Nonlinear Dynamics and Systems Theory, **7** (1) (2007) 1–21

Prior-free Inference for Objective Bayesian Analysis and Model Selection

Koki Kyo *

School of Agriculture,
Obihiro University of Agriculture and Veterinary Medicine,
Inada-cho, Obihiro, Hokkaido 080-8555, Japan

Received: April 18, 2005; Revised: December 25, 2005

Abstract: A new approach to Bayesian inference, named the *prior-free inference*, is introduced for developing objective Bayesian analysis based on information-theoretic approach. This new approach is essentially a Bayesian method but it does not depend on a prior distribution for unknown parameters. Thus, this approach not only has the advantages of the Bayesian approach but also can avoid the difficulty, the traditional Bayesian approach encounters due to a lack of prior information. Several examples are illustrated to show the procedure and the performance of the prior-free inference. A new information criterion, named *prior-free information criterion* (PFIC), is introduced as an extension of the procedure of the prior-free inference. Then, minimum PFIC method for model selection is developed based on the use of PFIC. Simulation results show that the minimum PFIC method performs very well.

Keywords: *Non-informative priors; prior-free inference; objective Bayesian analysis; model selection; information criterion.*

Mathematics Subject Classification (2000): 62B10, 62F15.

1 Introduction

A necessary condition of the traditional Bayesian analysis is the use of a prior distribution. As pointed out by Akaike [3], however, in practical applications of Bayesian analysis the available prior information is not usually sufficient to completely specify the prior distribution. For that reason, various procedures of objective Bayesian inference using non-informative or ignorance priors have been developed.

The pioneers in the accomplishment of Bayesian analysis such as Bayes and Laplace developed Bayesian procedure using uniform prior distribution for objectivity [4, 25].

* Corresponding author: x.q.jiang@m2.dion.ne.jp

However, sometimes such procedure encounters difficulties because of a lack of invariance under transformation of unknown parameters [15]. 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 [12, 13, 14]. Essentially, Fisher's fiducial approach is somewhat in the category of Bayesian, but it is not necessary to suppose a prior distribution. Unfortunately, Fisher's fiducial approach ultimately cannot be achieved as a systematized methodology for statistical inference.

Criticisms of the use of uniform prior distribution caused Jeffreys to develop his ignorance prior distribution [16]. The definition of Jeffreys prior is based on the concept of invariance of the distribution by a transformation of unknown parameters. Lindley applied Shannon entropy to introduce an information-theoretic analysis of the structure of Bayesian modeling [28]. Zellner and Bernardo developed objective Bayesian procedures using the maximal data information prior distribution and the reference prior distribution respectively [33, 34, 7]. These work prompted the work by Akaike on the problem of specifying a prior distribution over a finite number of data distributions [3].

The main concern with objective Bayesian procedures is that they often utilize improper prior distributions, and so do not automatically have desirable Bayesian properties, such as coherency [31]. Also, the use of improper priors may lead to some difficulties of utilizing information-theoretic approach to identification of priors. Thus recent studies of objective Bayesian procedures are mostly about to ensure that such problems do not arise [6, 8].

In this paper, we attempt to contribute to objective Bayesian theory by developing a new approach which is called *prior-free inference*. The remainder of the paper is organized as follows. In Section 2 we explain the procedural and mathematical background and motivation of the present study. In Section 3 we show the procedure of the prior-free inference and related theoretic results. In Section 4 we illustrate the procedure and the performance of the prior-free inference by several examples. In Section 5 we develop a methodology for model selection based on the prior-free inference. Finally, concluding remarks are given in Section 6.

2 Settings and motivation

2.1 Settings

In the present paper, we attempt to introduce a new approach to Bayesian inference for a vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^{\mathsf{t}}$, of k continuous parameters. Let $X(1:n) = \{X_1, X_2, \ldots, X_n\}$ be a sample of size n with each X_i being univariate continuous random variable, where $n > k$. Generally, suppose we have a statistical model of $X(1:n)$ given θ that is defined by a joint probability density $f_{X(1:n)}(x(1:n)|\theta)$. Based on $f_{X(1:n)}(x(1:n)|\theta)$ we can obtain a model density of X_i in the conditional density form, $f_{X_i}(x_i|x(1:i-1),\theta)$, given the observations $x(1:i-1) = \{x_1, x_2, \ldots, x_{i-1}\}$ of $X(1:i-1)$ for $i = 1, 2, \ldots, n$. Thus, by defining $f_{X_1}(x_1|x(1:0),\theta) = f_{X_1}(x_1|\theta)$, the model density $f_{X(1:n)}(x(1:n)|\theta)$ can be expressed by

$$f_{X(1:n)}(x(1:n)|\theta) = f_{X_1}(x_1|\theta) f_{X_2}(x_2|x(1:1),\theta) \cdots f_{X_n}(x_n|x(1:n-1),\theta). \quad (1)$$

