International Journal of
Applied & Experimental Mathematics
Research Article Open Access

Non-informative Prior with Maximum Divergence for Non-regular Bayesian Estimation

Shintaro Hashimoto[1,*] **and Ken-ichi Koike**[2]
Department of Mathematics, Hiroshima University, 1-3-1 Kagamiyama, Higashihiroshima, Hiroshima, 739-8511, Japan
Institute of Mathematics, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8571, Japan

Abstract

This paper presents a non-informative prior which maximizes a general divergence (called the -divergence) between the prior and the corresponding posterior distribution for non-regular family of distributions whose support depends on unknown parameter. This result is a general- ization of the result of Ghosal and Samanta [1] based on the Kullback-Leibler divergence. In a non-regular case, the prior which is different from the Jeffreys prior is obtained under certain conditions. Further, we show that the prior which maximizes the chi-square divergence does not exist for non-regular family in general. This result differs from the regular case by Ghosh et al. [2]. The comparison between regular and non-regular cases is also discussed.

Mathematics Subject Classication: Primary 62F10; Secondary 62F12

Publication History:

Received: September 22, 2016
Accepted: December 27, 2016
Published: December 29, 2016

Keywords:

Asymptotic expansion, Bayesian inference, Jeffreys prior, Posterior distribution, Trun-cation family

Introduction

In Bayesian inference, the selection of priors has been an important and much discussed problem. The concept of prior distribution is useful when we have much prior information for unknown parameter. However, we often have a little prior information for real situations. In such cases, we need to consider `non-informative' or `objective' prior. One of the most widely used non-informative priors is a uniform distribution over the parameter space. However, a uniform prior lacks invariance under smooth one-to-one transformation. For example, if we do not have information about the parameter θ, and we do not have information about $1/\theta$, either. Thus, a uniform prior lacks invariance under smooth one-to-one transformation in such a situation. To overcome this difficulty, Jeffreys [3] proposes the prior which is proportional to the positive square root of the Fisher information number in one dimensional case, which is known as the Jeffreys prior. This prior is invariant under smooth one-to-one transformation. Jeffreys [3] also generalizes this to multidimensional case by letting the prior be proportional to the positive square root of the determinant of the Fisher information matrix. However, in the presence of nuisance parameters, this prior suffers from many problems (see Bernardo and Smith [4]).

Another class of non-informative priors is the reference prior, which was proposed by Bernardo [5] and was extended by Berger and Bernardo [6]. The reference prior is defined by maximizing the Kullback-Leibler (KL) divergence between the prior and the posterior under some regularity conditions. This prior maximizes the expected posterior information to the prior, i.e., the prior is the `least informative' prior in some aspects. In regular case, the reference prior coincides with the Jeffreys prior in one-dimensional case, but does not in the presence of nuisance parameters (Bernardo and Smith [4]). These results are derived rigorously by Clarke and Barron [7, 8]. Besides these, there are many methods defining non-informative prior, e.g., the entropy maximizer of Jaynes [9], and the probability matching prior of Welch and Peers [10] (see also Tibshirani [11], and Datta and Mukerjee [12]) among others. We also refer to Kass and Wasserman [13] and Ghosh [14] as reviews on non-informative priors.

In the context of the reference priors, Ghosh et al. [2] derives the priors which asymptotically maximize the more general divergence measure (called the α-divergence) between the prior and the corresponding posterior. We note that the KL divergence, the Bhattacharyya-Hellinger divergence and the chi-square divergence are special cases of the α-divergence. Ghosh et al. [2] also shows that maximizing the divergence yields the Jeffreys prior with the exception of the case of the chi-square divergence. Maximizing the chi-square divergence yields a prior different from the Jeffreys prior.

However, Ghosh et al. [2] deals with the case of regular distributions and the result is not applied for non-regular distributions whose support depends on unknown parameter. The asymptotic expansion of the posterior distribution for non-regular distribution is derived by Ghosh et al.[15] and Ghosal and Samanta [16]. They show that the first order asymptotic distribution of the posterior distribution is an exponential distribution, that is, the asymptotic normality of the posterior distribution does not hold in non-regular case. Ghosal and Samanta [1] derives the prior which asymptotically maximizes the KL divergence in non-regular case. In non-regular case, the prior which is different from the Jeffreys prior is derived.

The aim of this paper is the generalization of the result of Ghosal and Samanta [1] by using the α-divergence. In the case of $-1 < \alpha < 1$, we obtain the same prior given by Ghosal and Samanta [1]. On the other hands, we show that the prior which maximizes the chi-square divergence ($\alpha = -1$) does not generally exist in non-regular case. This result differs from the regular case in Ghosh et al. [2].

Asymptotic expansion of the posterior distribution and the shrink-age argument

Let $X_1,\ldots, X_n$ be independent and identically distributed observations from a density $f(x,\theta)$ $(\theta \in \Theta \subset R)$ with respect to the Lebesgue measure.

