Prior-free Bayesian Inference based on Information-theoretic Approach

Xing-Qi Jiang
Asahikawa University, Department of Economics

1. Introduction

Although Bayesian statistical analysis has many advantages, a number of objections was made to it because its prior-dependency. As pointed out by Akaike (1983), in practical applications of Bayesian analysis the available prior information is not usually sufficient to completely specify the prior distribution. Ultimately, the traditional Bayesian inference often depends on objective priors. This may be a standard objection to the Bayesian approach.

Such problem was emphasized at the early stage of development of Bayesian approach. The pioneers in the accomplishment of Bayesian analysis such as Bayes (see Bayes 1763) and Laplace (see Laplace 1812) developed Bayesian procedure using uniform prior distribution for objectivity. However, sometimes such procedure encounters difficulties because of a lack of invariance under transformation of unknown parameters (see Jaynes 1983). Fisher did not accept Bayesian procedure mainly due to the use of uniform prior distribution, he attempted to make statistical inference by proposing the concept of inverse probability and his fiducial approach (see Fisher 1930, 1933, and 1935). Fisher's fiducial approach aimed to achieve advantages of the Bayesian approach without the assumption of a prior distribution. Unfortunately, Fisher's fiducial approach ultimately cannot be achieved as a systematized methodology for statistical inference.

The main concern with the use of uniform prior distribution is that it may be variant under a transformation of unknown parameters. This caused Jeffreys to develop his ignorance prior distribution (see Jeffreys 1946). The definition of Jeffreys prior is based on the concept of invariance of the distribution by a transformation of unknown parameters. Lindley (1956) applied Shannon entropy to introduce an information-theoretic analysis of the structure of Bayesian modeling. Zellner (1971, 1977) and Bernardo (1979) developed objective Bayesian procedures using the maximal data information prior distribution and the reference prior distribution respectively. These works prompted the work by Akaike (1983) on the problem of specifying a prior distribution over a finite number of data distributions.

Another problem in practical applications of objective Bayesian procedures is that they often utilize improper prior distributions, and so do not automatically have desirable Bayesian properties, such as coherency (see Stone 1976). Also, a poor choice of improper priors can even lead to marginalization paradoxes (see Dawid, Stone and Zidek 1973) and improper posteriors (see Ye and Berger 1991). Thus recent studies of objective Bayesian procedures are mostly about to ensure that such problems do not arise (see Berger and Bernardo 1992, Bernardo 1999).

In this paper, we attempt to contribute to objective Bayesian theory by developing a new approach called *prior-free inference*. This paper is organized as follows. In the next section we give procedural and mathematical background and motivation of present study. In Section 3 we show results obtained from information-theoretic optimization, and in Section 4 a procedure for this approach is given. Finally, concluding remarks are given in Section 5.

2. Procedural and mathematical background and motivation

2.1 Procedural and mathematical Background

In the present paper, we are concerned with statistical inference for a k-dimensional vector, $\theta=(\theta_1,\theta_2,\ldots,\theta_k)^t$, of continuous parameters based on a sample, $X(1:n)=\{X_1,X_2,\ldots,X_n\}$, of size $n\ (>k)$ with each X_i being univariate continuous random variable. Generally, we assume a model density $f_{X(1:n)}(x(1:n)|\theta)$ for $X(1:n)$ given θ. Then, the conditional model density for X_i given $x(1:i-1)$ and θ is obtained as

$$f_{X_i}(x_i|x(1:i-1),\theta)=\frac{f_{X(1:i)}(x(1:i)|\theta)}{f_{X(1:i-1)}(x(1:i-1)|\theta)} \tag{1}$$

for i=2,3,..., k, where $f_{X(1:i)}(x(1:i)|\theta)$ is given by

$$f_{X(1:i)}(x(1:i)|\theta)=\int f_{X(1:n)}(x(1:n)|\theta)dx(i+1:n).$$

If we define $X(1:0)=x(1:0)=\phi$, the empty set, and $f_{X(1:0)}(x(1:0)|\theta)=1$, then equation (1) holds also for $i=1$, i.e.

$$f_{X_1}(x_1|x(1:0),\theta)=\frac{f_{X(1:1)}(x(1:1)|\theta)}{f_{X(1:0)}(x(1:0)|\theta)}=f_{X_1}(x_1|\theta).$$

Thus, the model density for $X(1:k)$ given θ can be expressed as follows:

$$f_{X(1:k)}\ (x(1:k)|\theta)=f_{X_1}(x_1|\theta)f_{X_2}(x_2|x(1:1),\theta)\cdots f_{X_k}(x_k|x(1:k-1),\theta). \tag{2}$$

Incidentally, when $X(1:n)$ is a random sample, we have $f_{X_i}(x_i|x(1:i-1),\theta)=f_{X_i}(x_i|\theta)$.