For the sake of further discussion, we introduce the definition of "support". The concept of support can be found in [26] and [32]. For a density function $u(x)$ of X, its

support is defined by the set $\boldsymbol{\mathcal{S}}(u) = \{x; u(x) > 0\}$. Further, for a conditional density function $v(x|y)$ of X given y, its support is defined by the set $\boldsymbol{\mathcal{S}}(v|y) = \{x : v(x|y) > 0\}$.

In Bayesian approach, the parameter vector θ can be regarded as a vector of given values of k random variables, say $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_k)^{\mathsf{t}}$. It is required to set up an initial probability distribution, called the prior distribution, for Θ. Let $\pi(\theta)$ be a prior density, and denote by $f_\Theta(\theta|x(1:k))$ 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:

$$f_\Theta(\theta|x(1:k))h(x(1:k)) = \pi(\theta)f_{X(1:k)}(x(1:k)|\theta), \tag{2}$$

where $h(x(1:k))$ denotes the marginal density of $X(1:k)$.

Let $\boldsymbol{\mathcal{S}}(\pi)$ and $\boldsymbol{\mathcal{S}}(h)$ be the supports of $\pi(\theta)$ and $h(x(1:x))$, respectively. Denote by $\boldsymbol{\mathcal{S}}(f_\Theta|x(1:k))$ the support of $f_\Theta(\theta|x(1:k))$ for $x(1:k) \in \boldsymbol{\mathcal{S}}(h)$, and denote by $\boldsymbol{\mathcal{S}}(f_{X(1:k)}|\theta)$ that of $f_{X(1:k)}(x(1:k)|\theta)$ for $\theta \in \boldsymbol{\mathcal{S}}(\pi)$. For a likelihood oriented inference, it is unnecessary to consider a value of $\theta \in \boldsymbol{\mathcal{S}}(\pi)$ that leads to $f_{X(1:k)}(x(1:k)|\theta) = 0$. So, from equation (2) we can bring the equality $\boldsymbol{\mathcal{S}}(\pi) = \boldsymbol{\mathcal{S}}(f_\Theta|x(1:k))$ for $x(1:k) \in \boldsymbol{\mathcal{S}}(h)$. Similarly, we can also assume that $\boldsymbol{\mathcal{S}}(h) = \boldsymbol{\mathcal{S}}(f_{X(1:k)}|\theta)$ for $\theta \in \boldsymbol{\mathcal{S}}(\pi)$. Suppose that both of the prior density $\pi(\theta)$ and the post data density $f_\Theta(\theta|x(1:k))$ are proper. Then, we can obtain the marginal density of $X(1:k)$ as

$$h(x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} f_{X(1:k)}(x(1:k)|\theta)\pi(\theta)d\theta, \tag{3}$$

which is also a proper density. From equation (2), we obtain the post data density by

$$f_\Theta(\theta|x(1:k)) = \frac{f_{X(1:k)}(x(1:k)|\theta)\pi(\theta)}{h(x(1:k))}, \tag{4}$$

which is called Bayes' theorem (see [9]).

Bayes' theorem allows us to continuously update information about Θ as more observations are obtained. Now, let $f_{X(k+1:n)}(x(k+1:n)|x(1:k),\theta)$ be the model density 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

$$f_\Theta(\theta|x(1:n)) = \frac{f_{X(k+1:n)}(x(k+1:n)|x(1:k),\theta)f_\Theta(\theta|x(1:k))}{g(x(k+1:n)|x(1:k))}, \tag{5}$$

where $g(x(k+1:n)|x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} f_{X(k+1:n)}(x(k+1:n)|x(1:k),\theta)f_\Theta(\theta|x(1:k))d\theta$. The expression (5) is precisely of the same form as equation (4) except that $f_\Theta(\theta|x(1:k))$ 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. As pointed out by Zellner [35], information processing based on Bayes' theorem does not cause loss of information. In this paper, we call $f_\Theta(\theta|x(1:k))$ and $f_\Theta(\theta|x(1:n))$ the initial and the final post data density, respectively.

