Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 45–65

# Selection of Divergence Priors in the Presence of Nuisance Parameters

**Ruitao Liu**[1] **and Malay Ghosh**[2]
[1] *NebulAi Education*
[2] *University of Florida*

Received: 30 March 2020; Revised: 13 April 2020; Accepted: 21 April 2020

**Abstract**

The selection of priors in the presence of nuisance parameters is an important topic in Bayesian statistics. Bernardo (1979) proposed a stepwise procedure for handling nuisance parameters. He obtained the prior by maximizing the expected Kullback-Leibler divergence between the prior of parameters of interest and the corresponding posterior. His procedure turns out to be very efficient and has been applied to many examples. In this paper, we consider selection of priors in the presence of nuisance parameters under a general divergence criterion, originally introduced by Renyi (1961), later followed by a host of researchers, most notably by Amari (1982) and Cressie and Read (1984). This general divergence measure includes the Kullback-Leibler, Bhattacharyya-Hellinger and Chi-square divergence. It turns out that Bernardo's prior maximizes this divergence in the interior of this class of divergence measures. On the boundary, the Chi-square divergence, the prior turns out to be different from Bernardo's prior for some common families of distributions. Also, outside the boundary, Bernardo's prior turns out to be the minimizer rather than maximizer of the divergence, and there does not exist any prior which maximizes the divergence between the prior and the posterior.

*Key words*: General divergence; Nuisance parameters; Optimal prior; Characterisation of optimal priors.

## 1. Introduction

The most important component in Bayesian statistics is the prior for the unknown parameters. The selection of prior has always been a popular topic since the birth of Bayesian statistics.

Ideally, if one has enough historical data, it is possible to elicit an appropriate prior which reflects one's belief about unknown parameters. This is a subjective prior. But the choice of subjective priors is difficult, especially when there is not enough historical information available. In practice, it is common to use the so-called 'objective' priors which are also referred to as 'non-informative priors' or 'default priors'. Those priors are determined

Corresponding Author: Malay Ghosh
Email: ghoshm@ufl.edu

by some objective or structural criterion. For decades, many statisticians worked on this topic. Consequently, numerous criteria for selecting objective priors have been proposed.

One major criterion was proposed by Bernardo (1979). There are two innovations in his paper. First, he introduced the notion of missing information. He used the expected Kullback-Leibler divergence as the measure of the missing information provided by the data and then found the prior which maximizes this divergence. In the absence of nuisance parameters, he found that the desired prior is Jeffreys' general rule prior (Jeffreys 1961) which is proportional to the square root of the determinant of the Fisher information matrix. Second, he used a stepwise procedure to find the prior in the presence of nuisance parameters. His procedure is as follows. First, he assigns a conditional density of nuisance parameters given parameters of interest. Then he obtains the prior for parameters of interest by maximizing the expected Kullback-Leibler divergence between the prior for parameters of interest and the corresponding posterior.

In many cases, Bernardo's procedure produced priors different from Jeffreys' general rule prior. It turns out that his stepwise procedure often yields more reasonable priors than Jeffreys' general rule prior. One good example to show the advantage of Bernardo's stepwise procedure is the Neyman-Scott problem (1948) considered by Berger and Bernardo (1992b), Datta and Ghosh (1995a). The data consist of $n$ pairs of observations: $X_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \cdots, n$, $j = 1, 2$. Consider all the parameters to be of equal importance. Then, one gets Jeffreys' general rule prior $\pi(\mu_1, \cdots, \mu_n, \sigma^2) \propto (\sigma^2)^{-n/2-1}$. So the posterior mean is $s^2/(2n-2)$, where $s^2 = \sum_{i=1}^n \sum_{j=1}^2 (x_{ij} - \bar{x}_i)^2$ and $\bar{x}_i = (x_{i1} + x_{i2})/2$. This is an inconsistent estimator of $\sigma^2$. On the contrary, by treating $\sigma$ as the parameter of interest, using Bernardo's procedure, one gets the prior $\pi(\mu_1, \cdots, \mu_n, \sigma^2) \propto \sigma^{-2}$. This gives a posterior mean of $s^2/(n-2)$ which is consistent.

Bernardo's procedure has been applied to many examples. For example, exponential regression (Ye and Berger 1991), multinomial models (Berger and Bernardo 1992a) and AR(1) models (Berger and Yang 1994).

In both cases, with or without nuisance parameters, Bernardo used the expected Kullback-Leibler divergence to develop priors. One may ask questions like: What will happen if we use another divergence? Will we get the same priors as Berdardo did? If not, what do the new priors look like?

Instead of the Kullback-Leibler divergence, Clarke and Sun (1997) considered the expected Chi-square divergence motivated by the classical Chi-square goodness-of-fit statistic. They showed that, for the one-parameter exponential family of distributions with the canonical parameter, maximization of the Chi-square divergence led to a prior different from Jeffreys' prior. For multi-parameter exponential family of distributions, they conjectured that the prior should also be of the same form as they got in one-parameter case. For the case where nuisance parameters are present, they gave brief discussion and left it as an open question.

Recently, for regular one-parameter family of distributions, Ghosh, Mergel and Liu (2011) considered a general divergence between prior and posterior which has been considered in other contexts by several authors (for example, Renyi, 1961; Amari, 1982; Cressie and Read, 1984). It is a family of divergence measures including the Kullback-Leibler divergence, the Bhattacharyya-Hellinger divergence (Bhattacharyya, 1943; Hellinger, 1909), and the Chi-square divergence. They showed that Jeffreys' general prior is the desired prior under each divergence measure that is in the interior in this class of divergence measures. On the boundary, namely for the Chi-square divergence, the prior turns out to be different from Jefferys' prior for some common families of distributions but still maintains the invariance property. Also, outside the boundary, Jeffreys' prior turns out to be the minimizer rather than maximizer of the divergence, and there does not exist any prior which maximizes the distance between the posterior and the prior. A more comprehensive set of results were later obtained by Liu, Chakrabarty, Samanta, Ghosh and Ghosh (2014) for one parameter family of distributions.

In this paper, we consider prior selection in the presence of nuisance parameters under the general divergence used by Ghosh, Mergel and Liu (2011). This is a generalization of previous work of Bernardo (1979). We characterize optimal priors for every member in this family of divergence measures by using the two-step procedure proposed by Bernardo (1979). Explicit expressions for the optimal priors under every divergence measure (except for the Chi-square divergence) are given. Specifically, for the Kullback-Leibler divergence, we get the same prior as found by Bernardo (1979). Under the Chi-square divergence, we have shown that the objective prior should be the solution to a set of partial differential equations. We also consider a special case when the parameter of interest is one dimensional. In this case, a closed form expression for the optimal prior is provided also under the Chi-square divergence.

The outline of the remaining sections is as follows. In Section 2 of this paper, we have provided a general scheme of deriving the asymptotic expansion of the expected general divergence. Section 3 is devoted to the derivation of optimal prior in the interior of the divergence class and non-existence of optimal priors outside the boundary of this class. Section 4 provides a characterization of optimal priors under the Chi-square divergence class followed by some examples. Some remarks are made in Section 5. The proof of the main result in Section 4 is deferred to the Appendix.

## 2. Derivation of Priors

Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, where the $X_i$ are independent and identically distributed with common pdf $f(x \,|\, \boldsymbol{\theta})$. Parameter vector $\boldsymbol{\theta}$ can be partitioned as,

$$\boldsymbol{\theta}^T = (\theta_1, \cdots, \theta_{d_1}, \theta_{d_1+1}, \cdots, \theta_d) = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T) \in \mathbf{R}^d,$$

where $\boldsymbol{\theta}_1$ are $d_1-$dimensional nuisance parameters and $\boldsymbol{\theta}_2$ are $d_2-$dimensional parameters of interest. Here, we assume the parameter space is a compact set in $\mathbf{R}^d$ and consider a prior $p(\boldsymbol{\theta}_2)$ which puts all its mass on a compact set in $\mathbf{R}^{d_2}$. One passes on to the limit eventually in many of the actual examples considered in the literature.

We apply the following two step procedure proposed by Bernardo (1979) to find the divergence priors for the parameters of interest.

First, for fixed $\boldsymbol{\theta}_2$, one assigns a conditional density $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ to the nuisance parameters $\boldsymbol{\theta}_1$.

