate normal distribution $N_d(0, \tilde{H}^{-1})$. We have that

$$\begin{aligned}
\int b(w, x)dw &= \tilde{\pi}(2\pi)^{d/2}|\tilde{H}|^{-1/2} \left[1 + n^{-1} \left(\frac{1}{2}E(R_2(W)) - \frac{1}{6}E(R_1(W)R_3(W)) \right. \right. \\
&\quad \left. \left. - \frac{1}{24}E(R_4(W)) + \frac{1}{72}E(R_3(W)^2) \right) \right] + o(n^{-1}), \quad (17)
\end{aligned}$$

being $E(R_1(W))$ and $E(R_3(W))$ both related to odd moments and thus equal to zero. Now,

$$E(R_2(W)) = \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} \tilde{h}^{ij}, \quad E(R_1(W)R_3(W)) = 3 \frac{\tilde{\pi}_i}{\tilde{\pi}} \frac{\tilde{S}_{jkh}}{n} c_{jr} c_{ks} c_{ht} \tilde{h}^{ir} \tilde{h}^{st},$$

$$E(R_4(W)) = 3 \frac{\tilde{S}_{ijkh}}{n} c_{ir} c_{js} c_{kt} c_{hu} \tilde{h}^{rs} \tilde{h}^{tu}$$

and

$$E(R_3(W)^2) = \frac{\tilde{S}_{ijk}\tilde{S}_{rst}}{n^2} (9c_{ia}c_{jb}c_{kc}c_{rd}c_{se}c_{tf}\tilde{h}^{ab}\tilde{h}^{cd}\tilde{h}^{ef} + 6c_{ia}c_{rb}c_{jc}c_{sd}c_{ke}c_{tf}\tilde{h}^{ab}\tilde{h}^{cd}\tilde{h}^{ef}).$$

Putting together expressions (16) and (17), we finally obtain the result with

$$A_1(w) = R_1(w) - \frac{1}{6}R_3(w)$$

and

$$A_2(w) = \frac{1}{2}(R_2(w) - E(R_2(W))) - \frac{1}{6}(R_1(w)R_3(w) - E(R_1(W)R_3(W))) - \frac{1}{24}(R_4(w) - E(R_4(W))) + \frac{1}{72}(R_3(w)^2 - E(R_3(W)^2)).$$

□

Proof of Theorem 4.1

Proof. The proof follows the same steps as in Liu *et al.* (2014), Section 3, generalized for the use of a SR-posterior (5) instead of the classic posterior distribution.

Let $\pi(\theta)$ be a prior distribution for θ and $p(x|\theta)$ be the conditional distribution of X given θ . Thus, the marginal distribution of X can be written as $p(x) = p(x|\theta)\pi(\theta)/\pi(\theta|x)$, with $\pi(\theta|x)$ the conditional distribution of θ given x . The functional (8) associated to an α -divergence between the prior $\pi(\theta)$ and the SR-posterior (5) can be written as

$$\begin{aligned} T(\pi) &= \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\mathcal{X}} \left(\int_{\Theta} \pi(\theta)^\alpha \pi_{SR}(\theta|x)^{1-\alpha} d\theta \right) p(x) dx \right] \\ &= \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\Theta} \left(\int_{\mathcal{X}} \pi(\theta)^\alpha \pi_{SR}(\theta|x)^{1-\alpha} \pi(\theta|x)^{-1} p(x|\theta) dx \right) \pi(\theta) d\theta \right] \\ &= \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\Theta} \pi(\theta)^{\alpha+1} E_\theta [\pi_{SR}(\theta|X)^{1-\alpha} \pi(\theta|X)^{-1}] d\theta \right], \end{aligned} \quad (18)$$

where $E_\theta(\cdot)$ denotes expectation with respect to the conditional distribution of X given θ . For $\alpha = 0$ or 1, we need to interpret $T(\pi)$ as its limiting value, when it exists.