Bayesian approach gives a basis for inference not only on unknown parameters but also on any unobserved random variable that follows a probability distribution depending on the parameters. In the concrete, for unobserved random variables, say Y, that follow the model density $f_Y(y|x(1:n),\theta)$, the predictive density $f_Y(y|x(1:n))$ of Y is given by

$$f_Y(y|x(1:n)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} f_Y(y|x(1:n),\theta)f_\Theta(\theta|x(1:n))d\theta. \tag{6}$$

From the above observations, we can see that the cruxes of the traditional Bayesian analysis are the model density for observed data and the prior density for the parameters. The model density and the prior density are as two inputs for Bayesian information processing [35], but it may be true that the model density should precedes the prior density, because without model density there can be no parameters hence it is not necessary to consider a prior density. In scientific research, setting up hypotheses is the main subject for researchers, the model (or a set of contending models) for observed data may be constructed along with the hypotheses. However, it may be more difficult to have knowledge about the parameters in the constructed model before analyzing the observed data.

2.2 Motivation

It can be seen from the discussion in Subsection 2.1 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 it is also the main cause of criticism to Bayesian statistics. As pointed out by [11], "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 [33, 34, 36], Bernardo [7], Akaike [2], Jaynes [15], Chuaqui [10], Berger and Bernardo [6], Berger [5], Li and Vitanyi [27]. Such approaches have provided solutions to mitigate the difficulty of the traditional Bayesian analysis.

In order to overcome the difficulty of the traditional Bayesian analysis caused by a lack of prior information, a new approach to objective Bayesian analysis will be introduced in the present paper. The main feature of this approach is that it is free of dependence on a prior distribution. Thus, we call Bayesian inference based on this approach *prior-free inference*. Contrastively, we call Bayesian inference beginning with construction of priors the traditional Bayesian approach. An outline of the prior-free inference is shown in [17] by the name of self-concluding inference, and it was further developed in [18]. Main results on information-theoretic approach to the prior-free inference were given in [19], and an application of the prior-free approach to estimation and identification of regression models was given in [20]. 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. As the first stage 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 stage 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.

The similarity between the prior-free inference approach and the reference priors approach is that both of these two approaches are developed based on an information-theoretic viewpoint. As will be mentioned in Section 3, however, for an improper prior density the Lindley's criterion functional, which lays the foundations of the reference priors approach, cannot be well-defined. Unfortunately, in objective Bayesian analysis the prior is obtained frequently in an improper form. This difficulty is avoided by introducing a new criterion functional which is utilized as the foundations of the prior-free inference approach.

Now, the model selection is always an important problem in statistical analysis. When several contending models are constructed, it is required to evaluate each model and select

one as the best among them. In the present paper, a methodology for model selection is also developed as a natural extension of the prior-free inference.

3 Prior-free inference

3.1 Definition of inferential functions

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

$$\varphi_i(x(1:i),\theta) = \int_{-\infty}^{x_i} f_{X_i}(t|x(1:i-1),\theta)dt \tag{7}$$

for $i = 1, 2, \ldots, k$. Obviously, the quantity $\varphi_i(x(1:i),\theta)$ defined by equation (7) is a function of $x(1:k)$ and θ.

In the case that $x(1:k)$ are given, the quantity $\varphi_i(x(1:i),\theta)$ defined by equation (7) becomes a function of θ only, so we express it as follows:

$$z_i = z_i(\theta) = \varphi_i(x(1:i),\theta)|_{x(1:i)} \quad (i = 1, 2, \ldots, k). \tag{8}$$

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

$$Z = (Z_1, Z_2, \ldots, Z_k)^{\mathtt{t}} = (z_1(\Theta), z_2(\Theta), \ldots, z_k(\Theta))^{\mathtt{t}}, \tag{9}$$

is defined. The functions defined by equation (9) together with equations (7) and (8) are important for the procedure of the prior-free inference, we call them the *inferential functions*.