*Corresponding Author: Dr. Shintaro Hashimoto, Department of Mathematics, Hiroshima University, 1 Chome-3-2 Kagamiyama, Higashihiroshima, Hiroshima Prefecture 739-8511, Japan, E-mail: s-hashimoto@hiroshima-u.ac.jp

Citation: Hashimoto S, Koike K (2016) Non-informative Prior with Maximum Divergence for Non-regular Bayesian Estimation. Int J Appl Exp Math 1: 106. doi: http://dx.doi.org/10.15344/ijaem/2016/111

We assume that for all $\theta \in \Theta$, $f(x, \theta)$ is strictly positive in a closed interval $S(\theta) := [a_1(\theta); a_2(\theta)]$ depending on unknown parameter θ and is zero outside $S(\theta)$. It is permitted that one of the endpoints is free from θ and may be plus or minus innity. We assume the following conditions on the density $f(x, \theta)$, which are the same conditions as Ghosal and Samanta [16].

(A1) The endpoints $a_1(\theta)$ and $a_2(\theta)$ of the support are continuously differentiable functions of θ.
(A2) On the set $\{\theta: a_1(\theta)\} < x < a_2(\theta)$, $f(x, \theta)$ is continuously differentiable in θ.
(A3) For each x, $\log f(x,\theta)$ is twice differentiable in θ on $\theta : a_1(\theta) < x < a_2(\theta)\}$. Further, the following holds:

(a) For all $\theta \in \Theta$, $c(\theta) := E_\theta(\partial/\partial\theta) \log f(X_1,\theta)] < \infty$ is differentiable in θ and $c(\theta) \neq 0$. Moreover, $d(\theta) := E_\theta[(\partial^2/\partial\theta^2) \log f(x_1,\theta)] < \infty$.

(b) There exist a neighborhood N_{θ_0} of the true parameter θ_0 and an integrable function $H_{\theta_0}(x)$ such that for all $\theta \in N_{\theta_0}$ and $x \in (a_1(\theta), a_2(\theta))$, $(\partial^2/\partial\theta^2) \log f(x_1, \theta) | < H_{\theta_0}(x)$.

(A4) For a sufficiently large $\lambda > 0, E_{\theta_0} = [\sup_{\theta<\theta_0-\lambda} \log\{f(X_1, \theta) / f(X_1, \theta_0)\}] < 0$.

(A5) $E_{\theta_0} \log f(X_1, \theta, \rho) \to E_{\theta_0} \log f(X_1, \theta)(\rho \to 0)$, where $f(x, \theta, \rho) = \sup\{f(x, \theta') : |\theta\text{-}\theta'| \le \rho\}$.

Further, we assume the following on the prior density.

(A6) The prior density $\pi(\theta)$ is twice differentiable in θ.

We note that conditions (A3)-(A5) ensure the validity of the asymptotic expansion of the posterior distribution (cf. Ghosal and Samanta [16]). Families such as uniform distribution $U(0, \theta)$, location family $f(x, \theta) = f_0(x\text{-}\theta)$, with a positive smooth density f_0 on $[0, \infty)$ and the truncation family $f(x,\theta) = g(x) / \bar{G}(\theta)$ $(x > \theta)$, where g is a positive smooth density on $[0,\infty)$ and $\bar{G}(\theta) = \int_x^\infty g(t)dt$, satisfy the above conditions.

In view of the results of Ghosh et al. [15], in order to have a limit of the posterior distributions, it is necessary that the set $S(\theta)$ is either increasing or decreasing in θ, that is, $S(\theta)$ satisfies either $S(\theta) \subseteq S(\theta + \varepsilon)$ for $\varepsilon > 0$ or $S(\theta) \subseteq S(\theta + \varepsilon)$ for $\varepsilon < 0$, respectively. For this reason, we may assume $S(\theta)$ is decreasing without loss of generality. Indeed, the case where $S(\theta)$ increases with θ may be reduced to the case where $S(\theta)$ decreases by the reparametrization θ (θ). When $S(\theta)$ is decreasing, the set $\{a_1(\theta) \le X_i \le a_2(\theta); i = 1; 2, \ldots, n\}$ can be expressed as $\{\hat{\theta}_n(X_1,\ldots,X_n) \ge \theta\}$ where $\hat{\theta}_n := \min\{a_1^{-1}(X_{(1)}), a_2^{-1}(X_{(n)})\}$ and $X_{(1)} := \min_{1\le i\le n} X_i$, $X_{(n)} := \max_{1\le i\le n} X_i$. If a_1 does not depend on θ, then we interpret the above $\hat{\theta}_n$ as $a_2^{-1}(X_{(n)})$ while it is interpreted as $a_2^{-1}(X_{(1)})$ if a_2 does not depend on θ. Note that $\hat{\theta}_n - \theta = O_p(n^{-1})(n \to \infty)$. Define

$$\sigma := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i, \hat{\theta}_n).$$