Now, we apply the shrinkage argument for evaluating the conditional expectation in (18), $E_\theta [\pi_{SR}(\theta|X)^{1-\alpha} \pi(\theta|X)^{-1}]$. See Datta and Mukerjee (2004) for details on the shrinkage method.

Let us first consider the case of $0 < |\alpha| < 1$.

The first step of the shrinkage method involves fixing a suitable prior distribution $\bar{\pi}(\theta)$ and calculating the expected value of $\pi_{SR}(\theta|x)^{1-\alpha} \pi(\theta|x)^{-1}$ with respect to the corresponding posterior $\bar{\pi}(\theta|x)$, i.e. $E^{\bar{\pi}} [\pi_{SR}(\theta|x)^{1-\alpha} \pi(\theta|x)^{-1}|x]$. Notice that here we use the classic posterior

based on the prior $\bar{\pi}(\theta)$, i.e. the conditional distribution of θ given x , $\bar{\pi}(\theta|x)$. Using up to first order the asymptotic expansion (7) for $\pi_{SR}(\theta|x)$ and similar expansions for $\bar{\pi}(\theta|x)$ and $\pi(\theta|x)$, we obtain

$$\pi_{SR}(\theta|X)^{1-\alpha} = \left(\frac{n^d |\tilde{H}|}{(2\pi)^d} \right)^{(1-\alpha)/2} \exp \left\{ \frac{n(1-\alpha)}{2} (\theta - \tilde{\theta})^T \tilde{H} (\theta - \tilde{\theta}) \right\} + O_p(n^{-1/2})$$

and

$$\pi(\theta|X)^{-1} \bar{\pi}(\theta|X) = 1 + O_p(n^{-1/2}).$$

Since \tilde{H} tends to G as $n \rightarrow \infty$, we have that

$$\begin{aligned} E^{\bar{\pi}} [\pi_{SR}(\theta|x)^{1-\alpha} \pi(\theta|x)^{-1} | x] &= \left(\frac{n^d |\tilde{H}|}{(2\pi)^d} \right)^{(1-\alpha)/2} \int_{\Theta} \exp \left\{ \frac{n(1-\alpha)}{2} (\theta - \tilde{\theta})^T \tilde{H} (\theta - \tilde{\theta}) \right\} d\theta \\ &\quad + O(n^{-1/2}) \\ &= \left(\frac{n^d |G(\theta)|}{(2\pi)^d} \right)^{-\alpha/2} (1-\alpha)^{-d/2} + O(n^{-1/2}). \end{aligned}$$

The second step of the shrinkage argument requires to integrate again with respect to the distribution of X given θ and with respect to the prior $\bar{\pi}(\theta)$, thus obtaining

$$\begin{aligned} &\int_{\Theta} E_{\theta} [E^{\bar{\pi}} [\pi_{SR}(\theta|X)^{1-\alpha} \pi(\theta|X)^{-1} | X]] \bar{\pi}(\theta) d\theta \\ &= \int_{\Theta} \left(\frac{n^d |G(\theta)|}{(2\pi)^d} \right)^{-\alpha/2} (1-\alpha)^{-d/2} \bar{\pi}(\theta) d\theta + O(n^{-1/2}). \end{aligned}$$

Finally, by letting the prior $\bar{\pi}(\theta)$ go to θ , we obtain

$$E_{\theta} [\pi_{SR}(\theta|X)^{1-\alpha} \pi(\theta|X)^{-1}] = \left(\frac{n}{2\pi} \right)^{-d\alpha/2} |G(\theta)|^{-\alpha/2} (1-\alpha)^{-d/2} + O(n^{-1/2}). \quad (19)$$

For $0 < |\alpha| < 1$, by substituting (19) in (18) we can see that the selection of a prior $\pi(\theta)$ corresponds to the minimization with respect to $\pi(\theta)$ of the functional

$$\frac{1}{\alpha(1-\alpha)} \int_{\Theta} \pi(\theta)^{\alpha+1} |G(\theta)|^{-\alpha/2} d\theta. \quad (20)$$

Notice that the preceding expression can be interpreted as an increasing transformation of the $(-\alpha)$ -divergence between a density that is proportional to $|G(\theta)|^{1/2}$ and the prior $\pi(\theta)$. Thus, it is minimized if and only if the two densities coincide.