For the sake of further discussion, we introduce the definition of “support.” The concept of support can be found in Zacks (1971, p.41) and Lehmann and Casella (1998, p.16). For a density function $g_X(x)$ for X, its it support is defined by the set

$$\mathcal{S}(gX)=\{x; gX(x)>0\}.$$

Further, for a conditional density function $fX(x|\upsilon)$ of X given υ, its support is defined by the set

$$\mathcal{S}(fX|\upsilon)=\{x; fX(x|\upsilon)>0\}.$$

Note that the concept of support can also be applied to multivariate densities.

In Bayesian approach, the parameter vector θ can be regarded as a set of given values of random parameters $\Theta=(\Theta_1,\Theta_2,\cdots,\Theta_k)^t$. So, an initial probability distribution called the prior distribution for Θ is required. Let $\pi(\theta)$ be a prior density and let $f_\Theta(\theta|x(1:k))$ be the corresponding posterior density or post data density for Θ given $x(1:k)$. We have the following relation between the prior density and the post data density:

$$f_\Theta(\theta|x(1:k))h_{X(1:k)}(x(1:k))=\pi(\theta)f_{X(1:k)}(x(1:k)|\theta), \tag{3}$$

where $h_{X(1:k)}(x(1:k))$ denotes the marginal density of $X(1:k)$.

Denote by $\mathcal{S}(\pi)$ and $\mathcal{S}(h_{X(1:k)})$ the supports of $\pi(\theta)$and $h_{X(1:k)}(x(1:x))$, respectively. Let $\mathcal{S}(f_\Theta|x(1:k))$ be the support of $f_\Theta(\theta|x(1:k))$ for $x(1:k)\in\mathcal{S}(h_{X(1:k)})$, and let $\mathcal{S}(f_{X(1:k)}|\theta)$ be that of $f_{X(1:k)}(x(1:k)|\theta)$ for $\theta\in\mathcal{S}(\pi)$. So, from equation (3) we have $\mathcal{S}(f_\Theta|x(1:k))\subseteq\mathcal{S}(\pi)$, because a necessary condition for $f_\Theta(\theta|x(1:k))>0$ is $\pi(\theta)>0$. Further, from equation (3) we can see that $f_{X(1:k)}(x(1:k)|\theta)=0$ is a necessary result if we maintain the assumption that $\pi(\theta)>0$ for $\theta\in\mathcal{S}^C(f_\Theta|x(1:k))$ and $x(1:k)\in\mathcal{S}(h_{X(1:k)})$, where $\mathcal{S}^C(f_\Theta|x(1:k))$ denotes the complement of $\mathcal{S}(f_\Theta|x(1:k))$ in $\mathcal{S}(\pi)$. It is unnecessary for a likelihood oriented inference. Thus, we bring the equality

$$\mathcal{S}(\pi)=\mathcal{S}(f_\Theta|x(1:k)) \tag{4}$$

as a fundamental assumption in the present paper. Similarly, we can also assume that $\mathcal{S}(h_{X(1:k)})$ = $\mathcal{S}(f_{X(1:k)}|\theta)$.

Suppose that for a prior density $\pi(\theta)$, the post data density $f_\Theta(\theta\,|x(1:k))$ is a proper density in which $f_\Theta(\theta|x(1:k))$ is integrable with respect to θ and satisfies the following condition:

$$\int_{\mathcal{S}(\pi)}f_\Theta(\theta|x(1:k))d\theta=1. \tag{5}$$

Then, from equations (3) and (5) we can obtain the marginal density of $X(1:k)$ by

$$h_{X(1:k)}(x(1:k))=\int_{\mathcal{S}(\pi)}\pi(\theta)f_{X(1:k)}(x(1:k)|\theta)d\theta. \tag{6}$$

Equation (3) is a fundamental relation in Bayesian inference. It is called the product rule of probability and is generally regarded as an axiom in probability theory. In Zellner (1988) the product rule was proved from a viewpoint of information-theoretic approach and was concluded as the Bayesian information processing rule. It is obvious that equations (5) and (6) are equivalent under the fundamental relation (3). From the fundamental relation (3), we obtain the post data density by the following equation:

$$f_\Theta(\theta|x(1:k)) = \frac{\pi(\theta) f_{X(1:k)}(x(1:k)|\theta)}{h_{X(1:k)}(x(1:k))}, \tag{7}$$

which is called Bayes' theorem (see for example Box and Tiao 1973).