Note that $\sigma - c(\theta) = O_p(n^{-1})(n \to \infty)$ (Ghosal and Samanta [16], Lemma 2.1). By Theorem 3.1 of Ghosal and Samanta [16], the posterior density of $u = n\sigma(\theta - \hat{\theta}_n)$ given $\boldsymbol{X} = (X_1, \ldots, X_n)$ has the asymptotic expansion

$$\pi\left(u\middle|X\right) = e^u\left[1 + \frac{1}{n}\left\{\frac{\pi'(\hat{\theta}_n)}{\sigma\pi(\hat{\theta}_n)}(u+1) + \frac{c_2}{\sigma_2}(u^2-2)\right\} + O(n^{-2})\right] \quad (u<0), \tag{1}$$

where

$$c_2 := \frac{1}{2n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i, \hat{\theta}_n).$$

Putting $\theta = u / n\sigma + \hat{\theta}_n$ in (1), we have

$$\pi\left(\theta\middle|X\right) = n|\sigma| e^{n\sigma(\theta-\hat{\theta}_n)}\left[1 + \frac{1}{n}\left\{\frac{\pi'(\hat{\theta}_n)}{\sigma\pi(\hat{\theta}_n)}\left(n\sigma(\theta-\hat{\theta}_n)+1\right) + \frac{c_2}{\sigma^2}\left((n\sigma)^2(\theta-\hat{\theta}_n)^2 - 2\right)\right\} + O(n^{-2})\right] \quad (\theta \le \hat{\theta}_n)$$

Here, the (expected) α-divergence between the prior and the posterior is dened by

$$R^a(\pi) := \frac{1 - \int\left[\int \pi^\alpha(\theta)\pi^{1-\alpha}\left(\theta\middle|\boldsymbol{x}\right)d\theta\right]m(\boldsymbol{x})d\boldsymbol{x}}{\alpha(1-\alpha)} \tag{2}$$

where $m(\boldsymbol{x})$ is the marginal density of $\boldsymbol{X} = (X_1, \ldots, X_n)$ (cf. Amari [17] and Ghosh et al. [2]). Note that the α-divergence smoothly connects the KL divergence (α→0), the squared Bhattacharyya-Hellinger divergence (α = 1/2), and the chi-square divergence (α = -1). Let $L_n(\theta)$ be the likelihood function of θ. From the relation $L_n(\theta)\pi(\theta) = \pi(\theta|\boldsymbol{x})\, m(\boldsymbol{x})$, we can express (2) as

$$\begin{aligned} R^\alpha(\pi) &:= \frac{1 - \int\int \pi^{\alpha+1}(\theta)\pi^{-\alpha}\left(\theta\middle|\boldsymbol{x}\right)L_n(\theta)d\boldsymbol{x}d\theta}{\alpha(1-\alpha)} \\ &= \frac{1 - \int \pi^{\alpha+1}(\theta)E_\theta\left[\pi^{-\alpha}\left(\theta\middle|x\right)\right]d\theta}{\alpha(1-\alpha)}, \end{aligned} \tag{3}$$

where $E_\theta[\cdot]$ denotes the conditional expectation of $\boldsymbol{X}$ given θ. In order to derive the prior which maximizes the expected α-divergence, we need to compute the expectation $E_\theta[\pi^\alpha(\theta|\boldsymbol{x})]$ in (3). Since the exact computation of this expectation is not easy, we consider the asymptotic approximation of $E_\theta[\pi^\alpha(\theta|\boldsymbol{x})]$. We have the following theorem.

Theorem 2.1. *Under the conditions (A1)-(A6), the asymptotic approximation of $E_\theta[\pi^\alpha(\theta|\boldsymbol{x})]$ for $\alpha < 1$ is*

$$E_\theta\left[\pi^{-\alpha}\left(\theta\middle|X\right)\right] = \frac{\{n|c(\theta)|\}^{-\alpha}}{1-\alpha}\left[1 + \frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)} + \psi_2(\theta)\right\}\right] + O(n^{-2-\alpha}) \tag{4}$$

as $n \to \infty$, where

$$\psi_2(\theta) := \psi_1(\theta) + \alpha(\alpha+1)c'(\theta), \ \psi_1(\theta) = \frac{2d(\theta)\alpha(\alpha+1)}{c(\theta)^2}$$

are continuous, not involving $\pi(\theta)$.

