

CONNECTIONS BETWEEN REFERENCE PRIOR THEORY AND GLOBAL SENSITIVITY ANALYSIS, AN ILLUSTRATION WITH f -DIVERGENCES.

Antoine Van Biesbroeck^{1,2}, Clément Gauchy³, Josselin Garnier¹, & Cyril Feau²

¹ *CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France*

antoine.van-biesbroeck@polytechnique.edu, josselin.garnier@polytechnique.edu

² *Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191 Gif-sur-Yvette, France*

cyril.feau@cea.fr

³ *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France*

clement.gauchy@cea.fr

Résumé. La théorie des priors de référence construit un cadre de réponse à la question du choix du prior en analyse Bayésienne, via l'introduction d'un critère d'information dite objective. Dans l'idée d'une définition plus globale d'un tel critère, nous mettons en avant un lien entre la définition de celui-ci et des outils d'analyse de sensibilité, ouvrant la voie à de nouvelles classes de mesures que nous jugeons utiles et interprétables pour le choix d'un prior de référence. Dans le cadre que nous étudions, nous étayons la force du prior de Jeffreys comme optimum de nos nouvelles métriques d'information objective. Celles-ci sont introduites rigoureusement et accompagnées d'une preuve qui démontre le caractère optimal du prior de Jeffreys sous les hypothèses adéquates.

Mots-clés. Prior de référence, prior de Jeffreys, information mutuelle, indices de sensibilité

Abstract. The reference prior theory constructs a framework which helps the resolution of the prior choice issue in Bayesian analysis, relying on the introduction of some information criteria called objective. Within the idea of a more global definition of such criteria, we discuss a connection between their definition and some sensitivity analysis tools, which opens the way to numerous measures for the reference prior choice. In our framework, we demonstrate the robustness of the Jeffreys prior as the optimum of our new objective information metrics. Those are rigorously introduced and accompanied with a proof of the optimal characteristic of the Jeffreys prior under appropriate assumptions.

Keywords. Reference prior, Jeffreys prior, mutual information, sensitivity indices

1 Introduction

Reference prior theory in Bayesian analysis uses information-theoretical tools to choose a prior distribution that limits the influence of the prior distribution on the posterior distribution. The original definition of reference priors date back to the seminal paper of [Bernardo \(1979\)](#) and was largely informal. A more rigorous definition was proposed in [Berger et al. \(2009\)](#). Reference priors are usually defined as the prior that maximizes the mutual information between the parameters and the data. Under asymptotic posterior normality, [Clarke and Barron \(1994\)](#) established regularity conditions such that the reference prior is the well-known Jeffreys prior.

Some generalizations of the mutual information metric have been proposed in the literature, and mostly consist into changes within the measure tool for the influence between the prior and the posterior distributions [Chen et al. \(2010\)](#); [Liu et al. \(2014\)](#). Those works highlight the role of the Jeffreys prior as the optimum for so-called objective information.

In this paper, we unveil a direct connection between the reference prior theory and global sensitivity analysis [Da Veiga \(2015\)](#), this will enable to enlarge the notion of reference prior for novel mutual information criteria and offer a clearer interpretation of the reference prior. From a theoretical perspective, we show that the reference prior for a mutual information defined using f -divergences [Csiszár \(1967\)](#) is also the Jeffreys prior.

The article is organized as follows. In [Section 2](#), we formulate a standard framework for Bayesian analysis and we remind the definition of the mutual information. Then, the connection with global sensitivity analysis is stated and discussed in [Section 3](#). From this connection, we suggest a definition for generalized mutual information criteria in [Section 4](#), and we formulate a theorem that determines the reference prior for these criteria in [Section 5](#) followed by a proof in [Section 6](#) under appropriate assumptions. Finally, a discussion about the perspectives and the applications of our work is given in [Section 8](#).

2 Bayesian framework and notations

Let $\mathcal{Y}$ be an observation space over which $\mathcal{Y}$ is a σ -algebra. A statistical model is then characterized by a collection of parameterized probability measures $(\mathbb{P}_{Y|T=\theta})_{\theta \in \Theta}$ on $\mathcal{Y}$. $\mathbf{Y} = (Y_1, \dots, Y_k)$ denotes the observed items, it is a random variable taking values in $\mathcal{Y}^k$. In the Bayesian framework, T is a random variable taking values in $(\Theta, \mathcal{T})$ which follows a prior distribution π , and the distribution of $\mathbf{Y}$ conditionally to T is $\mathbb{P}_{\mathbf{Y}|T} = \mathbb{P}_{Y|T}^{\otimes k}$. From this we can express the posterior distribution $\mathbb{P}_{T|\mathbf{Y}}$ and the marginal distribution $\mathbb{P}_{\mathbf{Y}}$ as:

$$\begin{aligned} \forall A \in \mathcal{Y}^{\otimes k}, \mathbb{P}_{\mathbf{Y}}(A) &= \int_{\Theta} \mathbb{P}_{Y|T=\theta}^{\otimes k}(A) d\pi(\theta), \\ \forall B \in \mathcal{T}, A \in \mathcal{Y}^{\otimes k}, \int_A \mathbb{P}_{T|\mathbf{Y}=\mathbf{y}}(B) d\mathbb{P}_{\mathbf{Y}}(\mathbf{y}) &= \int_B \mathbb{P}_{Y|T=\theta}^{\otimes k}(A) d\pi(\theta). \end{aligned}$$