Bayes' theorem allows us to continuously update information about Θ as more observations are obtained. Now, assume that we have the model density $f_{X(k+1:n)}(x(k+1:n)|x(1:k),\theta)$ for $X(k+1:n)$ $= \{X_{k+1}, X_{k+2}, \ldots, X_n\}$ given $x(1:k)$ and θ. Then, we can obtain the post data density for Θ given $x(1:n)$ as

$$f_\Theta(\theta|x(1:n)) = \frac{f_\Theta(\theta|x(1:k)) f_{X(k+1:n)}(x(k+1:n)|x(1:k),\theta)}{h_{X(k+1:n)}(x(k+1:n)|x(1:k))}, \tag{8}$$

where $h_{X(k+1:n)}(x(k+1:n)|x(1:k)) = \int_{S(\pi)} f_\Theta(\theta|x(1:k)) f_X(x(k+1:n)|x(1:k),\theta)d\theta$. The expression (8) is precisely of the same form as equation (7) except that $f_\Theta(\theta|x(1:k))$, the post data density for Θ given $x(1:k)$, plays the role of the prior density for the succeeding observations $x(k+1:n)$. Obviously, this process can be repeated times. Thus, Bayes' theorem describes the process of updating the distribution of Θ as learning from data, and shows how information about Θ is continuously modified as new data become available. Therefore, we call $f_\Theta(\theta|x(1:k))$ and $f_\Theta(\theta|x(1:n))$ the initial post data density and the final post data density for Θ, respectively.

From the above observations, we can see that the cruxes of the traditional Bayesian analysis are the model for observed data and the prior density for the parameters. They are as two inputs for Bayesian information processing (Zellner 1988), but it may be true that the former should precedes the latter, because without model there can be no parameters hence there can be no prior. In scientific research, setting up hypotheses is the main subject for researchers, the model for observed data may be constructed along with the hypotheses. However, as usual case the construction of priors may be more difficult.

2.2 Motivation

It can be seen from the discussion in the previous subsection that a feature of the traditional Bayesian approach is the prior-dependency. It leads to a difficulty in applications of Bayesian inference when the prior information is unavailable. This difficulty may be fatal for most situations of scientific research and is the main cause of criticism to Bayesian statistics. As pointed out by Fine (1999), "The Bayesian methodology, while enjoying good properties (e.g., admissibility and consistency), is peculiar, in that it requires the user to postulate a prior distribution that is basically as complex as the quantities being inferred, if not more so." There are a number of studies on evaluating priors by using model and observed data, e.g., Zellner (1971, 1977, 1991), Bernardo (1979), Akaike (1980), Jaynes (1983), Chuaqui (1991), Berger and Bernardo (1992), Berger (1994), Li and Vitanyi (1997). Such approaches have provided solutions to mitigate the difficulty in the traditional Bayesian analysis.

In order to overcome the difficulty of the traditional Bayesian analysis caused by a lack of prior information, a quite different approach to objective Bayesian inference will be introduced in the present paper. The feature of this new approach is that it is free of dependence on a prior distribution. Thus, we call Bayesian inference based on this approach the *prior-free inference*. An outline of this approach is shown in Jiang (2000) by the name of self-concluding inference, and it was further developed in Jiang (2002). The key idea of the prior-free inference is as follows. The presupposition of the prior-free inference is that we have a model density for the observed data given parameters. As the first step of the procedure, we derive an initial post data density $f_\Theta(\theta|x(1:k))$ of Θ given $x(1:k)$, from the given model density for $X(1:k)$ directly. Then, in the second step we apply $f_\Theta(\theta|x(1:k))$ as the prior density for the observations of the remaining sample $X(k+1:n)$ to obtain the final post data density $f_\Theta(\theta|x(1:n))$ by using Bayes' theorem.

3. Prior-free Bayesian inference

3.1 Probability integral transformations

First of all, we define a set of probability integral transformations as

$$\varphi_i(x_i, x(1:i-1), \theta) = \int_{-\infty}^{x_i} f_{X_i}(t|x(1:i-1), \theta)\, dt \tag{9}$$

for $i = 1, 2, \ldots, k$. Obviously, the quantity $\varphi_i(x_i, x(1:i-1), \theta)$ defined by equation (9) is a function of both $x(1:k)$ and θ. Here, we consider the situation that θ is fixed. So, when $x(1:i-1)$ is also given, $\varphi_i(x_i, x(1:i-1), \theta)$ becomes a cumulative distribution function for the model density of X_i given $x(1:i-1)$. It is a function of x_i only and we denote it by

$$\omega_i = \omega_i(x_i) = \varphi_i(x_i, x(1:i-1), \theta)\Big|_{x(1:i-1),\theta} \tag{10}$$

for $i = 1, 2, \ldots, k$. When the value of x_i in equation (10) is replaced with the corresponding random quantity

X_i for $i=1,2,\dots,k$, a set of random variables, say $W_i=\omega_i(X_i)\,(i=1,2,\dots,k)$, can be defined.

Let $\boldsymbol{S}\left(f_{X_i}\middle|x(1:i-1),\theta\right)$ be the support of $f_{X_i}\left(x_i\middle|x(1:i-1),\theta\right)$ for $i=1,2,\dots,k$. The following assumption will be required and can be satisfied for most continuous random variables.

(A1) Given $x(1:i-1)$, the conditional model density $f_{X_i}\left(x_i\middle|x(1:i-1),\theta\right)$ is a continuous function of $\theta\in\boldsymbol{S}(\pi)$ and $x_i\in\boldsymbol{S}\left(f_{X_i}\middle|x(1:i-1),\theta\right)$ for $i=1,2,\dots,k$.