Second, let $p(\boldsymbol{\theta}_2)$ denote the marginal density function of $\boldsymbol{\theta}_2$. Then the divergence prior $p(\boldsymbol{\theta}_2)$ for the parameter of interest $\boldsymbol{\theta}_2$ is chosen by maximizing the asymptotic expected general divergence $R^\beta(p(\boldsymbol{\theta}_2))$ between the prior $p(\boldsymbol{\theta}_2)$ and the corresponding posterior, that is

$$R^\beta(p(\boldsymbol{\theta}_2)) = \frac{1 - \int \left[\int p^\beta(\boldsymbol{\theta}_2) p^{1-\beta}(\boldsymbol{\theta}_2|\, \boldsymbol{x}_n)\, d\boldsymbol{\theta}_2\right] m(\boldsymbol{x}_n)\mu(d\boldsymbol{x}_n)}{\beta(1-\beta)},$$

where $\mu(d\boldsymbol{x}_n)$ is a dominating measure and $m(\boldsymbol{x}_n)$ is the marginal density of $\boldsymbol{x}_n$

Note 1. The expected general divergence criterion as introduced by Renyi (1961), Amari (1982) and Cressie and Read (1984) is a family of divergences with index parameter $\beta$. When $\beta = 1/2$, this is the Bhattacharyya-Hellinger distance, and $\beta = -1$ amounts to the Chi-square distance. For $\beta = 0$ or 1, we need to interpret $R^\beta(p(\boldsymbol{\theta}))$ as its limiting value (when it exists). In particular,

$$R^0(p(\boldsymbol{\theta})) = \iint \left\{\log \frac{p(\boldsymbol{\theta}\,|\,\boldsymbol{x}_n)}{p(\boldsymbol{\theta})}\right\} p(\boldsymbol{\theta}\,|\,\boldsymbol{x}_n) m(\boldsymbol{x}_n)\, d\boldsymbol{\theta}\mu(d\boldsymbol{x}_n), \tag{1}$$

which is the KL divergence between the prior and the posterior considered in Bernardo (1979).

Note 2. In Step 1, for $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$, Bernardo recommends using the conditional Jeffreys' general rule prior which is proportional to $\sqrt{|I_{11}(\boldsymbol{\theta})|}$, where $I_{11}(\boldsymbol{\theta})$ is the part of Fisher information matrix $I(\boldsymbol{\theta})$ corresponding to the nuisance parametes.

Note 3. In Step 2, with the choice of $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$, we first find an asymptotic expansion of $R^\beta(p(\boldsymbol{\theta}_2))$ and then obtain the priors by maximizing that expansion.

Derivation of divergence priors in the presence of nuisance parameters is complicated. First, we give a general scheme of deriving the asymptotic expansion of the expected general divergence. Then, in Sections (3) and (4), by using the asymptotic expansion with different order of the remainder terms, we consider the prior selection for two cases when $\beta \neq -1$ and $\beta = -1$ separately.

By the relation $f_n(\boldsymbol{x}_n|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2) = p(\boldsymbol{\theta}_2|\boldsymbol{x}_n)m(\boldsymbol{x}_n)$, one can rewrite $R^\beta(p(\boldsymbol{\theta}_2))$ as

$$\begin{aligned} R^\beta(p(\boldsymbol{\theta}_2)) &= \frac{1 - \iint p^{\beta+1}(\boldsymbol{\theta}_2) p^{-\beta}(\boldsymbol{\theta}_2\,|\,\boldsymbol{x}_n) f_n(\boldsymbol{x}_n|\boldsymbol{\theta}_2)\, \mu(d\boldsymbol{x}_n)\, d\boldsymbol{\theta}_2}{\beta(1-\beta)} \\ &= \frac{1 - \int p^{\beta+1}(\boldsymbol{\theta}_2) E\left[p^{-\beta}(\boldsymbol{\theta}_2\,|\,\boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right]\, d\boldsymbol{\theta}_2}{\beta(1-\beta)}, \end{aligned} \tag{2}$$

where $f_n(\boldsymbol{x}_n|\boldsymbol{\theta}_2)$ is the joint density function of $\boldsymbol{x}_n = (x_1, \cdots, x_n)$ given $\boldsymbol{\theta}_2$.

By using the shrinkage argument proposed by Ghosh (1994), one can find the asymptotic expansion to $E\left[p^{-\beta}(\boldsymbol{\theta}_2\,|\,\boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right]$ and then find the asymptotic expansion to $R^\beta(p(\boldsymbol{\theta}_2))$. The shrinkage argument is discussed in details in Datta and Mukerjee (2004).

Here is the general scheme of deriving the asymptotic expansion:

- **Step 0:** Given the choice of $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$, for prior $p(\boldsymbol{\theta}_2)$, find posterior density $p(\boldsymbol{\theta}_2|\boldsymbol{x}_n)$ of $\boldsymbol{\theta}_2$ given $\boldsymbol{x}_n$:
$$p(\boldsymbol{\theta}_2|\boldsymbol{x}_n) = \int_{\boldsymbol{\theta}_1} \pi(\boldsymbol{\theta}|\boldsymbol{x}_n)d\boldsymbol{\theta}_1,$$
where $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n) \propto p(\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)\prod_{i=1}^n f(x_i|\boldsymbol{\theta})$.

- **Step 1:** Consider a proper prior density $\bar{p}(\boldsymbol{\theta}_2)$, such that the support of $\bar{p}(\boldsymbol{\theta}_2)$ is a compact rectangle in the parameter space and $\bar{p}(\boldsymbol{\theta}_2)$ vanishes on the boundary of the support while remaining positive in the interior. Consider the posterior density of $\boldsymbol{\theta}_2$ under $\bar{p}(\boldsymbol{\theta}_2)$, and obtain
$$G(\boldsymbol{x}_n) = \int p^{-\beta}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)\bar{p}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)d\boldsymbol{\theta}_2.$$

- **Step 2:** For $\boldsymbol{\theta}_2$ in the interior of the support of $\bar{p}(\boldsymbol{\theta}_2)$, compute $\lambda(\boldsymbol{\theta}_2)$ defined as
$$\begin{aligned}\lambda(\boldsymbol{\theta}_2) &= \int G(\boldsymbol{x}_n)f_n(\boldsymbol{x}_n|\boldsymbol{\theta}_2)d\boldsymbol{x}_n \\ &= \int G(\boldsymbol{x}_n)\left[\int \prod_{i=1}^n f(x_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1\right]d\boldsymbol{x}_n \\ &= \int \lambda_0(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1,\end{aligned}$$
where $\lambda_0(\boldsymbol{\theta}) = \int G(\boldsymbol{x}_n)\prod_{i=1}^n f(x_i|\boldsymbol{\theta})d\boldsymbol{x}_n$.

- **Step 3:** Integrate $\lambda(\boldsymbol{\theta}_2)$ with respect to $\bar{p}(\boldsymbol{\theta}_2)$ and then allow $\bar{p}(\boldsymbol{\theta}_2)$ to converge weakly to the degenerate prior at the true $\boldsymbol{\theta}_2$, supposing that the true $\boldsymbol{\theta}_2$ is an interior point of the support of $\bar{p}(\boldsymbol{\theta}_2)$. This yields $E\left[p^{-\beta}(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right]$.

By using the above procedure and equation (2), one can get an asymptotic approximation to $R^\beta(p(\boldsymbol{\theta}_2))$. Furthermore, the divergence priors are obtained by maximizing the approximation.

In the next two sections, according to different values of $\beta$, we derive two approximations to $R^\beta(p(\boldsymbol{\theta}_2))$ and call them the first order approximation and the second order approximation respectively. For most of values of $\beta$, the derivation of divergence prior only requires the first order approximation and it will be addressed in section (3). In section (4), we will discuss the derivation of divergence prior when $\beta = -1$; that is the only case which needs the second order approximation.

## 3. Divergence Priors for $\beta \neq -1$

In this section, we consider the prior selection for general expected divergence with $\beta \neq -1$. To begin with, we derive the first order approximation to $E\left[p^{-\beta}(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right]$. Then, in view of (2), we get the first order approximation to $R^\beta(p(\boldsymbol{\theta}_2))$. Finally, we discuss the divergence priors according to different values of $\beta$ such that $\beta \neq -1$.

First, by assuming the same regularity conditions as in Johnson (1970) and Bhattacharya and Ghosh (1978), one gets the following theorem which gives the first order expansion $E\left[p^{-\beta}(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right]$.