Given the number k of observed realizations, the mutual information issued by the prior

π is then defined as:

$$I(\pi|k) = \int_{\mathcal{Y}^k} KL(\mathbb{P}_{T|\mathbf{Y}}||\pi)d\mathbb{P}_{\mathbf{Y}}(\mathbf{y}) \quad (1)$$

with $\mathbb{P}_{T|\mathbf{Y}}$ being a short notation for $\mathbb{P}_{T|\mathbf{Y}=\mathbf{y}}$, KL denotes the Kullback-Leibler divergence.

Suppose now that the model admits a likelihood: there exists a reference measure μ over $\mathcal{Y}$ and density functions $(\ell(\cdot|\theta))_{\theta \in \Theta}$ such that

$$\forall A \in \mathcal{Y}, \forall \theta \in \Theta, \mathbb{P}_{Y|\theta}(A) = \int_A \ell(y|\theta)d\mu(y),$$

($\mathbb{P}_{Y|\theta}$ being a short notation for $\mathbb{P}_{Y|T=\theta}$). Thus $\mathbb{P}_{\mathbf{Y}}$ and $\mathbb{P}_{T|\mathbf{Y}}$ respectively admit the densities $p_{\mathbf{Y}}$ w.r.t. $\mu^{\otimes k}$ and $p(\cdot|\mathbf{y})$ w.r.t. π defined as

$$\begin{aligned} \forall \mathbf{y} \in \mathcal{Y}^k, p_{\mathbf{Y}}(\mathbf{y}) &= \int_{\Theta} \prod_{i=1}^k \ell(y_i|\theta) d\pi(\theta) = \int_{\Theta} \ell_k(\mathbf{y}|\theta) d\pi(\theta), \\ \forall \mathbf{y} \in \mathcal{Y}^k, \forall \theta \in \Theta, p(\theta|\mathbf{y}) &= \frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})}, \end{aligned}$$

so that $\forall \mathbf{y} \in \mathcal{Y}^k, \forall B \in \mathcal{T}, \mathbb{P}_{T|\mathbf{Y}}(B) = \int_B p(\theta|\mathbf{y}) d\pi(\theta)$. This way, while the integral in equation (1) is finite, the two following ones are finite as well and the equality holds:

$$I(\pi|k) = \int_{\mathcal{Y}^k} KL(\mathbb{P}_{T|\mathbf{Y}}||\pi)d\mathbb{P}_{\mathbf{Y}}(\mathbf{y}) = \int_{\Theta} KL(\mathbb{P}_{\mathbf{Y}|\theta}||\mathbb{P}_{\mathbf{Y}})d\pi(\theta),$$

with $\mathbb{P}_{\mathbf{Y}|\theta}$ denoting shortly $\mathbb{P}_{\mathbf{Y}|T=\theta}$. The right-hand side of that equation could also be written as an expectation as follows:

$$I(\pi|k) = \mathbb{E}_{T \sim \pi}[KL(\mathbb{P}_{\mathbf{Y}|T}||\mathbb{P}_{\mathbf{Y}})]. \quad (2)$$

Equation (2) shows that the mutual information $I(\pi|k)$ can be interpreted as a global sensitivity index in the sense of [Da Veiga \(2015\)](#). We thus propose in Section 3 a brief review of global sensitivity analysis.

3 Global sensitivity analysis, a brief review

Since the early work of Sobol' (see [Sobol' \(1993\)](#)), Global Sensitivity Analysis (GSA) has been thoroughly studied by the computer experiment community. Its main objective is to measure the impact of the input parameters uncertainty of a system (deterministic or stochastic) onto its output uncertainty by taking into account the overall uncertainty ranges of the input parameters. More formally, denote by $Y = \eta(X_1, \dots, X_p)$ the system studied which is considered as a function of the p input random variables $(X_i)_{1 \leq i \leq p}$ where $\eta : \mathbb{R}^p \mapsto \mathbb{R}$ is supposed continuous. In a classical GSA setting, the X_i 's have a known probability distribution and are mutually independent. A general class of global sensitivity indices is proposed in [Da Veiga \(2015\)](#) to measure the impact of X_i to Y :

$$S_i = \mathbb{E}_{X_i}[D(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})], \quad (3)$$

where D is a dissimilarity measure between two probability distributions. This definition covers a wide range of well-known quantities for sensitivity analysis.

A sub-class of them is considered in this work: the f -divergence d_f [Csiszár \(1967\)](#) is expressed as

$$d_f(\mathbb{P}_Y || \mathbb{P}_{Y|X_i}) = \int_{\mathcal{Y}} f\left(\frac{p_Y(y)}{p_{Y|X_i}(y)}\right) p_{Y|X_i}(y) dy,$$

where f is usually a convex function, and p_Y , $p_{Y|X_i}$ respectively denote the probability distributions of Y and of $Y|X_i$. One would notice that setting $f = -\log$ leads to the Kullback-Leibler divergence from $\mathbb{P}_Y$ to $\mathbb{P}_{Y|X_i}$.