This result cannot be extended to the case of $\alpha \geq 1$, as is evident from the right-hand-side of (19). For $\alpha < -1$, (20) turns out to be equivalent to a decreasing transformation of the $(-\alpha)$ -divergence between a density that is proportional to $|G(\theta)|^{1/2}$ and the prior $\pi(\theta)$. Thus a maximizer for the expected α -divergence in this case does not exist.

For $\alpha \rightarrow 0$, following the same steps as for $0 < |\alpha| < 1$, it can be easily shown that maximization of the average Kullback-Leibler divergence is asymptotically equivalent to minimization of

$$\int_{\Theta} \pi(\theta) \log \left(\frac{\pi(\theta)}{|G(\theta)|^{1/2}} \right) d\theta,$$

which is attained by choosing again a Jeffreys-type prior proportional to $|G(\theta)|^{1/2}$. Indeed, the preceding expression is equal up to an additive constant to Kullback-Leibler divergence between a density that is proportional to $|G(\theta)|^{1/2}$ and the prior $\pi(\theta)$.

The case $\alpha = -1$ corresponds to the Chi-square divergence. For this case the proof requires higher order terms in the expansion of the scoring rule posterior distribution and also of both the conditional distributions of θ given x calculated with respect to the priors $\pi(\theta)$ and $\bar{\pi}(\theta)$. Thus, we use (7) up to order $O_p(n^{-3/2})$. In particular, let $w = n^{1/2}(\theta - \tilde{\theta})$ and $w_1 = n^{1/2}(\theta - \hat{\theta})$, where $\tilde{\theta}$ is the minimum scoring rule estimator and $\hat{\theta}$ the maximum likelihood estimator for θ . The different posterior distributions for w and w_1 given x can be written as

$$\pi_{SR}(w|x) = \phi_d(w; \tilde{H}^{-1}) [1 + n^{-1/2}A_1(w) + n^{-1}A_2(w)] + O_p(n^{-3/2}),$$

$$\pi(w_1|x) = \phi_d(w_1; (\hat{J}^\ell)^{-1}) [1 + n^{-1/2}A_1^\ell(w_1) + n^{-1}A_2^\ell(w_1)] + O_p(n^{-3/2}),$$

$$\bar{\pi}(w_1|x) = \phi_d(w_1; (\hat{J}^\ell)^{-1}) [1 + n^{-1/2}\bar{A}_1^\ell(w_1) + n^{-1}\bar{A}_2^\ell(w_1)] + O_p(n^{-3/2}),$$

where A_1 and A_2 are defined as in Theorem 3.1, $J^\ell(\theta) = (-\partial^2 \ell / \partial \theta \partial \theta^T) / n$ is the observed information and $A_1^\ell, \bar{A}_1^\ell, A_2^\ell, \bar{A}_2^\ell$ are obtained by calculating A_1 and A_2 with the logarithmic score as scoring rule and π and $\bar{\pi}$ as priors; for instance,

$$A_1^\ell(w_1) = \frac{\hat{\pi}_i}{\hat{\pi}} w_1^i + \frac{1}{6} \frac{\hat{\ell}_{ijk}}{n} w_1^i w_1^j w_1^k, \quad \text{and} \quad \bar{A}_1^\ell(w_1) = \frac{\hat{\pi}_i}{\hat{\pi}} w_1^i + \frac{1}{6} \frac{\hat{\ell}_{ijk}}{n} w_1^i w_1^j w_1^k,$$

with ℓ the log-likelihood and ℓ_{ijk} its third order derivatives. As usual, a tilde or a hat over a quantity indicate that it is calculated at $\tilde{\theta}$ and $\hat{\theta}$ respectively.