Then the following fact can be verified under the assumption A1. When $x(1:i-1)$ and θ are fixed, the support of the density $f_{W_i}\left(\omega_i\middle|\omega(1:i-1),\theta\right)$ for W_i given $\omega(1:i-1)=\{\omega_1,\omega_2,\dots,\omega_{i-1}\}$ and θ is [0, 1] for $i=1,2,\dots,k$. Hence, for $x_i\in\boldsymbol{S}\left(f_{X_i}\middle|x(1:i-1),\theta\right)$, we have

$$\frac{\partial\omega_i}{\partial x_i}=f_{X_i}\left(x_i\middle|x(1:i-1),\theta\right)>0\quad(i=1,2,\dots,k).\tag{11}$$

Thus, equation (10) is a one-to-one transformation from $\boldsymbol{S}\left(f_{X_i}\middle|x(1:i-1),\theta\right)$ to [0, 1] for $i=1,2,\dots,k$. Therefore, under the assumption A1 we have

$$f_{X_i}\left(x_i\middle|x(1:i-1),\theta\right)=f_{W_i}\left(\omega_i\middle|\omega(1:i-1),\theta\right)\left|\frac{\partial\omega_i}{\partial x_i}\right|,\tag{12}$$

for $x_i\in\boldsymbol{S}\left(f_{X_i}\middle|x(1:i-1),\theta\right)$ and $i=1,2,\dots,k$, where $\left|\frac{\partial\omega_i}{\partial x_i}\right|$ denotes the absolute value of $\frac{\partial\omega_i}{\partial x_i}$.

The following lemmas and corollary are easily proved.

Lemma 1. Under the assumption A1, the density $f_{W_i}\left(\omega_i\middle|\omega(1:i-1),\theta\right)$ for W_i given $\omega(1:i-1)$ and θ is as

$$f_{W_i}\left(\omega_i\middle|\omega(1:i-1),\theta\right)=1,\quad\omega_i\in[0,\,1]\tag{13}$$

for $i=1,2,\dots,k$.

Lemma 2. Under the assumption A1, we have

$$f_{X(1:k)}\left(x(1:k)\middle|\theta\right)=\prod_{i=1}^{k}\left|\frac{\partial\omega_i}{\partial x_i}\right|.$$

Corollary 1. Under the assumption A1, the joint density for $W(1:k)=\{W_1, W_2,\dots, W_k\}$ given θ is as follows:

$$f_{W(1:k)}(\omega(1:k)|\theta)=1.$$

3.2 Definition of inferential functions

Here, we consider the property of the probability integral transformations for a set of given observations $x(1:k)$. In this case, the quantity $\varphi_i(x_i,x(1:i-1),\theta)$ defined by equation (9) becomes a function of θ only for $i=1,2,\dots,k$. It is expressed by

$$z_i=z_i(\theta)=\varphi_i(x_i,x(1:i-1),\theta)|_{x(1:i)}=\varphi_i(x(1:i),\theta)|_{x(1:i)} \quad (i=1,2,\dots,k). \tag{14}$$

Further, when θ is replaced with Θ, a set of random variables, say

$$Z=(Z_1,Z_2,\dots,Z_k)^{\mathrm{t}}=(z_1(\Theta),z_2(\Theta),\dots,z_k(\Theta))^{\mathrm{t}}, \tag{15}$$

is newly defined. The functions defined by equation (15) together with equation (14) are important for the procedure of prior-free inference, we call them the *inferential functions*.

Let $fz(z|x(1:k))$ be a post data density for Z given $x(1:k)$, and let $\mathcal{S}(fz|x(1:k))$ denote its support. The inferential functions can be regarded as a set of transformations from $\mathcal{S}(\pi)$ to $\mathcal{S}(fz|x(1:k))$ with

$$J=\left(\frac{\partial z_i}{\partial \theta_j}\right) \tag{16}$$

being the Jacobian matrix. Further, when both $x(1:i)$ and θ are given z_i is the cumulative probability, thus we can see that $\mathcal{S}(fz|x(1:k))\subseteq [0,\ 1]\times[0,\ 1]\times\cdots\times[0,\ 1]$.

For the usual case that $\mathcal{S}(\pi)$ is not empty, we call the inferential functions are *informative* under the observations $x(1:k)$ if they satisfy the following assumptions:

(A2) The partial differential, $\dfrac{\partial z_i}{\partial \theta_j}$, is a continuous function of θ at all points of $\mathcal{S}(\pi)$ for $i,j=1,2,\dots,k$.

(A3) The Jacobian matrix defined by equation (16) is a nonsingular matrix at all points of $\mathcal{S}(\pi)$.

If the inferential functions are informative, then they play the role of one-to-one transformations between $\mathcal{S}(\pi)$ and $\mathcal{S}(fz|x(1:k))$. Hence they have a property shown by the following lemma.