Consider now a stochastic system where the output is the random vector $\mathbf{Y} = (Y_1, \dots, Y_k)$ and the input parameters are gathered in the random variable T . Then, the mutual information defined in equation (2) is a global sensitivity index in the sense of [Da Veiga \(2015\)](#), with the Kullback-Leibler divergence as dissimilarity measure and that quantifies the impact of T on the random vector $\mathbf{Y}$ of k observations distributed w.r.t. $\mathbb{P}_{\mathbf{Y}}$.

4 A generalized mutual information

Following the similarity between equation (2) for the mutual information and the similarity indices of equation (3), in particular under the consideration of f -divergence as dissimilarity measures as we discussed in previous section, we propose a new range of mutual information measures.

Definition 1 (Generalized mutual information). Consider the Bayesian framework introduced in section 2. Consider a dissimilarity measure D . The D -mutual information of a prior π under k observations is

$$I_D(\pi|k) = \mathbb{E}_{T \sim \pi}[D(\mathbb{P}_{\mathbf{Y}} || \mathbb{P}_{\mathbf{Y}|T})].$$

Definition 2 (f -divergence mutual information). Denote $\text{inv} : x \mapsto 1/x$. Suppose $f \circ \text{inv}$ is a concave function. The f -divergence-mutual information of π is defined as

$$I_{d_f}(\pi|k) = \mathbb{E}_{T \sim \pi}[d_f(\mathbb{P}_{\mathbf{Y}} || \mathbb{P}_{\mathbf{Y}|T})] = \int_{\Theta} \int_{\mathcal{Y}^k} f\left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)}\right) \ell_k(\mathbf{y}|\theta) d\mu^{\otimes k}(\mathbf{y}) d\pi(\theta).$$

This proposition for generalization of mutual information interprets the reference prior as the prior which maximizes its sensitivity to the data. Written as in equation (2) or more globally as in the previous definition, a recursive optimization is possible to find numerically the reference prior [Nalisnick and Smyth \(2017\)](#). An appropriate choice of the dissimilarity measure D in this criterion could lead to an enhancement of the convergence of such methods, or to an easier approximation of the mutual information for a better computational complexity, taking benefit of the numerous estimators for the different sensitivity indices that exist in the literature [Da Veiga \(2015\)](#). [Clarke and Barron \(1994\)](#) demonstrated an asymptotic maximization of the standard mutual information by the Jeffreys prior (i.e. with $f = -\log$). In this work, our target is to develop a similar result for a general f -divergence mutual information.

5 Main result

In this section we formulate a theorem which expresses an analogous reference prior under our metric from the classical definition of [Berger et al. \(2009\)](#). For this purpose, we work under the regularity assumptions of [Clarke and Barron \(1994\)](#), notably, the Fisher information matrix is continuous and positive definite as well as the Hessian matrix of the negative log likelihood. We suppose Θ is a compact subset of $\mathbb{R}^d$. Also, an adaptation of ([Clarke and Barron, 1994](#), Condition 1) is required in our proof:

Assumption 1. There exist $\xi > 0$ and $\delta > 0$ such that the function

$$\int_{\mathcal{Y}} \exp \left(\xi \sup_{\|\tilde{\theta} - \theta\| < \delta} \|\nabla_{\tilde{\theta}}^2 \log \ell(y|\tilde{\theta})\| \right) \ell(y|\theta) d\mu(y)$$

is continuous w.r.t. θ . The norm $\|\cdot\|$ denotes the Euclidean norm when applied to a vector in $\mathbb{R}^d$ and the associated operator norm when applied to a matrix in $\mathbb{R}^{d \times d}$ (i.e. the largest singular value of the matrix).

This way, under appropriate assumptions about the function f , the following theorem can be seen as a definition of a reference prior for the f -divergence mutual information.

Theorem 1. Assume that f satisfies the two following conditions:

$$f(x) \underset{x \rightarrow 0^+}{=} \alpha x^\beta + o(x^\beta), \quad (4)$$

$$f(x) \underset{x \rightarrow +\infty}{=} O(x), \quad (5)$$

for some α, β , such that $\beta < 0$. For any prior π positive on Θ and absolutely continuous w.r.t. the Lebesgue measure with continuous and positive Radon–Nikodym derivative denoted by π as well, the quantity $k^{d\beta/2} I_{d_f}(\pi|k)$ has a positive limit when $k \rightarrow \infty$:

$$\lim_{k \rightarrow \infty} k^{d\beta/2} I_{d_f}(\pi|k) = \alpha C_\beta \int_{\Theta} \pi(\theta)^{1+\beta} |\mathcal{I}(\theta)|^{-\beta/2} d\theta, \quad (6)$$

where $C_\beta = (2\pi)^{d\beta/2} (1 - \beta)^{-d/2}$. Moreover, if $\alpha(\beta + 1) > 0$, then

$$\lim_{k \rightarrow \infty} k^{d\beta/2} (I_{d_f}(J|k) - I_{d_f}(\pi|k)) \geq 0, \quad (7)$$

where $J(\theta) = |\mathcal{I}(\theta)|^{1/2} / \int_{\Theta} |\mathcal{I}(\tilde{\theta})|^{1/2} d\tilde{\theta}$ denotes the Jeffreys prior. The equality stands iff $\pi = J$.