**Theorem 1:** For $\beta < 1$ and $\beta \neq -1$, $E\left[p^{-\beta}(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)|\boldsymbol{\theta}_2\right]$ can be expressed as:

$$E\left[p^{-\beta}(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)\middle|\, \boldsymbol{\theta}_2\right] = n^{-\frac{d_2\beta}{2}}\left[\int K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 + o(n^{-1})\right], \tag{3}$$

where

$$K(\boldsymbol{\theta}) = \left|I^{22}\right|^{\beta/2}(2\pi)^{d_2\beta/2}(1-\beta)^{-d_2/2},$$

$$I^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}.$$

**Proof of Theorem 1:**
The proof uses the shrinkage argument as mentioned before.

Let $\boldsymbol{h}^T = (h_1, \cdots, h_{d_1}, h_{d_1+1}, \cdots h_d) = (\boldsymbol{h}_1^T, \boldsymbol{h}_2^T) = \sqrt{n}(\boldsymbol{\theta}_1^T - \hat{\boldsymbol{\theta}}_1^T, \boldsymbol{\theta}_2^T - \hat{\boldsymbol{\theta}}_2^T)$, where $\hat{\boldsymbol{\theta}}^T = (\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T)$ is MLE of $\boldsymbol{\theta}$. For prior $p(\boldsymbol{\theta})$, from Datta and Mukerjee (1994), one gets the corresponding posterior density

$$p(\boldsymbol{h}|\boldsymbol{x}_n) = \phi_d(\boldsymbol{h}, C^{-1})\left[1 + n^{-\frac{1}{2}}\left\{R_1(\boldsymbol{h}) + \frac{1}{6}R_3(\boldsymbol{h})\right\}\right] + o(n^{-1}), \tag{4}$$

where $C$ is the observed Fisher information matrix, $\phi_d(\boldsymbol{h}, C^{-1})$ is the $d-$variate normal density with the null mean vector and dispersion matrix $C^{-1}$,

$$R_1(\boldsymbol{h}) = \sum_{j=1}^{d}\hat{p}_j h_j/\hat{p}, \quad R_3(\boldsymbol{h}) = \sum_{j=1}^{d}\sum_{r=1}^{d}\sum_{s=1}^{d} a_{jrs}h_j h_r h_s,$$

$$a_{jrs} = \frac{1}{n}\sum_{u=1}^{d}\frac{\partial \log f(X_u|\boldsymbol{\theta})}{\partial\theta_j\theta_r\theta_s} \text{ and } \hat{p} = p(\hat{\boldsymbol{\theta}}), \quad \hat{p}_j = \left.\frac{\partial p(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Let $N_{d_2}(\boldsymbol{h}_2|\boldsymbol{\mu}, \Sigma)$ denote the density function of multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and dispersion matrix $\Sigma$. Also, corresponding to the partition of the parameter vector $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$, we partition the matrix $C^{-1}$ as:

$$C^{-1} = \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix}.$$

It is easy to establish the relation

$$\phi_d(\boldsymbol{h}, C^{-1}) = N_{d_2}(\boldsymbol{h}_2|0, C^{22}) \times f_{d_1}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2),$$

where $f_{d_1}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) = N_{d_1}(\boldsymbol{h}_1|C^{12}[C^{22}]^{-1}\boldsymbol{h}_2, C^{11} - C^{12}[C^{22}]^{-1}C^{21})$.

In **Step 0**, for prior $p(\boldsymbol{\theta}_2)$, one gets

$$p(\boldsymbol{h}_2|\boldsymbol{x}_n) =N_{d_2}(\boldsymbol{h}_2|0, C^{22}) \left[1 + n^{-\frac{1}{2}} \left\{L_1(\boldsymbol{h}_2) + \frac{1}{6}L_3(\boldsymbol{h}_2)\right\}\right] + o(n^{-1}), \tag{5}$$

where

$$L_1(\boldsymbol{h}_2) = \int R_1(\boldsymbol{h}) f_{d_1}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1, \quad L_3(\boldsymbol{h}_2) = \int R_3(\boldsymbol{h}) f_{d_1}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1.$$

**Step 1.** We find an asymptotic expansion for

$$G(\boldsymbol{x}_n) = \int p^{-\beta}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)\bar{p}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)d\boldsymbol{\theta}_2.$$

With the general expansion

$$\left(\frac{1}{b_1 + \frac{b_2}{\sqrt{n}} + \frac{b_3}{n} + o(n^{-1})}\right)^{\beta} = b_1^{-\beta}\left(1 - \beta\frac{b_2}{b_1\sqrt{n}} + \frac{\beta}{n}\left(\frac{\beta+1}{2}\frac{b_2^2}{b_1^2} - \frac{b_3}{b_1}\right)\right) + o(n^{-1}),$$

one gets

$$p^{-\beta}(\boldsymbol{h}|\boldsymbol{x}_n) = N_{d_2}^{-\beta}(\boldsymbol{h}_2|0, C^{22}) \left[1 - \beta n^{-\frac{1}{2}} \left\{L_1(\boldsymbol{h}_2) + \frac{1}{6}L_3(\boldsymbol{h}_2)\right\}\right] + o(n^{-1}). \tag{6}$$

Using (5) and (6), for any arbitrary thrice differentiable prior $\bar{p}(\boldsymbol{\theta}_2)$ vanishing outside a compact set, one gets

$$\begin{aligned} p^{-\beta}(\boldsymbol{h}_2|\boldsymbol{x}_n)\bar{p}(\boldsymbol{h}_2|\boldsymbol{x}_n) = N_{d_2}^{1-\beta}(\boldsymbol{h}_2|0, C^{22}) &\left[1 + n^{-\frac{1}{2}} \left\{\bar{L}_1(\boldsymbol{h}_2) + \frac{1}{6}L_3(\boldsymbol{h}_2)\right.\right. \\ - \beta L_1(\boldsymbol{h}_2) - \frac{\beta}{6}L_3(\boldsymbol{h}_2)&\left.\left.\right\}\right] + o(n^{-1}), \end{aligned} \tag{7}$$

where

$$\bar{L}_1(\boldsymbol{h}_2) = \int \bar{R}_1(\boldsymbol{h}) f_{d_1}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1.$$

Here are some observations: (i) $N_{d_2}^{1-\beta}(\boldsymbol{h}_2|0, C^{22}) = K(\hat{\boldsymbol{\theta}}) \times N_{d_2}(\boldsymbol{h}_2|0, \frac{C^{22}}{1-\beta})$,
(ii) $f_{d_1}(\boldsymbol{h}_1|\boldsymbol{h}_2) \times N_{d_2}(\boldsymbol{h}_2|0, \frac{C^{22}}{1-\beta}) = N_d(\boldsymbol{h}|0, Q)$, where

$$K(\hat{\boldsymbol{\theta}}) = (2\pi)^{\frac{d_2\beta}{2}}|C^{22}|^{\frac{\beta}{2}}(1-\beta)^{-\frac{d_2}{2}},$$

$$Q = (q^{jr})_{d\times d} = \begin{pmatrix} C^{11} + \frac{\beta}{1-\beta}C^{12}[C^{22}]^{-1}C^{21} & C^{12}/(1-\beta) \\ C^{21}/(1-\beta) & C^{22}/(1-\beta) \end{pmatrix}.$$

With the above observations and the relation $\boldsymbol{\theta} = \boldsymbol{h}/\sqrt{n} + \hat{\boldsymbol{\theta}}$, and noting that

$$\int \bar{L}_1(\boldsymbol{h}_2)d\boldsymbol{h}_2 = \int L_1(\boldsymbol{h}_2)d\boldsymbol{h}_2 = \int L_3(\boldsymbol{h}_2)d\boldsymbol{h}_2 = 0,$$

one gets

$$\begin{aligned} G(\boldsymbol{x}_n) &= \int p^{-\beta}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)\bar{p}(\boldsymbol{\theta}_2|\boldsymbol{x}_n)d\boldsymbol{\theta}_2 \\ &= n^{-\frac{d_2\beta}{2}}\int p^{-\beta}(\boldsymbol{h}_2|\boldsymbol{x}_n)\bar{p}(\boldsymbol{h}_2|\boldsymbol{x}_n)d\boldsymbol{h}_2 = n^{-\frac{d_2\beta}{2}}\left[K(\hat{\boldsymbol{\theta}}) + o_P(n^{-1})\right] \end{aligned} \tag{8}$$