Lemma 3. If the inferential functions are informative, then the quantity defined by

$$\lambda = \int_{\boldsymbol{S}(\pi)} \left|\det(J)\right| d\theta \tag{17}$$

satisfies the inequality $0 < \lambda \leq 1$, where $\det(J)$ denotes the determinant of the Jacobian matrix J defined by equation (16), and $\left|\det(J)\right|$ does its absolute value.

Proof. Under the assumptions A2 and A3, we have

$$\lambda = \int_{\boldsymbol{S}(fz|x(1:k))} dz$$

from equation (17). Thus, the proof is completed from the fact that $\boldsymbol{S}(fz|x(1:k)) \subseteq$ [0, 1]×[0, 1]×⋯×[0, 1].

It is important that if the inferential functions are informative under $x(1:k)$, then the initial post data density $f_\Theta(\theta|x(1:k))$ for Θ given $x(1:k)$ can be defined in terms of the initial post data density $fz(z|x(1:k))$ for Z by

$$f_\Theta(\theta|x(1:k)) = fz(z|x(1:k))\left|\det(J)\right|. \tag{18}$$

Thus, we can determine $f_\Theta(\theta|x(1:k))$ through $fz(z|x(1:k))$.

3.3 Determination of initial post data density

In this subsection, we show how to determine the initial post data density $fz(z|x(1:k))$ for Z, or equivalently the initial post data density $f_\Theta(\theta|x(1:k))$ for Θ, by utilizing an information-theoretic approach.

Let $s(x)$ and $t(x)$ be two kinds of densities for X, the Kullback-Leibler information of $s(x)$ with respect to $t(x)$ is defined by

$$I_{KL}(s;t) = \int \ln\left\{\frac{s(x)}{t(x)}\right\} s(x)\,dx. \tag{19}$$

Following Kullback (1959), a necessary condition (but not sufficient) to guarantee the finiteness of $I_{KL}(s;t)$ is that the probability measures defined on $s(x)$ and $t(x)$ are absolutely continuous with respect to one another. Further, for the same purpose it is also required that both the densities $s(x)$ and $t(x)$ are proper, that is, $\int s(x)dx = 1$ and $\int t(x)dx = 1$. It is well-known that $I_{KL}(s;t) \geq 0$ and $I_{KL}(s;t) = 0$ if and only if

$t(x)=s(x)$ almost everywhere. So, $I_{KL}(s;t)$ is a functional that measures the "distance" between $s(x)$ and $t(x)$. That is, $I_{KL}(s;t)$ is a measure to evaluate how $t(x)$ is divergent from $s(x)$ in which $s(x)$ is regarded as the "standard density". Note that the Kullback-Leibler information can also be defined for multivariate densities.

Lindley (1956) utilized the Kullback-Leibler information in Bayesian inference in order to introduce his criterion functional. By the notation of the present paper, a kind of Lindley's criterion functional is as follows:

$$F_L(\pi)=\int_{\boldsymbol{S}(\pi)\times \boldsymbol{S}(h_{X(1:k)})} \ln\left\{\frac{f_\Theta(\theta|x(1:k))}{\pi(\theta)}\right\} \times f_\Theta(\theta|x(1:k))h_{X(1:k)}(x(1:k))d\theta dx(1:k) \tag{20}$$

which measures the missing information about the parameters θ under the condition that the model density $f_{X(1:k)}(x(1:k)|\theta)$ is given. Bernardo (1979) developed his reference prior procedure that derives a prior density as a solution to maximizing $F_L(\pi)$. In Bernardo (1979), such prior density is regarded as a prior that describes vague initial information about θ.

Obviously, Lindley's criterion functional can be rewritten as follows:

$$F_L(\pi)=\int_{\boldsymbol{S}(\pi)\times \boldsymbol{S}(h_{X(1:k)})} \ln\left\{\frac{f_\Theta(\theta|x(1:k))h_{X(1:k)}(x(1:k))}{\pi(\theta)h_{X(1:k)}(x(1:k))}\right\} \times f_\Theta(\theta|x(1:k))h_{X(1:k)}(x(1:k))d\theta dx(1:k).$$

So, it is obvious that

$$F_L(\pi)=I_{KL}(s;t) \tag{21}$$

by putting

$$s(x(1:k),\theta)=f_\Theta(\theta|x(1:k))h_{X(1:k)}(x(1:k)), \tag{22}$$

$$t(x(1:k),\theta)=\pi(\theta)h_X(x(1:k)). \tag{23}$$

As shown in equations (22) and (23), $s(x(1:k),\theta)$ denotes the joint density of $X(1:k)$ and Θ under the assumption that $X(1:k)$ and Θ are correlated, and $t(x,y)$ denotes another joint density under the assumption that $X(1:k)$ and Θ are independent of each other. That is, Lindley's criterion functional measures the distance between $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$ by regarding $s(x(1:k),\theta)$ as the standard density.

Lindley (1956) concluded that $F_L(\pi)$ is as a concave functional of $\pi(\theta)$. By against, we have the following theorem.