Note that the condition $\alpha(\beta + 1) > 0$ on α, β ensures the concavity of $x \mapsto f(1/x)$ in the neighborhood of 0. This concavity and the convexity of f are close properties, but one does not systematically imply the other. While the f -divergence is generally defined with a convex f , our proof shows that it is the concavity of $x \mapsto f(1/x)$ (in a neighborhood of 0) that is actually required in the framework of a general f -divergence-mutual mutual information.

Remark also that if $\beta = -1$ the function $x \mapsto x^{-\beta}$ is both convex and concave. A look at our proof in that case shows that the limit (6) in the theorem above does not depend on the

prior. This is the case with $f(x) = 1/x$ making therefore d_f corresponding to the chi-squared distance (see [Clarke and Sun \(1997\)](#)). A conclusion about the optimum of such generalized mutual information thus requires stronger regularity assumptions on the likelihood to push the asymptotic analysis further. [Clarke and Sun \(1997\)](#) show in a particular context that this could lead to the inverse of Jeffreys prior as the optimum.

6 Proof of theorem 1

For any $\theta \in \Theta$, we denote by $\mathbb{P}_\theta$ the probability distribution such that $\mathbb{P}_\theta(\mathbf{Y} \in B_1 \times \dots \times B_k) = \prod_{i=1}^k \mathbb{P}_{Y|T=\theta}(B_i)$ for any $B_1, \dots, B_k \in \mathcal{Y}$. We denote by $\mathbb{E}_\theta$ the expectation w.r.t. $\mathbb{P}_\theta$.

First, we note that for any $\theta, \tilde{\theta} \in \Theta$, there exists $\xi(\tilde{\theta}, \theta)$ on the segment between θ and $\tilde{\theta}$ such that

$$\log \left(\frac{\prod_{i=1}^k \ell(y_i|\tilde{\theta})}{\prod_{i=1}^k \ell(y_i|\theta)} \right) = (\tilde{\theta} - \theta)^T \sum_{i=1}^k \nabla_\theta \log \ell(y_i|\theta) + \frac{1}{2}(\tilde{\theta} - \theta)^T \sum_{i=1}^k \nabla_\theta^2 \log \ell(y_i|\xi(\tilde{\theta}, \theta))(\tilde{\theta} - \theta).$$

Let $\varepsilon, \tilde{\varepsilon} > 0$ and fix $\theta \in \Theta$. Denote $S_k = \frac{1}{\sqrt{k}} \sum_{i=1}^k \nabla_\theta \log \ell(y_i|\theta)$. For any $\theta \in \Theta$, the central limit theorem gives that while $\mathbf{y} \sim \mathbb{P}_{\mathbf{Y}|T=\theta}$, S_k converges in distribution to $\mathcal{N}(0, \mathcal{I}(\theta))$.

Choose $M > 0$ and denote $A_k = \{\|S_k\| > M\}$. We can choose M to have $\mathbb{P}_\theta(A_k) \xrightarrow[k \rightarrow \infty]{} \varepsilon/4$.

On another hand, denote $K_i = \{-x^T \nabla_\theta^2 \log \ell(y_i|\theta')x, \theta' \in \Theta, \|x\| = 1\}$ which is a compact subset of $(0, +\infty)$ as the Hessian matrix $\nabla_\theta^2 \log \ell(y_i|\theta')$ is assumed to be negative definite. We have $\frac{1}{k} \sum_{i=1}^k \inf_{\theta' \in \Theta} K_i \xrightarrow[k \rightarrow \infty]{} \mathbb{E}_\theta \inf_{\theta', x} -x^T \nabla_\theta^2 \log \ell(y|\theta')x > 0$ $\mathbb{P}_\theta$ -a.s. We choose an $m > 0$ smaller than this limit and we denote $B_k = \{\frac{1}{k} \sum_{i=1}^k \inf K_i < m\}$. Then $\mathbb{P}_\theta(B_k) \xrightarrow[k \rightarrow \infty]{} 0$.

Denote

$$C = \sup_{k>0} \sup_{\tilde{\theta}, \theta \in \Theta} \exp(\sqrt{k}M\|\tilde{\theta} - \theta\|) \exp\left(-\frac{mk}{2}\|\tilde{\theta} - \theta\|^2\right) ; \quad c = \inf_{\theta \in \Theta} \pi(\theta)(2\pi)^{d/2}|\mathcal{I}(\theta)|^{-1/2}$$

which are finite and positive. There exists a ν and a η such that for any $c < |x| < C$, $|l| < \nu$ and $|h| < \eta$, $|f(xh + lh) - h^\beta x^\beta \alpha| \leq \tilde{\varepsilon}|h^\beta|$. Using the $\mathbb{P}_\theta$ -a.s. convergence of $-\frac{1}{k} \nabla_\theta^2 \log \ell_k(\mathbf{y}|\theta)$ to $\mathcal{I}(\theta)$, we have $\mathbb{P}_\theta(C_k) \xrightarrow[k \rightarrow \infty]{} 0$ where

$$C_k = \left\{ \forall \|x\| \leq M, \left| \pi(\theta)(2\pi)^{d/2}|k\mathcal{I}(\theta)|^{-1/2} \exp\left(\frac{1}{2}x^T \mathcal{I}(\theta)^{-1}x\right) - \pi(\theta)(2\pi)^{d/2}|\nabla_\theta^2 \ell_k(\mathbf{y}|\theta)|^{-1/2} \exp\left(-\frac{k}{2}x^T \nabla_\theta^2 \log \ell_k(\mathbf{y}|\theta)^{-1}x\right) \right| \geq \nu k^{-d/2} \right\}.$$