Following the shrinkage argument, we need first to evaluate

$$E^{\bar{\pi}} [\pi_{SR}(\theta|x)^2 \pi(\theta|x)^{-1} | x] = n^{d/2} \int \pi_{SR}(w|x)^2 \pi(w|x)^{-1} \bar{\pi}(w|x) dw.$$

For doing this, we use the fact that

$$\begin{aligned} \pi_{SR}(w|x)^2 &= (2\pi)^{-d/2} |\tilde{H}|^{1/2} 2^{-d/2} \phi_d(w; \tilde{H}^{-1}/2) \\ &\quad [1 + n^{-1/2} 2A_1(w) + n^{-1}(2A_2(w) + A_1(w)^2)] + O_p(n^{-3/2}) \end{aligned}$$

and

$$\begin{aligned} \pi(w_1|x)^{-1} \bar{\pi}(w_1|x) &= 1 + n^{-1/2}(\bar{A}_1^\ell(w_1) - A_1^\ell(w_1)) \\ &\quad + n^{-1}(A_1^\ell(w_1)^2 + \bar{A}_2^\ell(w_1) - A_2^\ell(w_1) - \bar{A}_1^\ell(w_1)A_1^\ell(w_1)) + O_p(n^{-3/2}) \end{aligned}$$

and we calculate the last expression in $w_1 = w + n^{1/2}(\tilde{\theta} - \hat{\theta})$.

After tedious calculations, we obtain

$$\begin{aligned}
& E^{\tilde{\pi}} [\pi_{SR}(\theta|x)^2 \pi(\theta|x)^{-1} |x] = \\
& = (2\pi)^{-d/2} \left(\frac{n}{2}\right)^{d/2} |\tilde{H}|^{1/2} \left\{ 1 + \frac{1}{n^{1/2}} \left(\frac{\hat{\pi}_i}{\hat{\pi}} - \frac{\tilde{\pi}_i}{\tilde{\pi}} \right) n^{1/2} (\tilde{\theta} - \hat{\theta})^i \right. \\
& \quad + \frac{1}{2n} \left[\left(-\frac{\hat{\pi}_{ij}}{\hat{\pi}} + \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} \right) \hat{\mathbf{j}}^{ij} \right. \\
& \quad + \left(-\frac{\tilde{\pi}_{ij}}{\tilde{\pi}} + \frac{\hat{\pi}_i \tilde{\pi}_j}{\tilde{\pi} \tilde{\pi}} + 2 \frac{\tilde{\pi}_i \hat{\pi}_j}{\tilde{\pi} \hat{\pi}} - 2 \frac{\tilde{\pi}_i \tilde{\pi}_j}{\tilde{\pi} \hat{\pi}} + \frac{\hat{\pi}_i \hat{\pi}_j}{\hat{\pi} \hat{\pi}} - \frac{\hat{\pi}_i \tilde{\pi}_j}{\hat{\pi} \tilde{\pi}} + \frac{1}{2} \frac{\hat{\pi}_{ij}}{\hat{\pi}} - \frac{1}{2} \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} \right) \tilde{h}^{ij} \\
& \quad + \left(\frac{\tilde{\pi}_i}{\tilde{\pi}} - \frac{1}{2} \frac{\hat{\pi}_i}{\hat{\pi}} + \frac{1}{2} \frac{\tilde{\pi}_i}{\tilde{\pi}} \right) \frac{\tilde{S}_{khl}}{n} c_{kj} c_{hs} c_{lt} \tilde{h}^{st} \tilde{h}^{ij} + \left(-\frac{\hat{\pi}_i}{\hat{\pi}} + \frac{\tilde{\pi}_i}{\tilde{\pi}} \right) \frac{\tilde{\ell}_{jhk}}{n} \hat{\mathbf{j}}^{hk} \hat{\mathbf{j}}^{ij} \\
& \quad \left. \left. + \left(-\frac{\hat{\pi}_{ij}}{\hat{\pi}} + \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} + 2 \frac{\hat{\pi}_i \hat{\pi}_j}{\hat{\pi} \hat{\pi}} - 2 \frac{\hat{\pi}_i \tilde{\pi}_j}{\hat{\pi} \tilde{\pi}} \right) n (\tilde{\theta} - \hat{\theta})^{ij} \right] + R \right\} + O(n^{-1}),
\end{aligned}$$