Theorem 1. For given model density $f_{X(1:k)}\left(x(1:k)\middle|\theta\right)$, if the initial post data density $f_\Theta\left(\theta\middle|x(1:k)\right)$ defined by equation (7) with a fixed prior density $\pi(\theta)\in\boldsymbol{S}(\pi)$ being a continuous function of $\boldsymbol{\theta}$, then under the assumptions A1 Lindley's criterion functional defined by (20) equals zero. That is, $F_L(\pi)=0$.

Before proving Theorem 1, we give the following lemma.

Lemma 4. Under the conditions of Theorem 1, we have

$$\int_{\boldsymbol{S}(\pi)\times\boldsymbol{S}(h_{X(1:k)})} f_\Theta(\theta|x(1:k))f_{X(1:k)}(x(1:k)|\theta)dx(1:k)d\theta=1. \tag{24}$$

Proof. From the equations (6), (7) and Lemma 2, we have

$$\begin{aligned}
&\int_{\boldsymbol{S}(\pi)\times\boldsymbol{S}(h_{X(1:k)})} f_\Theta(\theta|x(1:k))f_{X(1:k)}(x(1:k)|\theta)dx(1:k)d\theta\\
&=\int_{\boldsymbol{S}(\pi)\times\boldsymbol{S}(h_{X(1:k)})}\frac{f_{X(1:k)}\left(x(1:k)\middle|\theta\right)\pi(\theta)}{h_{X(1:k)}\left(x(1:k)\right)}\prod_{i=1}^{k}\left|\frac{\partial\omega_i}{\partial x_i}\right|dx(1:k)d\theta\\
&=\int_{\boldsymbol{S}(\pi)}\left(\int_0^1\cdots\int_0^1\frac{f_{X(1:k)}\left(x(1:k)\middle|\theta\right)\pi(\theta)}{h_{X(1:k)}\left(x(1:k)\right)}d\omega(1:k)\right)d\theta\\
&=\int_0^1\cdots\int_0^1\left(\int_{\boldsymbol{S}(\pi)}\frac{f_{X(1:k)}\left(x(1:k)\middle|\theta\right)\pi(\theta)}{h_{X(1:k)}\left(x(1:k)\right)}d\theta\right)d\omega(1:k)\\
&=\int_0^1\cdots\int_0^1 d\omega(1:k)=1
\end{aligned}$$

which completes the proof.

Proof of Theorem 1. From the properties of the Kullback-Leibler information and equation (21), the following inequality is straightforward:

$$F_L(\pi)\geq 0. \tag{25}$$

On the other hand, by using Bayes' theorem and Jensen's inequality we have

$$F_L(\pi)=\int_{S(\pi)\times S(h_{X(1:k)})}\ln\left\{\frac{f_\Theta(\theta|x(1:k))}{\pi(\theta)}\right\}f_{X(1:k)}(x(1:k)|\theta)\pi(\theta)dx(1:k)d\theta$$
$$\le\ln\left(\int_{S(\pi)\times S(h_{X(1:k)})}\frac{f_\Theta(\theta|x(1:k))}{\pi(\theta)}f_{X(1:k)}(x(1:k)|\theta)\pi(\theta)dx(1:k)d\theta\right)$$
$$=\ln\left(\int_{S(\pi)\times S(h_{x(1:k)})}f_\Theta(\theta|x(1:k))f_{X(1:k)}(x(1:k)|\theta)dx(1:k)d\theta\right).$$

Further, by applying Lemma 4 to the above inequality we obtain

$$F_L(\pi)\le 0. \tag{26}$$

By combining inequality (25) with inequality (26), we have $F_L(\pi)=0$, which completes the proof of Theorem 1.

Theorem 1 implies that it may be difficult to specify a prior as a solution to maximization of Lindley's criterion functional. It prompts us to utilize the following newly-introduced criterion functional to determine an initial post data density:

$$F(f_\Theta)=\int_{S(\pi)\times S(h_{X(1:k)})}\ln\left\{\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))}\right\}\pi(\theta)h_{X(1:k)}(x(1:k))d\theta dx(1:k). \tag{27}$$

When we define the Kullback-Leibler information of $\pi(\theta)$ with respect to $f_\Theta(\theta|x(1:k))$ for given $x(1:k)$ by

$$I_{KL}(\pi;f_\Theta|x(1:k))=\int_{S(\pi)}\ln\left\{\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))}\right\}\pi(\theta)d\theta, \tag{28}$$

the criterion functional $F(f_\Theta)$ can be expressed as follows

$$F(f_\Theta)=\int_{S(h_{X(1:k)})}I_{KL}(\pi;f_\Theta|x(1:k))h_{X(1:k)}(x(1:k))dx(1:k). \tag{29}$$

That is, our newly-introduced criterion functional is the expected information of $\pi(\theta)$ with respect to $f_\Theta(\theta|x(1:k))$.