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

$$J = (\frac{\partial z_i}{\partial \theta_j}) \tag{10}$$

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

For given $x(1:k)$ we call inferential functions *informative* if they satisfy the following conditions:

(C1) The partial differential, $\frac{\partial z_i}{\partial \theta_j}$, is a continuous function of θ at all points of $\boldsymbol{\mathcal{S}}(\pi)$ for $i, j = 1, 2, \ldots, k$.

(C2) The Jacobian matrix defined by equation (10) is a nonsingular matrix at all points of $\boldsymbol{\mathcal{S}}(\pi)$.

When inferential functions are informative, they play the role of one-to-one transformations between $\boldsymbol{\mathcal{S}}(\pi)$ and $\boldsymbol{\mathcal{S}}(f_Z|x(1:k))$. Thus, they have a property shown by the following lemma (see Appendix A for proof):

Lemma 3.1 *If the inferential functions are informative, then the quantity defined by*

$$\lambda = \int_{\boldsymbol{\mathcal{S}}(\pi)} |\det(J)| d\theta \tag{11}$$

satisfies the inequality $0 < \lambda \leq 1$, *where* $\det(J)$ *denotes the determinant of the Jacobian matrix defined by equation* (10), *and* $|\det(J)|$ *denotes its absolute value.*

We can classify the informative inferential functions into two types. For the quantity λ defined by equation (11), the informative inferential functions are called *fully informative* if $\lambda = 1$, and they are called *partially informative* if $0 < \lambda < 1$. It can be verified that if the inferential functions are fully informative, then $\boldsymbol{\mathcal{S}}(f_Z|x(1:k)) = [0,1] \times [0,1] \times \cdots \times [0,1]$; and if they are partially informative, then $\boldsymbol{\mathcal{S}}(f_Z|x(1:k)) \subset [0,1] \times [0,1] \times \cdots \times [0,1]$. If the inferential functions are informative under $x(1:k)$, then the initial post data density $f_\Theta(\theta|x(1:k))$ for Θ can be defined in terms of the post data density $f_Z(z|x(1:k))$ for Z by

$$f_\Theta(\theta|x(1:k)) = f_Z(z|x(1:k)) \; |\det(J)|. \tag{12}$$

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

3.2 Determination of initial post data density

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

For random variable Y, which is possibly multivariate, let $u(y)$ and $v(y)$ be two density functions, the Kullback-Leibler information of $u(y)$ with respect to $v(y)$ is defined by

$$I_K(u;v) = \int \ln\{\frac{u(y)}{v(y)}\}u(y)dy. \tag{13}$$

It is well-known that $I_K(u;v) \geq 0$, and $I_K(u;v) = 0$ if and only if $v(y) = u(y)$ almost everywhere. $I_K(u;v)$ is as a functional of $u(y)$ and $v(y)$ that measures the "distance" between $u(y)$ and $v(y)$ by regarding $v(y)$ as the *reference distribution*. If the reference distribution $v(y)$ is improper and $u(y)$ is proper, then the probability measures defined on $u(y)$ and $v(y)$ cannot be absolutely continuous with respect to one another, hence $I_K(u;v)$ cannot be finite (see [24]). Thus, $I_K(u;v)$ must be infinite as long as $v(y)$ is improper.

Lindley applied the Kullback-Leibler information to Bayesian inference in order to introduce his criterion functional [28]. By the notation, an expression of Lindley's criterion functional is given by

$$F_L(\pi|f_{X(1:k)}) = \int_{\boldsymbol{\mathcal{S}}(h_{X(1:k)})} I_K^C(f_\Theta;\pi|x(1:k))h_{X(1:k)}(x(1:k))dx(1:k), \tag{14}$$

which measures the missing information about Θ for a given model density. In equation (14),

$$I_K^C(f_\Theta;\pi|x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\frac{f_\Theta(\theta|x(1:k))}{\pi(\theta)}\}f_\Theta(\theta|x(1:k))d\theta \tag{15}$$

is the Kullback-Leibler information between $f_\Theta(\theta|x(1:k))$ and $\pi(\theta)$ given $x(1:k) \in \boldsymbol{\mathcal{S}}(h)$. Bernardo [7] developed his reference priors approach that derives a prior density as a solution to maximizing $F_L(\pi|f_{X(1:k)})$. In [7], such solution is regarded as a prior that describes vague initial knowledge about θ.