where $\hat{\mathbf{j}}^{ij}$ are the components of $(\hat{J}^\ell)^{-1}$ and $R = R(\tilde{\theta}, \hat{\theta})$ is a function that does not involve the prior nor its derivatives.

Now, let $E_\theta(S_{ijk}/n) = B_{ijk}^S(\theta) + o(n^{-1/2})$, $E_\theta(\ell_{ijk}/n) = B_{ijk}^\ell(\theta) + o(n^{-1/2})$ and recall that $E_\theta(\tilde{H}) = G(\theta) + o(n^{-1/2})$ and $E_\theta(\hat{J}^\ell) = I(\theta) + o(n^{-1/2})$, where I is the expected information matrix. Moreover, note that $E_\theta[(\tilde{\theta} - \hat{\theta})^i] = o(n^{-1/2})$ and $E_\theta[n(\tilde{\theta} - \hat{\theta})^{ij}] = g^{ij} + \mathbf{i}^{ij} - 2\sigma^{ij} + o(n^{-1/2})$, where g_{ij} and g^{ij} are the components of G and G^{-1} respectively, \mathbf{i}^{ij} are the components of I^{-1} and $\sigma^{ij} = n \text{Cov}_\theta(\tilde{\theta}, \hat{\theta})^{ij}$. Furthermore, let $a_j^S = B_{khl}^S c_{kj} c_{hs} c_{lt} g^{st}$, $a_j^\ell = B_{jhk}^\ell \mathbf{i}^{hk}$. By integrating again the above expression with respect to the distribution of X given θ we get

$$\begin{aligned}
& E_\theta [E^{\tilde{\pi}} [\pi_{SR}(\theta|X)^2 \pi(\theta|X)^{-1} |X]] = \\
& = (2\pi)^{-d/2} \left(\frac{n}{2}\right)^{d/2} |G|^{1/2} \left\{ 1 + \frac{1}{2n} \left[\left(-\frac{\tilde{\pi}_{ij}}{\tilde{\pi}} + \frac{\pi_{ij}}{\pi} \right) \mathbf{i}^{ij} \right. \right. \\
& \quad + \left(-\frac{3}{2} \frac{\pi_{ij}}{\pi} + \frac{\pi_i \tilde{\pi}_j}{\pi \tilde{\pi}} + \frac{1}{2} \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} \right) g^{ij} \\
& \quad + \left(\frac{3}{2} \frac{\pi_i}{\pi} - \frac{1}{2} \frac{\tilde{\pi}_i}{\tilde{\pi}} \right) a_j^S g^{ij} - \left(\frac{\tilde{\pi}_i}{\tilde{\pi}} - \frac{\pi_i}{\pi} \right) a_j^\ell \mathbf{i}^{ij} \\
& \quad \left. \left. + \left(-\frac{\pi_{ij}}{\pi} + \frac{\tilde{\pi}_{ij}}{\tilde{\pi}} + 2 \frac{\pi_i \pi_j}{\pi \pi} - 2 \frac{\tilde{\pi}_i \tilde{\pi}_j}{\tilde{\pi} \tilde{\pi}} \right) (g^{ij} + \mathbf{i}^{ij} - 2\sigma^{ij}) \right] + F \right\} + O(n^{-1}),
\end{aligned}$$

where $F = F(\theta)$ is a function that does not involve the prior nor its derivatives.