We write $d_f(\mathbb{P}_\mathbf{Y}||\mathbb{P}_{\mathbf{Y}|T=\theta}) = \mathbb{E}_\theta \left[f\left(\int_\Theta h_\mathbf{y}(\tilde{\theta}, \theta) \pi(\tilde{\theta}) d\tilde{\theta}\right) \right]$ where

$$h_\mathbf{y}(\tilde{\theta}, \theta) = \exp\left((\tilde{\theta} - \theta)^T \sum_{i=1}^k \nabla_\theta \log \ell(y_i|\theta)\right) \exp\left(\frac{1}{2}(\tilde{\theta} - \theta)^T \sum_{i=1}^k \nabla_\theta^2 \log \ell(y_i|\xi(\tilde{\theta}, \theta))(\tilde{\theta} - \theta)\right).$$

Firstly, we focus on what happens upon the event $A_k^c \cap B_k^c \cap C_k^c$, and we bound $h_{\mathbf{y}}(\tilde{\theta}, \theta)$ under it:

$$h_{\mathbf{y}}(\tilde{\theta}, \theta) \leq \exp(M\sqrt{k}\|\tilde{\theta} - \theta\|) \exp\left(-\frac{k}{2}m\|\tilde{\theta} - \theta\|^2\right) \leq C.$$

Therefore, consider $\delta_k = Dk^{-1/2}$ and notice that $D > 0$ can be chosen such that $\int_{\|x\| > \delta_k} e^{M\sqrt{k}x} e^{-\frac{k}{2}x^2} \sup_{\Theta} |\pi| dx < k^{-d/2}\nu$, to have both

$$\begin{aligned} \int_{\|\tilde{\theta} - \theta\| > \delta_k} h_{\mathbf{y}}(\tilde{\theta}, \theta) \pi(\tilde{\theta}) d\tilde{\theta} &< \nu k^{-d/2} \\ \int_{\|\tilde{\theta} - \theta\| > \delta_k} \exp((\tilde{\theta} - \theta)^T \sqrt{k} S_k) \exp\left(\frac{1}{2}(\tilde{\theta} - \theta)^T \nabla_{\theta}^2 \log \ell_k(\mathbf{y}|\theta)(\tilde{\theta} - \theta)\right) \pi(\theta) d\tilde{\theta} &< \nu k^{-d/2}. \end{aligned}$$

By using the uniform continuity on Θ of π one can choose k such that for any $\|\theta - \tilde{\theta}\| < \delta_k$, $|\pi(\theta) - \pi(\tilde{\theta})| \leq \nu/(CD^d)$. We then have

$$\begin{aligned} &\left| \int_{\|\tilde{\theta} - \theta\| < \delta_k} h_{\mathbf{y}}(\tilde{\theta}, \theta) \pi(\tilde{\theta}) d\tilde{\theta} \right. \\ &\quad \left. - \int_{\|\tilde{\theta} - \theta\| < \delta_k} \exp((\tilde{\theta} - \theta)^T \sqrt{k} S_k) \exp\left(\frac{1}{2}(\tilde{\theta} - \theta)^T \nabla_{\theta}^2 \log \ell_k(\mathbf{y}|\theta)(\tilde{\theta} - \theta)\right) \pi(\theta) d\tilde{\theta} \right| \leq \nu k^{-d/2}. \end{aligned}$$

Also, as

$$\begin{aligned} &\int_{\mathbb{R}^d} \exp((\tilde{\theta} - \theta)^T \sqrt{k} S_k) \exp\left(\frac{1}{2}(\tilde{\theta} - \theta)^T \nabla_{\theta}^2 \log \ell_k(\mathbf{y}|\theta)(\tilde{\theta} - \theta)\right) \pi(\theta) d\tilde{\theta} \\ &\quad = \pi(\theta) (2\pi)^{d/2} |\nabla_{\theta}^2 \log \ell_k(\mathbf{y}|\theta)|^{-1/2} \exp\left(-\frac{k}{2} S_k^T \nabla_{\theta}^2 \log \ell_k(\mathbf{y}|\theta)^{-1} S_k\right), \quad (8) \end{aligned}$$

we have under C_k^c that $|(8) - \pi(\theta) (2\pi)^{d/2} |k\mathcal{I}(\theta)|^{-1/2} \exp\left(\frac{1}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k\right)| \leq \nu k^{-d/2}$. Eventually, if k is large enough so that $k^{-d/2} < \eta/4$, then we have under $A_k^c \cap B_k^c \cap C_k^c$:

$$\left| f\left(\int_{\Theta} h_{\mathbf{y}}(\tilde{\theta}, \theta) \pi(\tilde{\theta}) d\tilde{\theta}\right) - \alpha k^{-d\beta/2} \left(\pi(\theta) (2\pi)^{d/2} |\mathcal{I}(\theta)|^{-1/2} \exp\left(\frac{1}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k\right)\right)^{\beta} \right| \leq \tilde{\varepsilon} k^{-d\beta/2}.$$