As shown above, $G(\boldsymbol{x}_n)$ can be written as a function of $\hat{\boldsymbol{\theta}}$. Also $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = o_p(n^{-1})$ $(P_{\boldsymbol{\theta}})$. Therefore, by using Taylor expansion, one gets

$$\lambda_0(\boldsymbol{\theta}) = \int G(\boldsymbol{x}_n)\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})d\boldsymbol{x}_n = n^{-\frac{d_2\beta}{2}}\left[K(\boldsymbol{\theta}) + o(n^{-1})\right], \tag{9}$$

where $K(\boldsymbol{\theta}) = (2\pi)^{\frac{d_2\beta}{2}}|I^{22}|^{\frac{\beta}{2}}(1-\beta)^{-\frac{d_2}{2}}$. Moreover, at the end of **Step 2**, one gets

$$\begin{aligned} \lambda(\boldsymbol{\theta}_2) &= \int \lambda_0(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 \\ &= n^{-\frac{d_2\beta}{2}}\left[\int K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 + o(n^{-1})\right]. \end{aligned} \tag{10}$$

In **Step 3**, integrating $\lambda(\boldsymbol{\theta}_2)$ with respect to $\bar{p}(\boldsymbol{\theta}_2)$ and allowing $\bar{p}(\boldsymbol{\theta}_2)$ weakly converge to the degenerate density of true $\boldsymbol{\theta}_2$, we obtain the final asymptotic expansion for $E\left[p^{-\beta}(\boldsymbol{\theta}_2 \mid \boldsymbol{X}_n)\middle| \boldsymbol{\theta}_2\right]$ as

$$\begin{aligned} E_{\boldsymbol{\theta}_2}\left[p^{-\beta}(\boldsymbol{\theta}_2 \mid \boldsymbol{X}_n)\right] &= n^{-\frac{d_2\beta}{2}}\left[\int K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 + o(n^{-1})\right] \\ &= n^{-\frac{d_2\beta}{2}}\left[(2\pi)^{\frac{d_2\beta}{2}}(1-\beta)^{-\frac{d_2}{2}}\int |I^{22}|^{\frac{\beta}{2}}\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 + o(n^{-1})\right]. \end{aligned} \tag{11}$$

This proves the theorem.

When $\beta < -1$ and $\beta \neq -1$, we can obtain the divergence priors by maximizing the first order approximation to the general expected divergence $R^{\beta}(p(\boldsymbol{\theta}_2))$. The approximation is derived by neglecting the $o(n^{-1})$ term in Theorem 1. That is:

$$R^{\beta}(p(\boldsymbol{\theta}_2)) \approx \frac{1}{\beta(1-\beta)}\left[1 - \left(\frac{2\pi}{n}\right)^{\frac{d_2\beta}{2}}(1-\beta)^{-\frac{d_2}{2}}\int\left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right]^{-\beta}p(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2\right], \tag{12}$$

where

$$\phi(\boldsymbol{\theta}_2) = \left[\int \left|I^{22}(\boldsymbol{\theta})\right|^{\frac{\beta}{2}}\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1\right]^{-\frac{1}{\beta}}.$$

One may think that the divergence priors should be different as $\beta$ takes on different values. Amazingly, it turns out that, in most cases, one gets the same prior. Here are the results for different values of $\beta$ $(\beta \neq -1)$.

**CASE I.** First consider the case $0 < \beta < 1$. Since then $\beta(1-\beta) > 0$, the problem of prior selection reduces to minimization of

$$\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right]^{-\beta} p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2.$$

Noting that $u^{-\beta}$ is a convex function of $u(>0)$ when $\beta(1-\beta) > 0$, by Jensen's inequality, one gets

$$\begin{aligned} &\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right]^{-\beta} p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2 \\ &\geq \left\{\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right] p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2\right\}^{-\beta} = \left\{\int \phi(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2\right\}^{-\beta} \end{aligned}$$

with equality if and only if $p(\boldsymbol{\theta}_2) \propto \phi(\boldsymbol{\theta}_2)$.

**CASE II.** Similarly, when $-1 < \beta < 0$, $\beta(1-\beta) < 0$ and now the problem reduces to maximization of

$$\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right]^{-\beta} p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2.$$

Noting that $u^{-\beta}$ is a concave function of $u(>0)$ when $-1 < \beta < 0$, again by Jensen's inequality, one gets

$$\begin{aligned} &\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right]^{-\beta} p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2 \\ &\leq \left\{\int \left[\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right] p(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2\right\}^{-\beta} = \left\{\int \phi(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2\right\}^{-\beta} \end{aligned}$$

with equality if and only if $p(\boldsymbol{\theta}_2) \propto \phi(\boldsymbol{\theta}_2)$ which is the same prior developed in the previous case.

**CASE III.** When $\beta \longrightarrow 0$, using either Theorem 1 or alternatively from Bernardo, one gets the first order approximation of the general expected divergence $R^0(p(\boldsymbol{\theta}_2))$ :

$$R^0(p(\boldsymbol{\theta}_2)) \approx K_n + \int p(\boldsymbol{\theta}_2) \log \frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} d\boldsymbol{\theta}_2,$$

where $K_n$ is a constant depending on $n$. Then, from the property of the Kullback-Leibler distance, $R^0(p(\boldsymbol{\theta}_2))$ is maximized up to first order of approximation by $p(\boldsymbol{\theta}_2) \propto \phi(\boldsymbol{\theta}_2)$.

From the above three cases, we can easily draw the conclusion that when $|\beta| < 1$, the desired divergence prior is proportional to $\phi(\boldsymbol{\theta}_2)$.

**CASE IV.** Next for $\beta < -1$, writing $\beta = -\lambda$, one can rewrite the first order approximation of the general divergence $R^\beta(p(\boldsymbol{\theta}_2))$ as :

$$R^{\beta}(p(\boldsymbol{\theta}_2)) = \frac{\left(\frac{2\pi}{n}\right)^{-d_2\lambda/2}(1+\lambda)^{-\frac{d_2}{2}}\int\left(\frac{\phi(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\right)^{\lambda}p(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2 - 1}{\lambda(1+\lambda)}, \quad \lambda > 1. \tag{13}$$

Hence it suffices to maximize

$$\int \{\phi(\boldsymbol{\theta}_2)/p(\boldsymbol{\theta}_2)\}^{\lambda}\,p(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2$$

subject to $\int p(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2 = 1$. Again, by Jensen's inequality,

$$\int \{\phi(\boldsymbol{\theta}_2)/p(\boldsymbol{\theta}_2)\}^{\lambda}\,p(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2 \geq \left[\int \{\phi(\boldsymbol{\theta}_2)/p(\boldsymbol{\theta}_2)\}\,p(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2\right]^{\lambda} = \left(\int \phi(\boldsymbol{\theta}_2)\,d\boldsymbol{\theta}_2\right)^{-\beta}$$

since $\lambda > 1$, equality holding if and only if

$$p(\boldsymbol{\theta}_2) \propto \phi(\boldsymbol{\theta}_2).$$

Thus in this case the prior $p(\boldsymbol{\theta}_2) \propto \phi(\boldsymbol{\theta}_2)$ is the minimizer rather then the maximizer of $R^{\beta}(p(\boldsymbol{\theta}_2))$. Also there is no maximizing prior in this case. In fact, one can use similar argument in the previous section to show that

$$\sup_{p}\int \phi(\boldsymbol{\theta}_2)^{\lambda}p^{1-\lambda}(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2 = +\infty$$

## 3. Divergence Priors for $\beta = -1$

It remains to consider the case $\beta = -1$, the Chi-square distance as considered in Clarke and Sun (1997) for the one parameter exponential family and in Ghosh, Mergel and Liu (2011) for the general one-parameter family of distributions. Here $p^{\beta+1}(\boldsymbol{\theta}_2) = 1$ so that the first order term appearing in Theorem 1 will not suffice in finding the prior $p(\boldsymbol{\theta}_2)$. We can mimic Theorem 1 to get the second order expansion to $E\left[p(\boldsymbol{\theta}_2 \mid \boldsymbol{X}_n)|\boldsymbol{\theta}_2\right]$. Here is the new Theorem:

**Theorem 2:** For $\beta = -1$, $E\left[p(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)|\boldsymbol{\theta}_2\right]$ can be expressed as:

$$
\begin{aligned}
&E\left[p(\boldsymbol{\theta}_2 \,|\, \boldsymbol{X}_n)|\, \boldsymbol{\theta}_2\right] = n^{\frac{d_2}{2}} \Bigg[ \int K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 + \\
&\frac{1}{n}\Bigg\{ \sum_{j=d_1+1}^{d} \sum_{r=1}^{d} \left[ \int \left( q^{ojr} - \frac{1}{2} I^{jr} \right) K(\boldsymbol{\theta}) \pi_r(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_j(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad - \frac{1}{2} \sum_{j=d_1+1}^{d} \sum_{r=1}^{d} \frac{\partial}{\partial\theta_r} \left( \int I^{jr} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right) \frac{p_j(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad + \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \left[ \int A_{jrs} (q^{ojr} q^{osu} + q^{oju} q^{ors} + q^{ojs} q^{oru}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad - \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \left[ \int A_{jrs} (I^{jr} I^{su} + I^{ju} I^{rs} + I^{js} I^{ru}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \qquad (14) \\
&\quad + \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \left[ \int A_{jrs} k_{jrsu}(\boldsymbol{\theta}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad - \sum_{j=d_1+1}^{d} \sum_{r=d_1+1}^{d} \left[ \int \left\{ I^{jr} - \frac{q^{ojr}}{2} \right\} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_{jr}(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad + \frac{1}{2} \sum_{j=d_1+1}^{d} \sum_{r=d_1+1}^{d} \left[ \int I^{jr} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \right] \frac{p_j(\boldsymbol{\theta}_2) p_r(\boldsymbol{\theta}_2)}{p^2(\boldsymbol{\theta}_2)} \\
&\quad + S(\boldsymbol{\theta}_2) \Bigg\} + o(n^{-1}) \Bigg],
\end{aligned}
$$

where

$$
K(\boldsymbol{\theta}) = \left| I^{22} \right|^{-1/2} (2\pi)^{-d_2/2} 2^{-d_2/2}, \quad A_{jrs} = E\left[ \frac{\partial^3 \log f(X|\boldsymbol{\theta})}{\partial\theta_j \partial\theta_r \partial\theta_s} \right]
$$

$$
(q^{ojr})_{d \times d} = \begin{pmatrix} I^{11} - \frac{1}{2} I^{12} [I^{22}]^{-1} I^{21} & I^{12}/2 \\ I^{21}/2 & I^{22}/2 \end{pmatrix},
$$

$I^{jr}$ is the $jr^{th}$ element of the Fisher information matrix, $k_{jrsu}(\boldsymbol{\theta})$ involves $p(\boldsymbol{\theta}_2)$ and its derivatives, but $S(\boldsymbol{\theta}_2)$ is only a function of $\boldsymbol{\theta}_2$.

The proof of Theorem 2 is long and involved, and is omitted. The details are available from the authors.

Since $\beta = -1$ so that $\beta(1-\beta) = -2$, neglecting all terms which do not involve $p(\boldsymbol{\theta}_2)$ or its derivatives and using the relation $\frac{p_{jr}(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} = \frac{\partial}{\partial\theta_r} \left( \frac{p_j(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \right) + \frac{p_j(\boldsymbol{\theta}_2) p_r(\boldsymbol{\theta}_2)}{p^2(\boldsymbol{\theta}_2)}$, it suffices to maximize up to the second order approximation,

$$
\begin{aligned}
&R^{(-1)}(p(\boldsymbol{\theta}_2)) \approx \frac{1}{2} n^{\frac{d_2}{2}} \int \Bigg[ \int K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 + \\
&\frac{1}{n}\Bigg\{ \sum_{j=d_1+1}^{d} \sum_{r=1}^{d} \Big[ \int \Big( q^{ojr} - \frac{I^{jr}}{2} \Big) K(\boldsymbol{\theta}) \pi_r(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{p_j(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad - \frac{1}{2} \sum_{j=d_1+1}^{d} \sum_{r=1}^{d} \frac{\partial}{\partial \theta_r} \Big( \int I^{jr} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big) \frac{p_j(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad + \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \Big[ \int A_{jrs} (q^{ojr} q^{osu} + q^{oju} q^{ors} + q^{ojs} q^{oru}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad - \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \Big[ \int A_{jrs} (I^{jr} I^{su} + I^{ju} I^{rs} + I^{js} I^{ru}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad + \frac{1}{6} \sum_{1 \le j,r,s \le d} \sum_{u=d_1+1}^{d} \Big[ \int A_{jrs} k_{jrsu}(\boldsymbol{\theta}) K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{p_u(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \\
&\quad + \sum_{j=d_1+1}^{d} \sum_{r=d_1+1}^{d} \Big[ \int \Big\{ \frac{(q^{ojr} - I^{jr})}{2} - \frac{I^{jr}}{2} \Big\} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{\partial p_j(\boldsymbol{\theta}_2)/p(\boldsymbol{\theta}_2)}{\partial \theta_r} \\
&\quad + \sum_{j=d_1+1}^{d} \sum_{r=d_1+1}^{d} \Big[ \int \Big\{ \frac{1}{2}(q^{ojr} - I^{jr}) \Big\} K(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 \Big] \frac{p_j(\boldsymbol{\theta}_2) p_r(\boldsymbol{\theta}_2)}{p^2(\boldsymbol{\theta}_2)} \Bigg\} \Bigg] d\boldsymbol{\theta}_2.
\end{aligned} \tag{15}
$$

Let

$$
\begin{aligned}
\boldsymbol{y}(\boldsymbol{\theta}_2) &= (\boldsymbol{y}_{d_1+1}(\boldsymbol{\theta}_2), \cdots, \boldsymbol{y}_d(\boldsymbol{\theta}_2)) = \left( \frac{p_{d_1+1}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_2)}, \cdots, \frac{p_d(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)} \right) \\
\nabla \boldsymbol{y}(\boldsymbol{\theta}_2) &= \left( \frac{\partial \boldsymbol{y}_{d_1+1}(\boldsymbol{\theta}_2)}{\partial \theta_{d_1+1}}, \cdots, \frac{\partial \boldsymbol{y}_{d_1+1}(\boldsymbol{\theta}_2)}{\partial \theta_d}, \cdots, \frac{\partial \boldsymbol{y}_d(\boldsymbol{\theta}_2)}{\partial \theta_{d_1+1}}, \cdots, \frac{\partial \boldsymbol{y}_d(\boldsymbol{\theta}_2)}{\partial \theta_d} \right).
\end{aligned}
$$

Note that (15) can be expressed as

$$
\int F(\boldsymbol{\theta}_2, \boldsymbol{y}(\boldsymbol{\theta}_2), \nabla \boldsymbol{y}(\boldsymbol{\theta}_2)) d\boldsymbol{\theta}_2, \tag{16}
$$

so we need find $\boldsymbol{y}(\boldsymbol{\theta}_2)$ to maximize the above integral. From Giaquinta (1983), the maximizer should satisfy the Euler-Lagrange equations:

$$
\frac{\partial F}{\partial \boldsymbol{y}_i(\boldsymbol{\theta}_2)} - \sum_{j=d_1+1}^{d} \frac{\partial}{\partial \theta_j} \left( \frac{\partial F}{\partial (\partial \boldsymbol{y}_i / \partial \theta_j)} \right) = 0 \quad i = d_1 + 1, \cdots, d. \tag{17}
$$

Equivalently, the Euler-Lagrange equations are:

$$\begin{aligned}&\sum_{r=d_1+1}^{d}\left[\int I^{ir}K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1\right]\frac{p_r(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2)}\\ =&-\frac{1}{2}\sum_{1\leq j,r,s\leq d}\int A_{jrs}m^{ojr}I^{si}K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1\\ &+\frac{1}{2}\sum_{r=d_1+1}^{d}\frac{\partial}{\partial\theta_r}\int I^{ir}K(\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_1,\quad i=d_1+1,\cdots,d.\end{aligned}\tag{18}$$

Here

$$m^{ojr}=\begin{cases}n^{ojr} & \text{when } j,r\in H_1\\ I^{jr} & \text{other}\end{cases},$$

where $n^{ojr}$ is the $jr^{th}$ element of $I^{12}(I^{22})^{-1}I^{21}$. $H_1$ is the set of indexes of nuisance parameters.

By solving these partial differential equations, one gets the divergence priors. Usually, with multi-dimensional parameters of interest, these equations are so complicated that it is impossible to give a general solution and sometimes, there is no solution to these equations.