By integrating again with respect to the prior $\tilde{\pi}(\theta)$ and by letting the prior go to θ , we

finally obtain the expected value needed in (18) for $\alpha = -1$:

$$\begin{aligned}
E_\theta [\pi_{SR}(\theta|X)^2 \pi(\theta|X)^{-1}] &= \\
&= (2\pi)^{-d/2} \left(\frac{n}{2}\right)^{d/2} |G|^{1/2} \left\{ 1 + \frac{1}{4n} \left[(-3g^{ij} + 4\mathbf{i}^{ij} - 4\sigma^{ij}) \frac{\pi_{ij}}{\pi} \right. \right. \\
&\quad \left. \left. + \left(3a_j^S g^{ij} + 2a_j^\ell \mathbf{i}^{ij} + |G|^{-1} \frac{\partial |G|}{\partial \theta^j} (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij}) + 2 \frac{\partial (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij})}{\partial \theta^j} \right) \frac{\pi_i}{\pi} \right. \right. \\
&\quad \left. \left. + 2g^{ij} \frac{\pi_i \pi_j}{\pi} + M \right] \right\} + O(n^{-1}), \tag{21}
\end{aligned}$$

where $M = M(\theta)$ is a function that does not involve the prior nor its derivatives.

Substituting (21) in (18) with $\alpha = -1$, we find that maximization of the average Chi-square divergence is equivalent to maximization of

$$\begin{aligned}
&\int_{\Theta} |G|^{1/2} \left[(-3g^{ij} + 4\mathbf{i}^{ij} - 4\sigma^{ij}) \frac{\pi_{ij}}{\pi} + 2g^{ij} \frac{\pi_i \pi_j}{\pi} \right. \\
&\quad \left. + \left(3a_j^S g^{ij} + 2a_j^\ell \mathbf{i}^{ij} + |G|^{-1} \frac{\partial |G|}{\partial \theta^j} (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij}) + 2 \frac{\partial (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij})}{\partial \theta^j} \right) \frac{\pi_i}{\pi} \right] d\theta.
\end{aligned}$$

Let $y = y(\theta) = (y_1(\theta), \dots, y_d(\theta))^T$, with $y_i(\theta) = \pi_i(\theta)/\pi(\theta)$, $i = 1, \dots, d$, and let y' be the matrix of partial derivatives of y with respect to θ , i.e. $y' = (y_{ij})_{ij}$, $y_{ij} = \partial y_i / \partial \theta^j$, $i, j = 1, \dots, d$. Then, $\pi_{ij}/\pi = y_{ij} + y_i y_j$ and the quantity to be maximized can be rewritten as

$$\begin{aligned}
&\int_{\Theta} |G|^{1/2} \left[(-3g^{ij} + 4\mathbf{i}^{ij} - 4\sigma^{ij}) y_{ij} + (-g^{ij} + 4\mathbf{i}^{ij} - 4\sigma^{ij}) y_i y_j \right. \\
&\quad \left. + \left(3a_j^S g^{ij} + 2a_j^\ell \mathbf{i}^{ij} + |G|^{-1} \frac{\partial |G|}{\partial \theta^j} (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij}) + 2 \frac{\partial (g^{ij} + 2\mathbf{i}^{ij} - 4\sigma^{ij})}{\partial \theta^j} \right) y_i \right] d\theta.
\end{aligned}$$

Let us denote the integrand function by $U(\theta, y, y')$. The solution to the maximization problem is found by solving the system of Euler-Lagrange equations:

$$\frac{\partial U}{\partial y_i} - \sum_{j=1}^d \frac{\partial}{\partial \theta^j} \left(\frac{\partial U}{\partial y_{ij}} \right) = 0, \quad i = 1, \dots, d.$$