Proof. **Step one**. We consider a proper prior density $\bar{\pi}(\cdot)$ such that the support of $\bar{\pi}(\cdot)$ is compact in the parameter space and $\bar{\pi}(\cdot)$ vanishes outside of the support while remaining positive in the interior. Next, we compute the expectation $E^{\bar{\pi}}\left[\pi^{-\alpha}(u|\mathcal{X})|\mathcal{X}\right]$, where $E^{\bar{\pi}}(\cdot)$ denotes the expectation under the posterior density $\bar{\pi}(\cdot|X)$. Since

$$\begin{aligned} \pi^{-\alpha}\left(\theta\middle|\boldsymbol{X}\right)\bar{\pi}\left(\theta\middle|\boldsymbol{X}\right) &= \left(n|\sigma|\right)^{1-\alpha} e^{(1-\alpha)n\sigma(\theta-\hat{\theta}_n)} \\ &\cdot\left[1 + \frac{1}{n}\left\{-\frac{\alpha\pi'(\hat{\theta}_n)}{\sigma\pi(\hat{\theta}_n)}\left(n\sigma(\theta-\hat{\theta}_n)+1\right) + \frac{\bar{\pi}'(\hat{\theta}_n)}{\sigma\bar{\pi}(\hat{\theta}_n)}\left(n\sigma(\theta-\hat{\theta}_n)+1\right)\right.\right. \\ &\left.\left. + \frac{c_2}{\sigma^2}(1-\alpha)\left((n\sigma)^2(\theta-\hat{\theta}_n)^2 - 1\right)\right\} + O(n^{-2})\right], \end{aligned}$$

the expectation of $\pi^{-\alpha}\left(\theta\middle|X\right)$ under $\bar{\pi}\left(\cdot\middle|X\right)$ is given by

$$E^{\bar{\pi}}\left[\pi^{-\alpha}\left(\theta\middle|X\right)\middle|X\right]=\int_{-\infty}^{\hat{\theta}_n}\pi^{-\alpha}\left(\theta\middle|x\right)\bar{\pi}\left(\theta\middle|x\right)d\theta$$

$$=\frac{\left(n|\sigma|\right)^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'\left(\hat{\theta}_n\right)}{\sigma\pi\left(\hat{\theta}_n\right)}-\frac{\alpha\bar{\pi}'\left(\hat{\theta}_n\right)}{\sigma\bar{\pi}\left(\hat{\theta}_n\right)}+\frac{2c_2\alpha(\alpha+1)}{\sigma^2}\right\}+O\left(n^{-2}\right)\right].$$

Note that in order to compute the above integration, we put $n\sigma\left(\theta-\hat{\theta}_n\right)$=-t and regard the integration as the expectation of the exponential distribution with mean parameter (1-α).

Step two. For an interior point θ of the support $\bar{\pi}(\cdot)$, we compute $\lambda(\theta)=E_\theta\left[E^{\bar{\pi}}\left[\pi^{-\alpha}\left(\theta\middle|X\right)\middle|X\right]\right]$, $E_\theta[\cdot]$ denotes the conditional expectation of $\boldsymbol{X}$ given. Since $\hat{\theta}_n-\theta=O_p\left(n^{-1}\right),\sigma-c(\theta)=O_p\left(n^{-1}\right),c_2-d(\theta)=O_p\left(n^{-1}\right)$, the method in Datta and Mukerjee [12] gives

$$\lambda(\theta)=\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)}-\frac{\alpha\bar{\pi}'(\theta)}{c(\theta)\bar{\pi}(\theta)}+\psi_1(\theta)\right\}\right]+O\left(n^{-2-\alpha}\right).$$

Where

$$\psi_1(\theta):=\frac{2d(\theta)\alpha(\alpha+1)}{c(\theta)^2}$$

is a continuous function, neither involving $\pi(\theta)$ nor $\bar{\pi}(\theta)$.

Step three. The nal step of this argument involves integrating λ(θ) with respect to $\bar{\pi}(\cdot)$ and then making $\bar{\pi}(\theta)$ degenerate at θ. We have

$$\int\lambda(\theta)\bar{\pi}(\theta)d\theta=\int\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)}-\frac{\alpha\bar{\pi}'(\theta)}{c(\theta)\bar{\pi}(\theta)}+\psi_1(\theta)\right\}\right]\bar{\pi}(\theta)d\theta+O\left(n^{-2-\alpha}\right)$$

$$=\int\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)}+\psi_1(\theta)\right\}\right]\bar{\pi}(\theta)d\theta$$

$$-\int\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\frac{\alpha\bar{\pi}'(\theta)}{c(\theta)(1-\alpha)}d\theta+O\left(n^{-2-\alpha}\right)\cdot \quad (5)$$

Here, integration by parts for the second term of (5) gives

$$\int\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\frac{\alpha\bar{\pi}'(\theta)}{c(\theta)(1-\alpha)}d\theta$$

$$=\int\frac{n^{-\alpha}}{1-\alpha}\frac{\alpha}{1-\alpha}\frac{d}{d\theta}\left\{c(\theta)^{-\alpha-1}\right\}\bar{\pi}(\theta)d\theta.$$

Now suppose that the support of $\bar{\pi}(\cdot)$ contains the true θ as an interior point. Then allowing $\bar{\pi}(\cdot)$to converge weakly to the degenerate prior at θ, we have

$$\int\lambda(\theta)\bar{\pi}(\theta)d\theta=\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)}+\psi_1(\theta)+\alpha(\alpha+1)c'\right\}\right]+O\left(n^{-2-\alpha}\right)$$

Hence, the asymptotic approximation of $E_\theta[\pi^\alpha(\theta|\boldsymbol{x})]$ is given by

$$E_\theta\left[\pi^{-\alpha}\left(\theta\middle|\mathcal{X}\right)\right]=\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}\left[1+\frac{1}{n(1-\alpha)}\left\{\frac{\alpha^2\pi'(\theta)}{c(\theta)\pi(\theta)}+\psi_2(\theta)\right\}\right]+O\left(n^{-2-\alpha}\right), \quad (6)$$

where

$$\psi_2(\theta):=\psi_1(\theta)+\alpha(\alpha+1)c'(\theta)$$

is a continuous function, not involving $\pi(\theta)$. This completes the proof.