In the following, we focus on a special case when the parameter of interest is one-dimensional. In this case, instead of several partial differential equations, we only need to solve one differential equation and easily get a general form of the divergence priors.

When the parameter of interest $\boldsymbol{\theta}_2$ is one dimensional, that is $\boldsymbol{\theta}_2=\theta_d$, then the Euler-Lagrange equation becomes:

$$\begin{aligned}&\left[\int (I^{dd})^{\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1\right]\frac{p'(\theta_d)}{p(\theta_d)}\\ =&-\frac{1}{2}\sum_{1\leq j,r,s\leq d}\int A_{jrs}m^{ojr}I^{sd}(I^{dd})^{-\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1\\ &+\frac{1}{2}\frac{\partial}{\partial\theta_d}\int (I^{dd})^{\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1.\end{aligned}\tag{19}$$

By solving (19), one gets the divergence prior $p(\theta_d)$ which is proportional to

$$Q(\theta_d)^{\frac{1}{2}}\times\exp\left[\int -T(\theta_d)d\theta_d\right],\tag{20}$$

where

$$Q(\theta_d)=\int (I^{dd})^{\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1, T(\theta_d)=\frac{\frac{1}{2}\sum_{1\leq j,r,s\leq d}\int A_{jrs}m^{ojr}I^{sd}(I^{dd})^{-\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1}{\int (I^{dd})^{\frac{1}{2}}\pi(\boldsymbol{\theta}_1|\theta_d)d\boldsymbol{\theta}_1}.\tag{21}$$

Here are several examples to illustrate how to find divergence priors for one dimensional parameter of interest.

**Example 1:** Consider general symmetric location-scale family of distributions with probability density function $f(x|\mu,\sigma) = \frac{1}{\sigma}p(\frac{x-\mu}{\sigma})$ where $p(x) = p(-x)$. Writing $h(x) = \log p(x)$ and noting that $h'(x) = -h'(-x)$, $h''(x) = h''(-x)$ and $h'''(x) = -h'''(-x)$, one gets

$$E\left[\left.\frac{\partial^2 \log f}{\partial \mu^2}\right|\mu,\sigma\right] = -\sigma^{-2}\int h''(x)p(x)dx, \quad E\left[\left.\frac{\partial^2 \log f}{\partial \mu \partial \sigma}\right|\mu,\sigma\right] = 0,$$

$$E\left[\left.\frac{\partial^2 \log f}{\partial \sigma^2}\right|\mu,\sigma\right] = -\sigma^{-2}\Big[1 + 2\int xh'(x)p(x)dx + \int x^2h''(x)p(x)dx\Big],$$

$$E\left[\left.\frac{\partial^3 \log f}{\partial \mu^3}\right|\mu,\sigma\right] = E\left[\left.\frac{\partial^3 \log f}{\partial \mu \partial \sigma^2}\right|\mu,\sigma\right] = 0,$$

$$E\left[\left.\frac{\partial^3 \log f}{\partial \mu^2 \partial \sigma}\right|\mu,\sigma\right] = -\sigma^{-3}\Big[2\int h''(x)p(x)dx + \int xh'''(x)p(x)dx\Big]$$

and

$$E\left[\left.\frac{\partial^3 \log f}{\partial \sigma^3}\right|\mu,\sigma\right] = -\sigma^{-3}\Big[2 + 6\int xh'(x)p(x)dx + 6\int x^2h''(x)p(x)dx + \int x^3h'''(x)p(x)dx\Big].$$

- If $\mu$ is parameter of interest, from (20), the prior should have the following form:

$$p(\mu) \propto \Big[\int \sigma\pi(\sigma|\mu)d\sigma\Big]^{\frac{1}{2}}.$$

- If $\sigma$ is parameter of interest, from (20), the prior should have the following form:

$$p(\sigma) \propto \sigma^{\frac{1}{2}+\frac{1+3\int xh'(x)p(x)dx+3\int x^2h''(x)p(x)dx+\frac{1}{2}\int x^3h'''(x)p(x)dx}{1+2\int xh'(x)p(x)dx+\int x^2h''(x)p(x)dx}}.$$

As special case, recall that for the $N(\mu,\sigma^2)$ distribution, $h'(x) = -x$, $h''(x) = -1$ and $h'''(x) = 0$. Hence $p(\sigma) \propto \sigma^3$.

**Example 2:** Consider the proper dispersion model introduced by Jorgensen (1997). The probability density function of this model is given by

$$f(x|\mu,\lambda) = a(\lambda)c(x)\exp[\lambda t(x,\mu)],$$

where $\mu$ and $\lambda$ are two parameters. Now observe that

$$\frac{\partial \log f}{\partial \mu} = \lambda\frac{\partial t}{\partial \mu}, \quad \frac{\partial \log f}{\partial \lambda} = u(\lambda) + t(x,\mu),$$

where $u(\lambda) = a'(\lambda)/a(\lambda)$. Accordingly,

$$\frac{\partial^2 \log f}{\partial \mu^2} = \lambda\frac{\partial^2 t}{\partial \mu^2}, \quad \frac{\partial^2 \log f}{\partial \mu \partial \lambda} = \frac{\partial t}{\partial \mu}, \quad \frac{\partial^2 \log f}{\partial \lambda^2} = u'(\lambda).$$

Since $E\left(\frac{\partial t}{\partial \mu}|\mu,\lambda\right)=0$, the Fisher information matrix is

$$I(\mu,\lambda)=\text{ Diag }[I_{\mu\mu}(\mu,\lambda),-u'(\lambda)],$$

where $I_{\mu\mu}(\mu,\lambda)=-\lambda E\left[\frac{\partial^2 t}{\partial\mu^2}|\mu,\lambda\right]$. Thus $\mu$ and $\lambda$ are orthogonal in the sense of Cox and Reid (1987). Further, $\frac{\partial^3\log f}{\partial\mu^3}=\lambda\frac{\partial^3 t}{\partial\mu^3}$, $\frac{\partial^3\log f}{\partial\mu^2\partial\lambda}=\frac{\partial^2 t}{\partial\mu^2}$, $\frac{\partial^3\log f}{\partial\mu\partial\lambda^2}=0$, $\frac{\partial^3\log f}{\partial\lambda^3}=u''(\lambda)$.

- When $\mu$ is the parameter of interest, it is easy to get that $I^{dd}=I_{\mu\mu}^{-1}(\mu,\lambda)$, $A_{ddd}=E\left[\frac{\partial^3\log f}{\partial\mu^3}|\mu,\lambda\right]=\lambda E\left[\frac{\partial^3 t}{\partial\mu^3}|\mu,\lambda\right]$ and all the $m^{jr}=0$ except $m^{dd}=I^{dd}$. Hence, by using (21), one gets

$$Q(\mu)=\int I_{\mu\mu}^{-1/2}(\mu,\lambda)\pi(\lambda|\mu)d\lambda,\ \ T(\mu)=\frac{\int\lambda E\left[\frac{\partial^3 t}{\partial\mu^3}|\mu,\lambda\right]I_{\mu\mu}^{-3/2}(\mu,\lambda)}{2Q(\mu)} \tag{22}$$

Then, by using (20), we can get the prior for $\mu$.

- When $\lambda$ is the parameter of interest, one gets that $I^{dd}=-\frac{1}{u'(\lambda)}$, $A_{ddd}=E\left[\frac{\partial^3\log f}{\partial\lambda^3}|\mu,\lambda\right]=u''(\lambda)$ and all the $m^{jr}=0$ except $m^{dd}=I^{dd}$. By using (21), it is easy to check that

$$Q(\lambda)=\left(-\frac{1}{u'(\lambda)}\right)^{1/2},\ \ T(\lambda)=-\frac{u''(\lambda)}{2u'(\lambda)}. \tag{23}$$

Therefore, the prior for $\lambda$ is proportional to

$$\left(-\frac{1}{u'(\lambda)}\right)^{1/4}\times\exp\left[\int\frac{u''(\lambda)}{2u'(\lambda)}d\lambda\right]=(-u'(\lambda))^{1/4} \tag{24}$$

Now, we consider several special cases of the above general result.