Perhaps, the intention to specify a prior by maximizing the Lindley's criterion functional is to make inference by using the traditional Bayesian approach with the most non-informative prior. Contrastively, the intention to obtain an initial post data density by maximizing our newly-introduced criterion functional is that we want to make inference by using the information from observations to the maximum for a given model density. It is easy to show that $F(f_\Theta) = I_{KL}(t;s)$ under the use of equations (22) and (23). Our criterion functional $F(f_\Theta)$ measures also the distance between $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$ by regarding $t(x(1:k),\theta)$ as the standard density. Thus, it can be seen that the greater the value of $F(f_\Theta)$ the larger the information about Θ from $x(1:k)$. Therefore, we determine the post data density $f_\Theta(\theta)$ for Θ given $x(1:k)$ directly by maximizing our criterion functional $F(f_\Theta)$.

From equation (18), we can see that the post data densities $f_\Theta(\theta|x(1:k))$ and $f_Z(z|x(1:k))$ correspond with each other for a given model density. So, our criterion functional $F(f_\Theta)$ is also a functional of $f_Z(z|x(1:k))$ and the maximization of $F(f_\Theta)$ with respect to $f_\Theta(\theta|x(1:k))$ will be equivalent to that with respect to $f_Z(z|x(1:k))$.

Theorem 2. Under equation (7), if the inferential functions are informative, then the criterion functional $F(f_\Theta)$ is maximized when

$$f_Z(z|x(1:k)) = c \tag{30}$$

for a given model density of $X(1:k)$, where c denotes a constant.

Proof. From equation (28), we have

$$I_{KL}(\pi; f_\Theta|x(1:k)) = \int_{S(\pi)} \ln\left\{\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))}\right\}\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))} f_\Theta(\theta|x(1:k))\,d\theta.$$

By applying equation (7), the above equation can be rewritten as

$$I_{KL}(\pi; f_\Theta|x(1:k)) = \int_{S(\pi)} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_\Theta(\theta|x(1:k))\,d\theta, \tag{31}$$

where

$$\phi(x(1:k),\theta) = \frac{h_{X(1:k)}(x(1:k))}{f_{X(1:k)}(x(1:k)|\theta)}.$$

For a given model density $f_{X(1:k)}(x(1:k)|\theta)$, if we fix the prior density $\pi(\theta)$, then the marginal density $h_{X(1:k)}(x(1:k))$ for $X(1:k)$ is fixed, hence the function $\phi(x(1:k),\theta)$ in equation (31) is also fixed. Thus, we can only maximize $I_{KL}(\pi; f_\Theta|x(1:k))$ through the initial post data density $f_\Theta(\theta|x(1:k))$. By applying equation (18) to equation (31), we obtain the following relation:

$$I_{KL}(\pi; f_\Theta | x(1:k)) = \int_{S(\pi)} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_Z(z|x(1:k))|\det(J)|d\theta$$
$$= \int_{S(f_Z|x(1:k))} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_Z(z|x(1:k))dz.$$

It is obvious that $I_{KL}(\pi; f_\Theta | x(1:k))$ is maximized when equation (30) holds. Moreover, from equation (29), we can see that $F(f_\Theta)$ is maximized as long as $I_{KL}(\pi; f_\Theta | x(1:k))$ is maximized. Thus, the theorem is proved.

Note that the maximizer of the criterion functional defined by equation (27) is free of dependence on a prior distribution. The following corollary is straightforward from Theorem 2, and equations (17) and (18).

Corollary 2. Under equation (7), if the inferential functions are informative, then the criterion functional $F(f_\Theta)$ is maximized when

$$f_\Theta(\theta | x(1:k)) = \frac{|\det(J)|}{\lambda}, \tag{32}$$

for a given model density of $X(1:k)$, where λ is defined by equation (17).

4. General procedure

In this section, we give a general procedure for prior-free inference based on the observations $x(1:n) = \{x_1, x_2, \ldots, x_n\}$ for the sample $X(1:n)$.

Suppose we can ensure that the inferential functions are informative under $x(1:k)$ by permuting the observations appropriately. Firstly, we give the post data density $f_\Theta(\theta | x(1:k))$ for Θ given $x(1:k)$ by using the equation (32). Then, we utilize $f_\Theta(\theta | x(1:k))$ as a prior density for the remaining observations $x(k+1:n)$ of the sample, and obtain the final post data density for Θ by using equation (8).

5. Concluding remarks

A new approach named by the prior-free inference to Bayesian inference was introduced for developing objective Bayesian analysis. The feature of this new approach is that it is essentially a Bayesian method but it may be free of dependence on a prior distribution for unknown parameters. So, this approach does not only have advantages of the Bayesian approach but also can avoid the difficulties when we have no prior information.

An important problem is the relation between our prior-free inference and Fisher's fiducial approach. It can be seen that if a model has a sufficient statistic for a single parameter, they can lead to the same result, otherwise our prior-free inference is better than Fisher's fiducial approach. Further, it is well-known that Fisher's fiducial approach is difficult for the multivariate parameter case.