Choosing k large enough let $\mathbb{P}(A_k \cup B_k \cup C_k) < \varepsilon$, this states the following limit in $\mathbb{P}_{|\theta}$ -probability:

$$k^{d\beta/2} \left| f\left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)}\right) - \alpha k^{-d\beta/2} \left(\pi(\theta) (2\pi)^{d/2} |\mathcal{I}(\theta)|^{-1/2} \exp\left(\frac{1}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k\right)\right)^{\beta} \right| \xrightarrow[k \rightarrow \infty]{\mathbb{P}_{|\theta}} 0. \quad (9)$$

As we look for a convergence in $L^1(\mathbb{P}_{|\theta})$ we must show that this family of random variables is equi-integrable in $L^1(\mathbb{P}_{|\theta})$. Choose an event A such that $\mathbb{P}_{|\theta}(A) \leq \rho$. Write

$$\begin{aligned} &\mathbb{E}_{|\theta} \left[\mathbb{1}_A k^{d\beta/2} \left| f\left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)}\right) - \alpha k^{-d\beta/2} \left(\pi(\theta) (2\pi)^{d/2} |\mathcal{I}(\theta)|^{-1/2} \exp\left(\frac{1}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k\right)\right)^{\beta} \right| \right] \\ &\quad \leq \mathbb{E}_{|\theta} \left[\mathbb{1}_A k^{d\beta/2} \left| f\left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)}\right) \right| \right] + H \mathbb{E}_{|\theta} \left[\mathbb{1}_A \exp\left(\frac{\beta}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k\right) \right], \quad (10) \end{aligned}$$

for some constant $H > 0$. Under the assumption that $\beta < 0$, the second expected value is bounded from above by $\rho\hat{H}$ for some constant $\hat{H} > 0$. In order to bound from above the first expected value, we use the following convergence in probability stated in [Clarke and Barron \(1990\)](#):

$$\log \frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})} - \frac{d}{2} \log \frac{k}{2\pi} + \log \pi(\theta) - \frac{1}{2} \log \det \mathcal{I}(\theta) + \frac{1}{2} S_k^T \mathcal{I}(\theta) S_k \xrightarrow[k \rightarrow \infty]{\mathbb{P}_\theta} 0.$$

Choose $\gamma > 0$ such that $\forall x > \gamma, |f(x)| \leq \hat{C}(1+x)$, $\forall x < \gamma, |f(x)| \leq \hat{C}'x^\beta$ and write $D_k = \{p_{\mathbf{Y}}(\mathbf{y}) > \gamma \ell_k(\mathbf{y}|\theta)\}$ to state

$$\mathbb{E}_\theta \left[\mathbb{1}_{A_k} k^{d\beta/2} \left| f \left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right) \right| \right] \leq \rho^{1/2} 2\hat{C} k^{\beta d/2} \mathbb{P}_\theta(D_k)^{1/2} + \hat{C}' \rho^{1/2} K$$

using Lemma 1 proved in Section 7. Finally, as $\beta < 0$, the resulting upper bound of the expected value from equation (10) makes the sequence from equation (9) equi-integrable. Thus, its convergence holds in $L^1(\mathbb{P}_\theta)$.

Using the convergence in distribution under $\mathbb{P}_\theta$ of S_k to a normal distribution, we get

$$d_f(\mathbb{P}_{\mathbf{Y}} || \mathbb{P}_{\mathbf{Y}|\theta}) \underset{k \rightarrow \infty}{=} \alpha k^{-d\beta/2} \pi(\theta)^\beta (2\pi)^{d\beta/2} |\mathcal{I}(\theta)|^{-\beta/2} \mathbb{E}[e^{\frac{\beta}{2} \|X\|^2}] + o(k^{-d\beta/2})$$

where $X \sim \mathcal{N}(0, I_d)$. We will now dominate those terms to integrate this limit w.r.t. θ . We choose $\gamma > 0$ such that $\forall x > \gamma, |f(x)| \leq \hat{C}(1+x)$, $\forall x < \gamma, |f(x)| \leq \hat{C}'x^\beta$. We have

$$\begin{aligned} k^{d\beta/2} d_f(\mathbb{P}_{\mathbf{Y}} || \mathbb{P}_{\mathbf{Y}|\theta}) &= k^{d\beta/2} \mathbb{E}_\theta \left[\mathbb{1}_{D_k} f \left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right) \right] + k^{d\beta/2} \mathbb{E}_\theta \left[\mathbb{1}_{D_k^c} f \left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right) \right] \\ &\leq 2\hat{C} k^{d\beta/2} + \hat{C}' K \leq 2\hat{C} + \hat{C}' K \end{aligned}$$

by Lemma 1. This domination allows to write

$$\lim_{k \rightarrow \infty} k^{d\beta/2} I_{d_f}(\pi|k) = \alpha (2\pi)^{\beta d/2} (1-\beta)^{-d/2} \int_{\Theta} \pi(\theta)^\beta |\mathcal{I}(\theta)|^{-\beta/2} \pi(\theta) d\theta$$

as $\mathbb{E}[e^{\frac{\beta}{2} \|X\|^2}] = (1-\beta)^{-d/2}$. This concludes the first assertion of the theorem.