**Example 3:** Consider the two-parameter Gamma probability density function

$$f(x|\mu,\lambda)=\exp\left(-\frac{\lambda}{\mu}x\right)\frac{\lambda^\lambda x^{\lambda-1}}{\mu^\lambda}\frac{1}{\Gamma(\lambda)}.$$

Here,

$$a(\lambda)=\frac{\lambda^\lambda}{\Gamma(\lambda)},\quad t(x,\mu)=-\frac{x}{\mu}+\log\frac{x}{\mu}.$$

- When $\mu$ is the parameter of interest, it is easy to get that

$$I_{\mu\mu}=-\lambda E\left[\frac{\partial^2 t}{\partial\mu^2}|\mu,\lambda\right]=-\frac{1}{\mu^2},\ \ E\left[\frac{\partial^3 t}{\partial\mu^3}|\mu,\lambda\right]=\frac{4}{\mu^3}.$$

Then, by using (22), one gets

$$Q(\mu)=\mu\int\lambda^{-\frac{1}{2}}\pi(\lambda|\mu)d\lambda,\ \ T(\mu)=\frac{2}{\mu}.$$

Therefore, from (20), the prior should have the following form:

$$p(\mu) \propto \mu^{-\frac{3}{2}} \left[\int \lambda^{-\frac{1}{2}} \pi(\lambda|\mu) d\lambda\right]^{1/2}.$$

When $\pi(\lambda|\mu)$ is independent of $\mu$ (for example, conditional Jeffreys' general rule prior suggested by Bernardo), the divergence prior for $\mu$ is proportional to $\mu^{-\frac{3}{2}}$.

- When $\lambda$ is the parameter of interest, one gets that $u(\lambda) = 1 + \log\lambda - \frac{d}{d\lambda}\log\Gamma(\lambda)$. Then, from (24), the prior should have the following form:

$$p(\lambda) \propto [-u'(\lambda)]^{\frac{1}{4}}.$$

**Example 4:** Consider the Inverse Gaussian distribution with probability density function

$$f(x|\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right].$$

Here,

$$a(\lambda) = \lambda^{1/2}, \quad t(x,\mu) = -\frac{(x-\mu)^2}{2\mu^2 x}.$$

- When $\mu$ is the parameter of interest, it is easy to get that

$$I_{\mu\mu} = -\lambda E\left[\frac{\partial^2 t}{\partial\mu^2}|\mu,\lambda\right] = \frac{\lambda}{\mu^3}, \quad E\left[\frac{\partial^3 t}{\partial\mu^3}|\mu,\lambda\right] = \frac{6}{\mu^4}.$$

Then, by using (22), one gets

$$Q(\mu) = \mu^{3/2}\int \lambda^{-\frac{1}{2}}\pi(\lambda|\mu)d\lambda, \quad T(\mu) = \frac{3}{\mu}.$$

Therefore, from (20), the prior should have the following form:

$$p(\mu) \propto \mu^{-\frac{9}{4}} \left[\int \lambda^{-\frac{1}{2}} \pi(\lambda|\mu) d\lambda\right]^{1/2}.$$

Similar to Example 3, when $\pi(\lambda|\mu)$ is independent of $\mu$, the divergence prior is proportional to $\mu^{-\frac{21}{4}}$.

- When $\lambda$ is the parameter of interest, one gets that $u(\lambda) = \frac{1}{2\lambda}$. Then, from (24), the prior should have the following form:

$$p(\lambda) \propto \lambda^{-\frac{1}{2}}.$$

**Example 5:** (Fisher von-Mises) The probability density function

$$f(x|\mu,\lambda) = \frac{\exp[\lambda\cos(x-\mu)]}{2\pi I_0(\lambda)}, \quad \text{where } I_0(\lambda) = \frac{1}{2\pi}\int \exp(\lambda\cos x)dx.$$

Then $t(x,\mu) = \cos(x-\mu)$, $a(\lambda) = I_0^{-1}(\lambda)$. Hence, $\frac{\partial t}{\partial \mu} = \sin(x-\mu)$, $\frac{\partial^2 t}{\partial \mu^2} = -\cos(x-\mu)$, $\frac{\partial^3 t}{\partial \mu^3} = \sin(x-\mu)$, so that $E\left(\frac{\partial^3 t}{\partial \mu^3}|\mu,\lambda\right) = E\left(\frac{\partial t}{\partial \mu}|\mu,\lambda\right) = 0$. Further $u(\lambda) = -\frac{I_0'(\lambda)}{I_0(\lambda)}$ and $u'(\lambda) = -\frac{d}{d\lambda}\left[\frac{I_0'(\lambda)}{I_0(\lambda)}\right]$. • When $\mu$ is the parameter of interest, it is easy to get that

$$I_{\mu\mu} = -\lambda E\left[\frac{\partial^2 t}{\partial \mu^2}|\mu,\lambda\right] = \lambda\frac{I_1(\lambda)}{I_0(\lambda)}, \;\; E\left[\frac{\partial^3 t}{\partial \mu^3}|\mu,\lambda\right] = 0.$$

Then, by using (22), one gets

$$Q(\mu) = \int \lambda^{-\frac{1}{2}}\left[\frac{I_0(\lambda)}{I_1(\lambda)}\right]^{1/2}\pi(\lambda|\mu)d\lambda \text{ (which is a constant)}, \;\; T(\mu) = 0.$$

Therefore, from (20), the prior for $\mu$ is uniform distribution.

- When $\lambda$ is the parameter of interest, one gets that $u(\lambda) = -\frac{I_0'(\lambda)}{I_0(\lambda)}$. Then, from (24), the prior should have the following form:

$$p(\lambda) \propto \left[\frac{d}{d\lambda}\left(\frac{I_0(\lambda)}{I_0(\lambda)}\right)\right]^{1/4}.$$

**Example 6:** Let's consider the selection of priors for the famous Neyman-Scott problem (Berger and Bernardo 1992a, 1992b). This problem can be formalized as a fixed effects one-way balanced ANOVA model.

Let $X_{i1},\cdots,X_{ik}|\mu_i$ be mutually independent $N(\mu_i,\sigma^2)$, $i = 1,\cdots,n$, $k \geq 2$. Here, $k$, the number of observations within each treatment $i$ is fixed, while the number of treatments, $n$, can grow to infinity. The Fisher Information matrix is

$$I(\mu_1,\cdots,\mu_n,\sigma^2) = k \text{ Diag}(\sigma^{-2},\cdots,\sigma^{-2},\frac{n\sigma^{-4}}{n}).$$

If we consider all the parameters of equal importance, then one gets Jeffreys' general rule prior

$$\pi_J(\mu_1,\cdots,\mu_n,\sigma^2) \propto (\sigma^2)^{-n/2-1}.$$

The corresponding marginal posterior distribution of $\sigma^2$ is an Inverse Gamma distribution with two parameters equaling to $\frac{nk}{2}, \frac{n(k-1)S}{2}$ respectively. Here $S = \frac{1}{n(k-1)}\sum_{i=1}^n\sum_{j=1}^k(X_{ij}-\bar{X}_i)^2$. Then the posterior mean of $\sigma^2$ is given by $n(k-1)S/(nk-2)$, while the posterior mode is $n(k-1)S/(nk+2)$. Both are inconsistent estimators of $\sigma^2$.

Now we use the expression (20) to construct prior for this problem and show that the corresponding Bayes estimators of $\sigma^2$ are consistent.

Here, $\sigma^2$ is the parameter of interest, while $\mu_1,\cdots,\mu_n$ are nuisance parameters. By (21), one gets

$$Q(\sigma^2) = \sqrt{\frac{2}{nk}\sigma^2}, \;\; T(\sigma^2) = \frac{2}{\sigma^2}.$$

Therefore, the desired prior for $\sigma^2$ obtained by (20) is proportional to $(\sigma^2)^{-3/2}$.

Letting the conditional prior for $(\mu_1, \cdots, \mu_n)$ given $\sigma^2$ equal to the conditional Jeffreys' general rule prior suggested by Bernardo (1979) (in this case, it is the uniform distribution), one gets the two-stage reference prior $\pi_R(\mu_1, \cdots, \mu_n, \sigma^2) \propto (\sigma^2)^{-3/2}$.

The corresponding marginal posterior of $\sigma^2$ is an Inverse Gamma distribution with two parameters equaling to $\frac{n(k-1)+1}{2}$, $\frac{n(k-1)S}{2}$ respectively. Then the posterior mean of $\sigma^2$ is given by $n(k-1)S/(nk-n-1)$, while the posterior mode is $n(k-1)S/(nk-n+3)$. Both are consistent estimators of $\sigma^2$.

Note. Datta and Ghosh (1995a) studied the same problem and developed their two-stage reference prior based on the Kullback-Leibler divergence. Their prior is proportional to $(\sigma^2)^{-1}$ which is slightly different from ours and also leads to consistent Bayes estimators of $\sigma^2$.