After some calculations, we obtain the solution to the variational problem as

$$y(\theta) = \frac{1}{4} \left[6a^S G^{-1} + 4a^\ell I^{-1} + |G|^{-1} \frac{\partial |G|}{\partial \theta} (5G^{-1} - 4\Sigma) + 2G \nabla \cdot (5G^{-1} - 4\Sigma) \right] \Gamma,$$

where $a^S = (a_1^S, \dots, a_d^S)^T$, $a^\ell = (a_1^\ell, \dots, a_d^\ell)^T$, $\Sigma = (\sigma^{ij})_{ij}$, $\Gamma = (G^{-1} - 4I^{-1} + 4\Sigma)^{-1}$, $\nabla \cdot G^{-1} = (\partial g^{1j} / \partial \theta^j, \dots, \partial g^{dj} / \partial \theta^j)^T$ and $\nabla \cdot \Sigma = (\partial \sigma^{1j} / \partial \theta^j, \dots, \partial \sigma^{dj} / \partial \theta^j)^T$. \square

References

- Barndorff-Nielsen, O.E. (1976). Plausibility inference. *J. Roy. Statist. Soc. B*, **38**, 103–131.
- 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.
- Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 1–17.
- Berger, J.O., Bernardo, J.M. (1992). On the development of reference priors (with discussion). In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 35–60. Oxford Univ. Press.
- Berger, J.O., Bernardo, J.M., Sun, D. (2009). The formal definition of reference priors. *Journal of the American Statistical Association*, **107**, 636–648.
- Berger, J.O., Bernardo, J.M., Sun, D. (2012). Overall objective priors. *Bayesian Analysis*, **10**, 189–221.
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B*, **41**, 113–147.
- Bernardo, J.M. (2005). Reference analysis. In: *Handbook of Bayesian Statistics*, **25** (D.K. Dey and C.R. Rao, eds.), North-Holland, Amsterdam: Elsevier, 17–90.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78**, 1–3.
- Chandler, R.E., Bate, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika*, **94**, 167–183.
- Chang, I., Mukerjee, R. (2008). Bayesian and frequentist confidence intervals arising from empirical-type likelihoods. *Biometrika*, **95**, 139–147.
- Corcuera, J.M., Giummolè, F. (1998). A characterization of monotone and regular divergences. *Ann. Inst. Statist. Math.*, **50**, 433–450.
- Datta, G.S., Mukerjee R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Lecture Notes in Statistics 178. Springer.
- Dawid, A.P. (1986). Probability forecasting. In: *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, and C. B. Read eds.), 210–218. Wiley-Interscience.
- Dawid, A.P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math.*, **59**, 77–93.

- Dawid, A.P., Lauritzen, S.L. (2005). The geometry of decision theory. In: *Proceedings of the Second International Symposium on Information Geometry and its Applications*, 22–28. University of Tokyo.
- Dawid, A.P., Musio, M. (2014). Theory and Applications of Proper Scoring Rules. *Metron*, **72**, 169–183.
- Dawid, A.P., Musio, M. (2015). Bayesian model selection based on proper scoring rules (with discussion). *Bayesian Analysis*, **10**, 479–521.
- Dawid, A.P., Musio, M., Ventura, L. (2016). Minimum scoring rule inference. *Scand. J. Statist.*, **43**, 123–138.
- Farcomeni, A., Ventura, L. (2012). An overview of robust methods in medical research. *Statistical Methods in Medical Research*, **21**, 111–133.
- Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1211.
- Good, I.J. (1952). Rational decisions. *J. Roy. Statist. Soc. B*, **14**, 107–114.
- Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. *Stat. Sci.*, **26**, 187–202.
- Ghosh, M., Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electr. J. Statist.*, **7**, 2420–2456.
- Ghosh, M., Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Ann. Inst. Stat. Math.*, **68**, 413–437.
- Ghosh, M., Mergel, V., Liu, R. (2011). A general divergence criterion for prior selection *Ann. Inst. Stat. Math.* 43–58.
- Ghosh, J.K., Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multi-parameter case. *J. Mult. Anal.*, **38**, 385–393.
- Greco, L., Racugno, W., Ventura L. (2008). Robust likelihood functions in Bayesian inference. *J. Statist. Plann. Inf.*, **138**, 1258–1270.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.P. (2009). *Robust methods in biostatistics*. Wiley, Chichester, UK.