Now let J be the Jeffreys prior on Θ : $J(\theta) = |\mathcal{I}(\theta)|^{1/2} / \int_{\Theta} |\mathcal{I}(\tilde{\theta})|^{1/2} d\tilde{\theta}$. We have

$$\begin{aligned} \lim_{k \rightarrow \infty} k^{d\beta/2} (I_{d_f}(J|k) - I_{d_f}(\pi|k)) &= -\alpha (2\pi)^{d\beta/2} (1-\beta)^{-d/2} \int_{\Theta} \pi(\theta)^\beta |\mathcal{I}(\theta)|^{-\beta/2} \pi(\theta) d\theta \\ &\quad + \alpha (2\pi)^{d\beta/2} (1-\beta)^{-d/2} \left(\int_{\Theta} |\mathcal{I}(\theta)|^{1/2} d\theta \right)^{-\beta}. \end{aligned}$$

Under the assumption of the concavity of $x \mapsto \alpha x^{-\beta}$ (i.e. $\alpha(\beta+1) > 0$) we can write

$$\alpha \int_{\Theta} \pi(\theta)^\beta |\mathcal{I}(\theta)|^{-\beta/2} \pi(\theta) d\theta \leq \alpha \left(\int_{\Theta} \pi(\theta)^{-1} |\mathcal{I}(\theta)|^{1/2} \pi(\theta) d\theta \right)^{-\beta} = \alpha \left(\int_{\Theta} |\mathcal{I}(\theta)|^{1/2} d\theta \right)^{-\beta}$$

and we get $\lim_{k \rightarrow \infty} k^{d\beta/2} (I_{d_f}(J|k) - I_{d_f}(\pi|k)) \geq 0$.

7 A technical lemma

In this section we prove the following lemma:

Lemma 1. *Let $\zeta < 0$. There exists a constant K independent of θ such that*

$$k^{d\zeta/2} \int_{\mathcal{Y}^k} \left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right)^\zeta \ell_k(\mathbf{y}|\theta) d\mu^{\otimes k}(\mathbf{y}) \leq K.$$

Set $\theta \in \Theta$, write $\delta_k = D/\sqrt{k}$ with $D > 0$ such that $-\zeta D^2/2 = \xi$ (see Assumption 1) and consider $k \geq k_0$ where k_0 is chosen such that $\delta_{k_0} < \delta$. We compute

$$\begin{aligned} \mathbb{E}_{|\theta} \left[\left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right)^\zeta \right] &\leq \mathbb{E}_{|\theta} \left[\left(\int_{\|\tilde{\theta}-\theta\| \leq \delta_k} \exp((\tilde{\theta}-\theta)^T \sqrt{k} S_k) \right. \right. \\ &\quad \cdot \exp \left(-\frac{k}{2} (\tilde{\theta}-\theta)^T (\hat{\mathcal{I}}_k(\xi(\tilde{\theta}, \theta)) - \mathcal{I}(\theta)) (\tilde{\theta}-\theta) \right) \exp \left(-\frac{k}{2} (\tilde{\theta}-\theta)^T \mathcal{I}(\theta) (\tilde{\theta}-\theta) \right) \pi(\tilde{\theta}) d\tilde{\theta} \Big)^\zeta \Big], \end{aligned}$$

with $\hat{\mathcal{I}}_k(\theta')$ denoting $-\frac{1}{k} \sum_{i=1}^k \nabla_{\theta'}^2 \log \ell(y_i|\theta')$. Denote $G_k = \sup_{\|\tilde{\theta}-\theta\| \leq \delta_k, \|x\|=1} x^T (\hat{\mathcal{I}}_k(\tilde{\theta}) - \mathcal{I}(\theta))x$, and $m_\pi = \inf_{\Theta} \pi$. We have

$$\begin{aligned} \mathbb{E}_{|\theta} \left[\left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right)^\zeta \right] &\leq \mathbb{E}_{|\theta} \left[m_\pi^\zeta e^{-\frac{D^2\zeta}{2} G_k} \left(k^{-d/2} \int_{\|x\| < D} e^{x^T S_k} e^{-\frac{1}{2} x^T \mathcal{I}(\theta) x} dx \right)^\zeta \right] \\ &\leq \mathbb{E}_{|\theta} \left[m_\pi^\zeta e^{-\frac{D^2\zeta}{2} G_k} k^{-d\zeta/2} (2\pi)^{2\zeta/2} |\mathcal{I}(\theta)|^{\zeta/2} e^{\frac{\zeta}{2} S_k^T \mathcal{I}(\theta)^{-1} S_k} e^{-\zeta D \|S_k\|} e^{-\frac{D^2\zeta}{2} \|\mathcal{I}(\theta)\|} D^{d\zeta} \right]. \end{aligned}$$