Remark 2.1. In the proof, we adopt the computation method called the shrinkage argument (Ghosh [18], Datta and Mukerjee [12]). This method is a Bayesian approach for frequentist computations. The shrinkage argument consists of three steps (For the details, see Datta and Mukerjee [12]).

Remark 2.2. Theorem 2.1 does not hold for α ≥ 1 as is evident from the right-hand-side expression in (4).

Non-informative priors as maximizer of expected α-divergences

In this section, we derive maximizing priors of the expected α-divergences between the prior and the corresponding posterior in a similar way to Ghosh et al. [2]. We assume the following condition concerning the integration of order.

(A7) From (4) the expectation $E_\theta[\pi^\alpha(\theta|\boldsymbol{x})]$ is expressed by

$$E_\theta\left[\pi^{-\alpha}\left(\theta\middle|x\right)\right]=\frac{\left|c(\theta)\right|^{-a}}{n^\alpha(1-\alpha)}+d(\theta),$$

where $d(\theta)=O(n^{-1-\alpha})$, then it holds

$$\int_\Theta d(\theta)\pi^{\alpha+1}(\theta)d\theta=O\left(n^{-1-\alpha}\right),$$

where $c(\theta)$ is a function of θ, and $d(\theta)$ is a function of θ and n.

Remark 3.1. In fact, since it is not easy to conrm the condition (A7), we do not discuss it here.

From (4) and (A7), for α < 1 and α ≠ 0 or -1, the first order approximation to (3) is given by

$$R^\alpha(\pi)\approx\frac{1-\int\pi^{\alpha+1}(\theta)\frac{\left\{n|c(\theta)|\right\}^{-\alpha}}{1-\alpha}d\theta}{\alpha(1-\alpha)}$$

$$=\frac{1-\frac{1}{n^\alpha(1-\alpha)}\int\left\{\frac{\pi(\theta)}{|c(\theta)|}\right\}^\alpha\pi(\theta)d\theta}{\alpha(1-\alpha)}. \quad (7)$$

We now consider maximization of (7) with respect to $\pi(\cdot)$ under $\int\pi(\theta)d\theta=1$ according to the value of α. The main theorem in this paper is the following.

Theorem 3.1. Under the conditions (A1)-(A7), non-informative prior based on the α-divergence is given by

$$\pi(\theta)\propto|c(\theta)|=\left|E_\theta\left[\frac{\partial}{\partial\theta}\log f\left(X_1,\theta\right)\right]\right| \quad (8)$$

for -1 < α < 1, and such prior generally does not exist for α ≤ -1.

Proof. The proof consists of the following five parts.

Case I $0 < \alpha < 1$.

In this case, we consider minimizing

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{a} \pi(\theta) d\theta$$

with respect to $\pi()$ under $\int \pi(\theta) d\theta = 1$. First, we recall the Hölder inequality, i.e.,

$$\int |f_1(\theta) f_2(\theta)| d\theta \le \left(\int |f_1(\theta)|^p d\theta \right)^{1/p} \left(\int |f_1(\theta)|^q d\theta \right)^{1/q}$$

where $p > 1$, $q > 1$ and $p^{-1}+q^{-1} = 1$. The equality holds if and only if $\{f_1(\theta)\}^p = a\{f_2(\theta)\}^q$, where a is a constant. Now putting $f_1(\theta) = \pi(\theta) |c(\theta)|^{-\alpha/(1+\alpha)}$, $f_2(\theta) = |c(\theta)|^{\alpha/(1+\alpha)}$, $p = 1+\alpha$ and $q = (1+\alpha)/\alpha$, we have

$$\left(\int \pi^{1+\alpha}(\theta) |c(\theta)|^{-\alpha} d\theta \right)^{1/(1+\alpha)} \left(\int |c(\theta)| d\theta \right)^{\alpha/(1+\alpha)} \ge \int \pi(\theta) d\theta = 1$$

or equivalently,

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{\alpha} \pi(\theta) d\theta \ge \left(\int |c(\theta)| d\theta \right)^{-\alpha}$$

The equality holds if and only if $\pi(\theta) \propto |c(\theta)|$, which is the maximizing prior of the expected α-divergence for $0 < \alpha < 1$.

Case II $-1 < \alpha < 0$.