Obviously, by definition we have

$$F_L(\pi|f_{X(1:k)}) = I_K(s;t), \tag{16}$$

where

$$s(x(1:k),\theta) = f_\Theta(\theta|x(1:k))h(x(1:k)), \quad (17)$$
$$t(x(1:k),\theta) = \pi(\theta)h(x(1:k)). \quad (18)$$

As shown in equations (17) and (18), $s(x(1:k),\theta)$ denotes the joint density for $X(1:k)$ and Θ under the assumption that $X(1:k)$ and Θ are correlated, and $t(x(1:k),\theta)$ is that for $X(1:k)$ and Θ under the assumption that $X(1:k)$ and Θ are independent of each other. So, Lindley's criterion functional measures the distance between $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$ by regarding $t(x(1:k),\theta)$ as the reference distribution. In the traditional Bayesian approach, if the model density is given, then both of the initial post data density and the marginal density for $X(1:k)$ are as functionals of the prior density, hence both of $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$ are functionals of the prior density $\pi(\theta)$. Therefore, the Lindley's criterion functional $F_L(\pi|f_{X(1:k)})$ is as a functional of the prior density.

A result given in [19] shows that it may be difficult to specify a prior as a solution to maximizing the Lindley's criterion functional. This fact prompts us to introduce another criterion functional for specifying an initial post data density. The newly-introduced criterion functional is defined by

$$F(f_\Theta,\pi|f_{X(1:k)}) = \int_{\boldsymbol{\mathcal{S}}(h_{X(1:k)})} I_K^C(\pi; f_\Theta|x(1:k))h_{X(1:k)}(x(1:k))dx(1:k), \quad (19)$$

where

$$I_K^C(\pi; f_\Theta|x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))}\}\pi(\theta)d\theta \quad (20)$$

defines the Kullback-Leibler information between $\pi(\theta)$ and $f_\Theta(\theta|x(1:k))$ given $x(1:k) \in \boldsymbol{\mathcal{S}}(h)$. It is obvious that

$$F(f_\Theta,\pi|f_{X(1:k)}) = I_K(t;s) \quad (21)$$

under the definitions in equations (17) and (18). The criterion functional $F(f_\Theta,\pi|f_{X(1:k)})$ measures the distance between $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$ by regarding $s(x(1:k),\theta)$ as the reference distribution. In the prior-free inference, we consider the criterion functional $F(f_\Theta,\pi|f_{X(1:k)})$ as a functional not only for the prior density but also for the initial post data density because we attempt to determine the initial post data density directly by maximizing $F(f_\Theta,\pi|f_{X(1:k)})$ for a given model density and any fixed prior density.

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 vague prior. Contrastively, the intention to obtain an initial post data density by maximizing the newly-introduced criterion functional is that we attempt to make post data inference by using the information contained in $x(1:k)$ to the maximum for a given model density and any fixed prior density that is regarded as a non-informative prior. Obviously, the greater the value of $F(f_\Theta,\pi|f_{X(1:k)})$ the larger the information about Θ contained in $x(1:k)$. Therefore, in order to obtain an initial post data density that has maximal information contained in $x(1:k)$, we derive the initial post data density directly by maximizing $F(f_\Theta,\pi|f_{X(1:k)})$. As a theoretical finding, we have the following theorem (see Appendix B for proof):

Theorem 3.1 *Under equation* (4)*, if the inferential functions are informative, then the criterion functional* $F(f_\Theta, \pi | f_{X(1:k)})$ *may have the following two maximizers:*

$$f_\Theta^{(1)}(\theta | x(1:k)) = \frac{1}{\psi}, \tag{22}$$

$$f_\Theta^{(2)}(\theta | x(1:k)) = \frac{1}{\lambda} \, |\det(J)|, \tag{23}$$

for a given model density of $X(1:k)$ *and any fixed prior density that is proper, where* $\psi = \int_{\boldsymbol{\mathcal{S}}(\pi)} d\theta$ *is a constant, and* λ *is calculated by using equation* (11)*.*

Note that the both of these two maximizers of the criterion functional $F(f_\Theta, \pi | f_{X(1:k)})$ are free of dependence on the prior density.