Let H'' be an upper bound of the function $x \mapsto e^{\frac{\zeta}{2} \inf_{\|x'\|=1, \theta'} (x'^T \mathcal{I}(\theta')^{-1} x') \|x\|^2} e^{-\zeta D \|x\|}$. It comes

$$\mathbb{E}_{|\theta} \left[\left(\frac{p_{\mathbf{Y}}(\mathbf{y})}{\ell_k(\mathbf{y}|\theta)} \right)^\zeta \right] \leq k^{-d\zeta/2} 2^{-\zeta} (2\pi)^{d\zeta/2} |\mathcal{I}(\theta)|^{\zeta/2} H'' e^{-\frac{D^2\zeta}{2} \|\mathcal{I}(\theta)\|} D^{d\zeta} \mathbb{E}_{|\theta} \left[e^{-\frac{D^2\zeta}{2} G_k} \right].$$

The first term of the right-hand side is a continuous function of θ . It remains to bound the last expected value to conclude. For any $\tilde{\theta}$ such that $\|\tilde{\theta}-\theta\| \leq \delta_k$,

$$x^T (\mathcal{I}(\theta) - \hat{\mathcal{I}}_k(\tilde{\theta}))x \leq \sup_{\Theta} \|\mathcal{I}\| + \frac{1}{k} \sum_{i=1}^k \sup_{\|\tilde{\theta}-\theta\| < \delta_k} \|\nabla_{\tilde{\theta}}^2 \log \ell_k(\mathbf{y}|\tilde{\theta})\|.$$

Therefore, $\mathbb{E}_{|\theta} \left[e^{-\frac{D^2\zeta}{2} G_k} \right] \leq e^{-\frac{\zeta D^2}{2} \sup_{\Theta} \|\mathcal{I}\|} \mathbb{E}_{|\theta} \left[\exp \left(-\frac{\zeta^2 D}{2} \sup_{\|\tilde{\theta}-\theta\| < \delta} \|\nabla_{\tilde{\theta}}^2 \log \ell(y_1|\tilde{\theta})\| \right) \right]$, which is bounded uniformly in $\theta \in \Theta$ by Assumption 1. This concludes the proof of the lemma.

8 Conclusion and perspectives

From the initial definition of the reference prior introduced by Bernardo (1979), we have constructed a generalization of the objective information criterion. Using a connection with

global sensitivity analysis, we suggest a different point of view for the interpretation of such reference prior. Within various areas of study, Bayesian analysis is a tool for estimation which relies on the prior choice. Our emphasis on how the Jeffreys prior as a reference choice beyond the standard mutual information metric reinforces its use for studies which seek for objectivity (e.g. [Van Biesbroeck et al. \(2023\)](#)). Also, the expression of our generalized criterion as a sensitivity index of the prior to the data simplifies the numerical methods for approximation of an optimal prior through a Neural Network based iterative process [Nalisnick and Smyth \(2017\)](#).

References

- Berger, J. O., Bernardo, J. M. and Sun, D. (2009), ‘The formal definition of reference priors’, *The Annals of statistics* **37**(2), 905–938.
- Bernardo, J. M. (1979), ‘Reference posterior distributions for Bayesian inference’, *Journal of the Royal Statistical Society. Series B* **41**(2), 113–147.
- Chen, M.-H., Dey, D. K., Müller, P., Sun, D. and Ye, K. (2010), *Objective Bayesian Inference with Applications*, Springer New York, New York, NY, pp. 31–68.
- Clarke, B. S. and Barron, A. R. (1990), ‘Information-theoretic asymptotics of Bayes methods’, *IEEE Transactions on Information Theory* **36**(3), 453–471.
- Clarke, B. S. and Barron, A. R. (1994), ‘Jeffreys’ prior is asymptotically least favorable under entropy risk’, *Journal of Statistical Planning and Inference* **41**(1), 37–60.
- Clarke, B. and Sun, D. (1997), ‘Reference priors under the chi-squared distance’, *Sankhyā: The Indian Journal of Statistics* **59**(2), 215–231.
- Csiszár, I. (1967), ‘Information-type measures of difference of probability distributions and indirect observation’, *Studia Scientiarum Mathematicarum Hungarica* **2**, 229–318.
- Da Veiga, S. (2015), ‘Global sensitivity analysis with dependence measures’, *Journal of Statistical Computation and Simulation* **85**(7), 1283–1305.
- Liu, R., Chakrabarti, A., Samanta, T., Ghosh, J. K. and Ghosh, M. (2014), ‘On Divergence Measures Leading to Jeffreys and Other Reference Priors’, *Bayesian Analysis* **9**(2), 331–370.
- Nalisnick, E. and Smyth, P. (2017), Learning approximately objective priors, in ‘Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)’.
- Sobol’, I. (1993), ‘Sensitivity estimates for non linear mathematical models’, *Mathematical Modeling and Computer Experiments* **1**, 407–414.
- Van Biesbroeck, A., Gauchy, C., Feau, C. and Garnier, J. (2023), ‘Reference prior for Bayesian estimation of seismic fragility curves’. [arxiv.2302.06935](#).