We need to maximize

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{\alpha} \pi(\theta) d\theta$$

with respect to $\pi(\cdot)$ under $\int \pi(\theta) d\theta = 1$, Here let $f_1(\theta) = \pi(\theta) |c(\theta)|^{-\alpha/(1+\alpha)}$, $f_2(\theta) = |c(\theta)|^{\alpha/(1+\alpha)}$, $p = 1+\alpha$ and $q = (1+\alpha)/\alpha$, By the Hölder inequality, it holds

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{\alpha} \pi(\theta) d\theta \le \left(\int \pi(\theta) \right)^{1+\alpha} \left(\int |c(\theta)| d\theta \right)^{-\alpha} = \left(\int |c(\theta)| d\theta \right)^{-\alpha}.$$

The equality holds if and only if $\pi(\theta) \propto |c(\theta)|$, which is the maximizing prior of the expected α-divergence for $-1 < \alpha < 0$.

Case III $\alpha < -1$.

Putting $\alpha = -\beta (\beta > 1)$, by (7), we have

$$R^{\alpha}(\pi) = \frac{1 - \frac{1}{n^{-\beta}(1+\beta)} \int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{-\beta} \pi(\theta) d\theta}{\beta(1+\beta)}.$$

Hence it suffices to maximize

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{-\beta} \pi(\theta) d\theta$$

with respect to $\pi(.)$ under $\int \pi(\theta) d\theta = 1$. We may write

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{-\beta} \pi(\theta) d\theta = E\left[\left\{ \frac{|c(\theta)|}{\pi(\theta)} \right\}^{\beta} \right],$$

where $E[\cdot]$ denotes the expectation under the prior $\pi(\cdot)$. Here we recall the Lyapounov inequality (DasGupta [19]), i.e.,

$$\left\{ E\left(X^b \right) \right\}^{1/b} \ge \left\{ E\left(X^a \right) \right\}^{1/a}, \quad 0 < a < b.$$

In this inequality, putting $a = 1$, $b = \beta$,

$$\left\{ E\left[\left\{ \frac{|c(\theta)|}{\pi(\theta)} \right\}^{\beta} \right] \right\}^{1/\beta} \ge E\left(\frac{|c(\theta)|}{\pi(\theta)} \right),$$

or equivalently,

$$\int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{\alpha} \pi(\theta) d\theta \ge \left(\int |c(\theta)| d\theta \right)^{\beta}.$$

The equality holds if and only if $\pi(\theta) \propto |c(\theta)|$. Hence in this case $\pi(\theta) \propto |c(\theta)|$ is the minimizer rather than the maximizer of (7). For $\alpha < -1$, we can show that there is no maximizing prior in the same way of Ghosh et al. [2]. For the Lagrange multiplier λ, we put $H_{\lambda}(\pi) := \left\{ \pi(\theta) / |c(\theta)| \right\}^{-\beta} \pi(\theta) - \lambda \pi(\theta)$. Since

$$\frac{\partial}{\partial \pi} H_{\lambda}(\pi) = (1-\beta) \pi^{-\beta}(\theta) |c(\theta)|^{\beta} - \lambda = 0,$$

the prior $\pi(\theta) \propto |c(\theta)|$ is the minimizer and not the maximizer.

Case IV $\alpha = 0$.

For $= 0$, we need to interpret (7) as its limiting value (when it exists). In this case, (7) corresponds to the KL divergence. Indeed, it suffices to compute the following

$$R^0(\pi) = \lim_{\alpha \to 0} R^{\alpha}(\pi) = \lim_{\alpha \to 0} \frac{1 - \frac{1}{n^{\alpha}(1-\alpha)} \int \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\}^{\alpha} \pi(\theta) d\theta}{\alpha(1-\alpha)}.$$

By L'Hospital's rule, we have

$$R^0(\pi) = \log n + 1 - \int \log \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\} d\theta.$$

Hence, we need to minimize

$$\int \pi(\theta) \log \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\} d\theta$$

with respect to $\pi(\cdot)$ under $\int \pi(\theta) d\theta = 1$. For the Lagrange multiplier η, we put $H_{\eta}(\pi) := \pi(\theta) \log\{\pi(\theta)/c(\theta)|\} + \eta\pi(\theta)$. Since

$$\frac{\partial}{\partial \pi} H_{\lambda}(\pi) = \log \left\{ \frac{\pi(\theta)}{|c(\theta)|} \right\} + 1 + \eta = 0,$$

we have $\pi(\theta) \propto |c(\theta)|$. This is the maximizing prior of the expected α-divergence for $\alpha = 0$ (see also Ghosal and Samanta [1]).

Case V $\alpha = -1$.