**Example 7:** One-way random effects model has been studied by many people through Bayesian approach. Now, we revisit this model and only consider the balanced model.

Let $Y_{ij} = \mu + \alpha_i + e_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$. Here the common mean $\mu$ is an unknown parameter, while $\alpha_i$'s and $e_{ij}$ are mutually independent with $\alpha_i$'s i.i.d $N(0, \sigma_\alpha^2)$ and $e_{ij}$ i.i.d $N(0, \sigma^2)$.

Berger and Bernardo (1992c) first found two-stage reference priors for variance components in this problem. Later Ye (1994) and Datta and Ghosh (1995a, 1995b) found reference priors under different parametrizations. Here, we follow the parametrization used in Ye (1994). Let $\phi = \frac{n\sigma_\alpha^2}{\sigma^2}$. So parameters are $\mu, \sigma^2$ and $\phi$. The likelihood function $L(\mu, \sigma^2, \phi)$ can be written as

$$L(\mu, \sigma^2, \phi) \propto \sigma^{-kn}(1+\phi)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}\left[S_2 + \frac{S_1 + kn(\bar{Y} - \mu)^2}{1+\phi}\right]\right\},$$

where

$$\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}, \quad \bar{Y} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij},$$

$$S_1 = n\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2, \quad S_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2.$$

Then the Fisher information matrix simplifies to

$$I(\mu, \sigma^2, \phi) = \begin{pmatrix} \frac{kn}{(1+\phi)\sigma^2} & 0 & 0 \\ 0 & \frac{kn}{2\sigma^4} & \frac{k}{2(1+\phi)\sigma^2} \\ 0 & \frac{k}{2(1+\phi)\sigma^2} & \frac{k}{2(1+\phi)^2} \end{pmatrix}.$$

The inverse matrix is

$$I^{-1}(\mu, \sigma^2, \phi) = \begin{pmatrix} \frac{\sigma^2(1+\phi)}{kn} & 0 & 0 \\ 0 & \frac{2\sigma^4}{k(n-1)} & -\frac{2(1+\phi)\sigma^2}{k(n-1)} \\ 0 & -\frac{2(1+\phi)\sigma^2}{k(n-1)} & \frac{2n(1+\phi)^2}{k(n-1)} \end{pmatrix}.$$

To illustrate our method, we consider the following two cases.

**1.** $\mu$ is the parameter of interest. The common mean $\mu$ is of great relevance in meta analysis (Morris and Normand 1992).

By (21), one gets

$$Q(\mu) = \int \left(\frac{(1+\phi)\sigma^2}{kn}\right)^{1/2} \pi(\sigma^2, \phi|\mu) d\sigma^2 d\phi \text{ (which is just a constant)}, \quad T(\mu) = 0.$$

Therefore, by using (20), the prior for $\mu$ is uniform distribution. If we take $\pi(\sigma^2, \phi|\mu)$ as the conditional Jeffreys' general prior which is proportional to $(1+\phi)^{-1}\sigma^{-2}$ in this case, then the two-stage reference prior is

$$\pi(\mu, \sigma^2, \phi) \propto (1+\phi)^{-1}\sigma^{-2}.$$

**2.** $\phi$ is the parameter of interest. As pointed out by Ye (1994), the variance ratio $\sigma_\alpha^2/\sigma^2$ is of great interest in genetic studies.

By (21), one gets

$$Q(\phi) = \int \left(\frac{2n(1+\phi)^2}{k(n-1)}\right)^{1/2} \pi(\mu, \sigma^2|\phi) d\sigma^2 d\phi = \sqrt{\frac{2n}{k(n-1)}}(1+\phi), \quad T(\phi) = \frac{2-1/n}{1+\phi}.$$

Therefore, by using (20), the prior for $\phi$ is proportional to $(1+\phi)^{-\frac{3}{2}+\frac{1}{n}}$. If we take $\pi(\mu, \sigma^2|\phi)$ as the conditional Jeffreys' general prior which is proportional to $(1+\phi)^{-1/2}\sigma^{-3}$ in this case, then the two-stage reference prior is

$$\pi(\mu, \sigma^2, \phi) \propto (1+\phi)^{-2+1/n}\sigma^{-3}.$$

## 4. Summary

In this paper, We generalize the idea from Bernardo (1979) to handle the problem of selection of priors in the presence of nuisance parameters. Instead of using Kullback-Leibler divergence which is studied by Bernardo (1979), we use a general divergence criterion to develop objective priors. This general divergence criterion is a family of divergence measures between prior and corresponding posterior including the Kullback-Leibler, Bhattacharyya-Hellinger and the Chi-square divergence. An interesting finding is that with one exception (the Chi-square divergence), for every divergence measure in the general divergence family, the desired divergence prior is the same prior as Bernardo found. Under the Chi-square divergence, we have shown that the objective prior should be the solution to a set of partial differential equations. We also consider a special case when the parameter of interest is one dimensional. In this case, the closed forms of the optimal priors are provided and also several examples are given.

# References

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

Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with application to the multinomial problem, *Biometrika,* **25,** 25-37,

Berger, J. O. and Bernardo, J. M. (1992b). On the development of reference priors (with discussion), *Bayesian Statistics 4,* Eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 35-60, Springer-Verlag, Berlin.

Berger, J. O. and Bernardo, J. M. (1992c). Reference Priors in a Variance Components Problem, *Bayesian Statistics in Statistics and Econometrics,* Eds P. K. Goel and N. S. Iyengar, pp. 177-194, Springer-Verlag, New York.

Berger, J. O. and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model, *Econometric Theory,* **10,** 461-482,

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference, *Journal of the Royal Statistical Society, Series B,* **41,** 113-147.

Bhattacharyya, A. K. (1943). On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society,* **35,** 99-109.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion, *The Annals of Statistics,* **6,** 434-451.

Clarke, B. and Sun, D. (1997). Reference priors under the chi-square distance, *Sankhya, Series A,* **59,** 215-231.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inferece (with discussion), *Journal of Royal Statistical Society, Series B,* **49,** 1-39.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B,* **46,** 440-464.

Datta, G. S. and Ghosh, M. (1995a). Some remarks on noninformative priors, *Journal of the American Statistical Association,* **90,** 1357-1363.

Datta, G. S. and Ghosh, M. (1995b). Hierachical Bayes estimators of the error variance in one-way ANOVA models, *Journal of Statistical Planning and Inference,* **45,** 399-411.

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics,* Springer, New York.

Ghosh, J. K. (1994). *Higher Order Asymptotics.* NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 4.

Ghosh, M. and Liu, R. (2009). Divergence Priors For Multiparameter Models, *Technical Report,* University of Florida

Ghosh, M., Mergel, V. and Liu, R. (2011). A general divergence criterion for prior selection, *Annals of the Institute of Statistical Mathematics,* **63,** 43-58.

Giaquinta, M. (1983). *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems* Princeton University Press, Princeton, New Jersey.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischen Formen von unendlichen vielen Veränderlichen. *Journal für Reine und Angewandte Mathematik,* **136** 210-271.

Jeffreys, H. (1961). *Theory of Probability.* (3rd edition) London: Oxford University Press.

Johnson, R.A. (1970). Asymptotic expansions associated with posterior distribution, *The Annals of Mathematical Statistics,* **41,** 851-864.

Jorgensen, B. (1997). *The Theory of Dispersion Models.* Chapman and Hall, New York.

Liu, R., Chakrabarty, A., Samanta, T., Ghosh, J. K. and Ghosh, M. (2014). On Divergwnce Measures Leading to Jeffreys and Other Reference Priors. *Bayesian Analysis* **9**, 331-370.

Morris, C. N. and Normand, S. L. (1992). Hierarchical Models for Combining Information and for Meta-analysis, *Bayesian Statistics 4,* Eds J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, pp. 321-344, Oxford. Univ. Press, New York.

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations, *Econometrika,* **16,** 1-32.

Renyi, A. (1961). Measures of Entropy and Information. In *Proceedings of the 4th Berkely Symposium„ V1.* Editors: J. Neyman and L. Lecam, pp 547-561.

Ye, K. and Berger, J. (1991). Noninformative priors for inference in exponential regression mdels, *Biometrika,* **78,** 645-656.

Ye, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model, *Journal of Statistical Planning and Inference,* **41,** 267-280.