- Huber, P.J., Ronchetti, E.M. (2009). *Robust Statistics*. Wiley, New York.
- Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.
- Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, New York.
- Lazar, N.A. (2003). Bayes empirical likelihood. *Biometrika*, **90**, 319–326.
- Leisen, F., Villa, C., Walker, S.G. (2017). On a global objective prior from score rules. arXiv: 1706.00599v1
- Lin, L. (2006). Quasi Bayesian likelihood. *Statistical Methodology*, **3**, 444–455.
- Liu, R., Chakrabarti, A., Samanta, T., Ghosh, J.K., Ghosh, M. (2014). On divergence measures leading to Jeffreys and other reference priors. *Bayesian Analysis*, **9**, 331–370.
- Machete, R. (2013). Contrasting probabilistic scoring rules. *J. Statist. Plann. Inf.*, **143**, 1781–1790.
- Mameli, V., Musio, M., Ventura, L. (2017). Bootstrap adjustments of signed scoring rule root statistics. *Comm. Statist. - Simul. Comput.*, to appear.
- Mameli, V., Ventura, L. (2015). Higher-order asymptotics for scoring rules. *J. Statist. Plann. Inf.*, **165**, 13–26.
- Mardia, K.V., Jupp, P.E. (2000). *Directional statistics*. Wiley, London.
- Mardia, K.V., Kent, J.T., Laha, A.K. (2016). Score matching estimators for directional distributions. ArXiv:1604.08470v1.
- Musio, M., Mameli, V., Ruli, E., Ventura, L. (2017). Bayesian Inference for directional data through ABC and homogeneous proper scoring rules. *Proceeding of 61st ISI World Statistics Congress*, to appear.
- Pace, L., Salvan, A., Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statist. Sin.*, **21**, 129–148.
- Parry, M., Dawid, A.P., Lauritzen, S.L. (2012). Proper local scoring rules. *Ann. Statist.*, **40**, 561–592.
- Pauli, F., Racugno, W., Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statist. Sin.*, **21**, 149–164.

- Riani, M., Atkinson, A.C., Perrotta, D. (2014). A parametric framework for the comparison of methods of very robust regression. *Statistical Science*, **29**, 128–143.
- Ribatet, M., Cooley, D. and Davison, A.C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sin.*, **22**, 813–845.
- Ruli, E., Sartori, N., Ventura, L. (2016). Approximate Bayesian Computation with composite score functions. *Statist. Comp.*, **26**, 679–692.
- Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *The Annals of Statistics*, **35**, 634–672.
- Smith, E.L., Stephenson, A.G. (2009). An extended Gaussian max-stable process model for spatial extremes. *J. Statist. Plann. Inf.*, **139**, 1266–1275.
- Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Physics*, **52**, 479–487.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sin.*, **21**, 5–42.
- Ventura, L., Cabras, S., Racugno, W. (2010). Default prior distributions from quasi- and quasi-profile likelihoods. *J. Statist. Plann. Inf.*, **140**, 2937–2942.
- Ventura, L., Racugno, W. (2016). Pseudo-likelihoods for Bayesian inference. In: *Topics on Methodological and Applied Statistical Inference, Series Studies in Theoretical and Applied Statistics*, Springer-Verlag, 205–220.
- Walker, S.G. (2016). Bayesian information in an experiment and the Fisher information distance. *Statistics and Probability Letters*, **112**, 5–9.
- Yang, Y., He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, **40**, 1102–1131.