In this case, the α-divergence corresponds to the chi-square divergence. Since $\pi^{\alpha+1}(\theta) = 1$, the rst order term in (7) is constant, and we need to consider the second order term. When $\alpha = -1$ in (4), we note that $\alpha(1-\alpha) = -2 < 0$. When $c(\theta) > 0$ for all $\theta \in \Theta$, it suffices to maximize

$$\int \frac{\pi'(\theta)}{\pi(\theta)} d\theta$$

with respect to $\pi(\cdot)$ under $\int \pi(\theta) d\theta = 1$. However, we can not find such $\pi(\theta)$ in general. For example, putting $\pi(\theta) = \sin\theta (0 \le \theta \le \pi/2)$, we have $\int \pi'(\theta) / \pi(\theta) d\theta = \int_0^{\pi/2} (\tan\theta) \, d\theta = \infty$. As in the case of $c(\theta) > 0$,

we can also show that the maximizing divergence prior generally does not exist in the case of $c(\theta) < 0$.

Comparison between Regular and Non-regular Cases

As previously stated, Ghosh et al.[2] derives non-informative prior as a maximizer of the α-divergence in regular case. In this section, we compare our result with Ghosh et al. [2]. The result is given in Table 1.

From Table 1, for $-1 < \alpha < 1$, the prior $\pi(\theta) \propto |c(\theta)|$ in non-regular case corresponds to the Jeffreys prior in regular case. This is the same as the reference prior in Ghosal and Samanta [1] and the probability matching prior in Ghosal [20]. The prior $\pi(\theta) \propto |c(\theta)|$ is also invariant under a smooth one-to-one transformation, that is, if we consider a smooth one-to-one transformation $\eta = \phi(\theta)$. the maximum divergence prior for is $|c^*(\eta)| := |c(\theta)||d\theta / d\eta|$. For $\alpha = -1$, i.e., the chi-square divergence, Ghosh et al. [2] derives a prior which is different from the Jeffreys prior in regular case. On the other hand, in non-regular case, we show that there is no maximizing prior of the chi-square divergence in general. For $\alpha = -1$, there is no maximizing prior in both cases.

Finally we show some typical non-regular examples and compute the prior $\pi(\theta) \propto |c(\theta)|$ for each examples. Note that Examples 4.1 and 4.3 are the same as Ghosal and Samanta [1].

α (<1)	regular case	non-regular case
$0 < \alpha < 1$	$\pi(\theta) \propto \sqrt{I(\theta)}$	$\pi(\theta) \propto \lvert c(\theta)\rvert$
$\alpha = 0$	$\pi(\theta) \propto \sqrt{I(\theta)}$	$\pi(\theta) \propto \lvert c(\theta)\rvert$
$-1 < \alpha < 0$	$\pi(\theta) \propto \sqrt{I(\theta)}$	$\pi(\theta) \propto \lvert c(\theta)\rvert$
$\alpha = -1$	$\pi(\theta) \propto \exp\left[\int \frac{2g_3(\theta) - I'(\theta)}{4I(\theta)} d\theta\right]^*$	–
$\alpha < -1$	–	–

Table 1: The comparison of the maximum α-divergence priors.

* $I(\theta) = E_\theta[\{(\partial / \partial\theta) \log f(X_1, \theta)^2] < \infty$ denotes the Fisher information number and $g_3(\theta) := E_\theta[(\partial^3 / \partial\theta^3) \log f(X_1, \theta)] < \infty$.

Example 4.1 (Location family). Let f_0 be a strictly positive density on $[0, \infty)$. Consider the location family of distribution $f(x, \theta) = f_0(x-\theta)$. In particular, the location family $f(x, \theta) = e^{-(x-\theta)}(x > \theta)$ of exponential distribution belongs to this. Since $|c(\theta)|$ is constant, the uniform prior is the maximum divergence prior.

Example 4.2 (Scale family). Let f_0 be a strictly positive density on $[0; 1]$. Consider the scale family of distribution $f(x, \theta) = \theta^{-1} f_0(x/\theta)(\theta > 0)$. In particular, the uniform distribution $U(0, \theta)(\theta > 0)$ belongs to this. Since $|c(\theta)| \propto \theta^{-1}$, $\pi(\theta) \propto \theta^{-1}$ is the maximum divergence prior.

Example 4.3 (Truncation family). Let $g(x)$ be a strictly positive density on $(0, \infty)$, and let $f(x, \theta) = g(x) / \bar{G}(\theta)(x > \theta > 0)$, where $\bar{G}(\theta) = \int_\theta^\infty g(t)\, dt$. Since $|c(\theta)| = g(\theta) / \bar{G}(\theta)$, the maximum divergence prior is $\pi(\theta) = g(\theta) / \bar{G}(\theta)$ which is the hazard rate of $g(x)$. In particular, this family corresponds to the one-sided truncated exponential distribution when $g(x) = e^{-x}$. In this case, non-informative prior of a truncation parameter θ is

$$\pi(\theta) = \frac{g(\theta)}{\bar{G}(\theta)} = \frac{e^{-\theta}}{\int_\theta^\infty e^{-t} dt} = 1,$$

that is, the uniform distribution. In this case, the Bayes estimator is easily calculated because the posterior distribution of θ is proportional to the likelihood function under the uniform prior (Ghosh et al. [21]). On the other hand, this family corresponds to the one-sided truncated normal distribution when $g(x) = \phi(x) = \left(1/\sqrt{2\pi}\right) e^{-x^2/2}$. In this case, non-informative prior of a truncation parameter θ is