Nowadays, most objective Bayesian analysis procedures use Jeffreys prior. However, a number of objections can be made to the Jeffreys prior, the most important of which is that it depends on the form of the observed data. Such objection is reasonable, perhaps, because the prior distribution should only represent the information prior to the observed data, it can not be influenced by the data. Also, sometimes, the Bayesian procedure using the Jeffreys prior will violate the Likelihood Principle, and it is difficult to apply Jeffreys' procedure to the multivariate parameter case. Such difficulties can be overcome by the use of the procedure proposed in this paper.

References

Akaike, H. (1980), Likelihood and the Bayes procedure, in: Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and A.F.M. Smith (eds), *Bayesian Statistics*, University Press, Valencia, pp.143-166.

Akaike, H. (1983), On minimum information prior distributions, *Ann. Inst. Statist. Math.*, Vol.35, pp.139-149.

Bayes, T. (1763), An essay towards solving a problem in the doctrine of chances, *Philos. Trans. Roy. Soc.*, Vol.53, pp.370-418.

Berger, J.O. (1994), An overview of robust Bayesian analysis, *Test*, Vol.3, pp.5-124.

Berger, J.O. and J.M. Bernardo (1992), On the development of reference priors (with discussion), in: Bernardo, J.M., Berger, J.O., Dawid, A.P. and A.F.M. Smith (eds), *Bayesian Statistics 4*, Oxford University Press, Oxford, pp.35-60.

Bernardo, J.M. (1979), Reference posterior distributions for Bayesian inference (with discussion), *J. Roy. Statist. Soc.* (Ser. B), Vol.41, pp.113-147.

Bernardo, J.M. (1999), Nested hypothesis testing: the Bayesian reference criterion (with discussion), in: Bernardo, J.M., Berger, J.O., Dawid, A.P. and A.F.M. Smith (eds), *Bayesian Statistics 6*, Oxford University Press, Oxford, pp.101-130.

Box, G.E.P. and G.C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Massachusetts.

Chuaqui, R. (1991), *Truth, Possibility and Probability: New Logical Foundations of Probability and Statistical inference*, North-Holland, Amsterdam.

Dawid, A.P., Stone, M. and J.V. Zidek (1973), Marginalization paradoxes in Bayesian and structural inference (with discussion), *J. Roy. Statist. Soc.* (Ser. B), Vol.35, pp.189-233.

Fine, T.L. (1999), Foundations of probability (update), in: Kotz, S. (ed), *Encyclopedia of Statistical Sciences,* *Up. Vol.3*, John Wiley and Sons, New York, pp.246-254.

Fisher, R.A. (1930), Inverse probability, *Proc. Camb. Phil. Soc.*, Vol.26, pp.528-535.

Fisher, R.A. (1933), The concepts of inverse probability and fiducial probability referring to unknown parameters, *Proc. Roy. Phil. Soc.*(Ser. A), Vol.193, pp.343-348.

Fisher, R.A. (1935), The fiducial argument in statistical inference, *Ann. Eugenics*, Vol.6, pp.391-398.

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London* (Ser. A), Vol.186, pp.453-461.

Jaynes, E.T. (1983), *Papers on Probability, Statistics, and Statistical Physics* (Rosenkrantz, R.D., ed), Kluwer, Dordrecht.

Jiang, X.Q. (2000), A general procedure of statistical inference based on information theory, in: Tokuyama, M. and H.E. Stanley (eds), *Statistical Physics,* American Institute of Physics, Melville, pp.642-644.

Jiang, X.Q. (2002), A new approach to objective Bayesian analysis, *The Journal of Asahikawa University*, No.53, pp.1-25.

Kullback, S. (1959), *Information Theory and Statistics*, John Wiley and Sons, New York.

Laplace, P.S. (1812), *Thēorie Analytique des Probabilitēs*, Courcier, Paris.

Lehmann, E.L. and G. Casella (1998), *Theory of Point Estimation* (Second Edition), Springer-Verlag, New York.

Li, M. and P. Vitanyi (1997), *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York.

Lindley, D.V. (1956), On a measure of the information provided by an experiment, *Ann. Math. Statist.*, Vol.27, pp.986-1005.

Stone, M. (1976), Strong inconsistency from uniform priors (with discussion), *J. Amer. Statist. Assoc.*, Vol.71, pp.114-125.

Ye, K.Y. and J.O. Berger (1991), Non-informative priors for inferences in exponential regression models, *Biometrika*, Vol.78, pp.645-656.

Zacks, S. (1971), *The Theory of Statistical Inference*, John Wiley and Sons, New York.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John-Wiley and Sons, New-York.

Zellner, A. (1977), Maximal data information prior distributions, in: Aykac, A. and C. Brumat (eds), *New Developments in the Applications of Bayesian Methods*, North-Holland, Amsterdam, pp.211-232.

Zellner, A. (1988), Optimal information processing and Bayes's theorem (with discussion), *The American Statistician*, Vol.42, pp.278-294.

Zellner, A. (1991), Bayesian methods and entropy in economics and econometrics, in: Grandy, Jr.W.T. and L.H. Schick (eds), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, pp.17-31.