$$\pi(\theta) = \frac{g(\theta)}{\bar{G}(\theta)} = \frac{\phi(\theta)}{\int_\theta^\infty \phi(t)\, dt} = \frac{\phi(\theta)}{1 - \Phi(\theta)}, \tag{9}$$

with $\Phi(x) = \int_{-\infty}^x \phi(t)\, dt$. Since the prior (9) involves a non-linear function $\Phi(\theta)$, when we actually compute the Bayes estimator under this prior (9), it may need to use the computational method like Markov chain Monte Carlo (MCMC) methods.

Remark 4.1. When we compute the posterior distribution based on non-informative prior, we often deal with improper priors and direct evaluation of these integrals over the entire parameter space is so difficult. An important point to note is that evaluation of all integrals is carried out over an increasing sequence of compact set K_i. For example, in the case of location exponential family of distribution in Example 4.1, the parameter space of θ is **R** and one can take the increasing sequence of compact sets $[-i, i]$ $(i \geq 1)$. Evaluations of these integrals are usually carried out by taking a sequence of priors π_i with compact support K_i, and nally using sufficiently large i (Berger and Bernardo [6], Ghosh [14]).

Conclusion

The generalization of the result of Ghosal and Samanta [1] by using the α-divergence was given. We showed that the prior which maximizes the chi-square divergence ($\alpha = -1$) does not generally exist in non-regular case. The comparison between regular and non-regular cases was also discussed. Finally, we gave some non-regular examples involving location, scale and truncation family of distributions. As a future plan, we need to construct non-informative priors in more complex situations. This paper can be extended to multiparametric family in the presence of nuisance parameters which involve the one-sided truncated exponential family of distributions (see Akahira [22]).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Both the authors substantially contributed to the study conception and design as well as the acquisition and interpretation of the data and drafting the manuscript.

References

1. Ghosal S, Samanta T (1997a) Expansion of Bayes risk for entropy loss and reference prior in non-regular cases. Statistics & Decisions 15: 129-140.
2. Ghosh M, Mergel V, Liu R (2011) A general divergence criterion for prior selection. Annals of the Institute of Statistical Mathematics 63: 49-58.
3. Jeffreys H (1961) Theory of Probability (3rd Edition), Oxford University Press, London.
4. Bernardo J, Smith A (1994) Bayesian Theory. Wiley, Chichester.
5. Bernardo JM (1979) Reference posterior distributions for Bayesian inference (with discussion). J R Statist Soc B41: 113-147.
6. Berger JO, Bernardo JM (1989) Estimating a product of means: Bayesian analysis with reference priors. Journal of the American Statistical Association 84: 200-207.

7. Clarke B, Barron A (1990) Information-theoretic asymptotics of Bayes methods. IEEE Transactions on Information Theory 36: 453-471.

8. Clarke B, Barron A (1994) Jeffreys' prior is asymptotically least favorable under entropy risk. Journal of Statistical Planning and Inference 41: 37-60.

9. Jaynes ET (1968) Prior probabilities. IEEE Transactions on Systems Science and Cyber-netics SSC-4: 227-241.12

10. Welch BL, Peers HW (1963) On formulae for condence points based on integrals of weighted likelihoods. J R Statst Soc 25: 318-329.

11. Tibshirani RJ (1989) Noninformative priors for one parameter of many. Biometrika 76: 604-608.

12. Datta GS, Mukarjee R (2004) Probability Matching Priors: Higher Order Asymptotics.Springer, New York.

13. Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. Journal of the American Statistical Association 91: 1343-1370.

14. Ghosh M (2011) Objective priors: an introduction for frequentists. Statistical Science 26: 187-202.

15. Ghosh JK, Ghosal S, Samanta T (1994) Stability and convergence of posterior in non-regular problems. Statistical Decision Theory and Related Topics V (eds. S. S. Gupta and J. O. Berger),183-199, Springer, New York.

16. Ghosal S, Samanta T (1997b) Asymptotic expansion of posterior distributions in non-regular case. Annals of the Institute of Statistical Mathematics 49: 181-197.

17. Amari S (1982) Differential geometry of curved exponential families curvatures and information loss. The Annals of Statistics 10: 357-387.

18. Ghosh JK (1994) Higher Order Asymptotics. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol.4, Institute of Mathematical Statistics, Hayward.

19. DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability. Springer, New York.

20. Ghosal S (1999) Probability matching priors for non-regular cases. Biometrika 86: 956-964.

21. Ghosh JK, Delampady M, Samanta T (2006) An Introduction to Bayesian Analysis. Springer, New York.

22. Akahira M (2016) Second order asymptotic comparison of the MLE and MCLE of a natu-ral parameter for a truncated exponential family of distributions. Annals of the Institute of Statistical Mathematics 68: 469-490.