Now, we have to choose one from the above alternative solutions to maximizing the criterion functional $F(f_\Theta, \pi | f_{X(1:k)})$. We employ here the concept of information. For given $x(1:k)$ the information of the initial post data density $f_\Theta(\theta | x(1:k))$ is defined by

$$I(f_\Theta | x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{f_\Theta(\theta | x(1:k))\} f_\Theta(\theta | x(1:k)) d\theta. \tag{24}$$

$I(f_\Theta | x(1:k))$ defined by equation (24) can be regarded as the negative conditional entropy of Θ with respect to $f_\Theta(\theta | x(1:k))$. The greater value of $I(f_\Theta | x(1:k))$ means that we have larger value of information to predict the value of Θ based on $x(1:k)$. It is desirable to find an initial post data density that maximizes the criterion functional $F(f_\Theta, \pi | f_{X(1:k)})$, and leads to a larger value of $I(f_\Theta | x(1:k))$. The following theorem gives us a strategy of determining the initial post data density (see Appendix C for proof):

Theorem 3.2 *Under the condition that the initial post data density is proper, we have*

$$I(f_\Theta^{(2)} | x(1:k)) \geq I(f_\Theta^{(1)} | x(1:k)), \tag{25}$$

where $I(f_\Theta^{(1)} | x(1:k))$ *and* $I(f_\Theta^{(2)} | x(1:k))$ *denote the values of information* $I(f_\Theta | x(1:k))$ *corresponding to equations* (22) *and* (23)*, respectively.*

Theorem 3.2 together with Theorem 3.1 implies that it is a better strategy to determine the initial post data density by using equation (23).

3.3 General procedure

Suppose we have observations $x(1:n)$ for a sample $X(1:n)$ of size n, and the model density for $X(1:n)$ is given by equation (1). Assume that we can ensure that the inferential functions are informative under $x(1:k)$ by an appropriate permutation of the observations $x(1:n)$. Based on the results obtained in the previous subsection, we obtain a general procedure for the prior-free inference as follows:

Firstly, we calculate the initial post data density $f_\Theta(\theta | x(1:k))$ by using equation (23) together with equation (11). Then, we utilize $f_\Theta(\theta | x(1:k))$ as the prior density for the remaining observations $x(k+1:n)$, and obtain the final post data density $f_\Theta(\theta | x(1:n))$ by using equation (5). Finally, if it is necessary we compute the predictive density for an unobserved random quantity Y that has the model density $f_Y(y | x(1:n), \theta)$ by using equation (6).

The reason to carry out the prior-free inference by using the two stage constructions of the post data density is as follows: In the stage of determining the initial post density, there may be information loss due to a lack of prior information. The information loss can be minimized by using the proposed procedure. In the stage of calculating the final post density, the information contained in the additional observations $x(k+1:n)$ can be fully employed, because the use of Bayes' theorem. Thus, it is desirable to save the observations for the second stage as many as possible. It should be emphasized that the number k of observations used in the first stage is the minimum requirement for ensuring the inferential functions to be informative.

3.4 Comparison between criterion functionals

It can be seen that the newly-introduced criterion functional, defined by equation (19) together with equation (20), lays the foundations of the frior-free inference. To show the necessity for introducing it instead of the Lindley's criterion functional, we compare the properties of these two criterion functionals as follows:

Firstly, as is shown in Theorem 3.1 and Theorem 3.2, the newly-introduced criterion functional is concave with respect to the initial post data density $f_\Theta(\theta|x(1:k))$ for a given model density and any fixed prior density that is proper. It was shown in [19], however, Lindley's criterion functional identically equals zero under some regular conditions. So, it seems to be difficult to specify a prior density by maximizing Lindley's criterion functional.

Secondly, the Bernardo's reference prior approach may lead to improper priors when at least one end point of the support of the prior density is not finite. In such case, a difficulty will arise because the Lindley's criterion functional cannot be well-defined. But this difficulty does not arise in the proposed approach because the maximizer of the newly-introduced criterion functional is free of independence on a prior, so the newly-introduced criterion functional can be defined well on any fixed prior density as long as it is proper.

Finally, as mentioned in Subsection 3.2 both of $s(x(1:k),\theta)$ and $t(x(1:k),\theta)$, defined by equations (17) and (18) respectively, are functionals of the prior density $\pi(\theta)$, so from equation (16) we can see that the Lindley's criterion functional is a more intricate functional of the prior. Thus, its maximization may be complicated. On the other hand, for a given model density and any fixed prior, $t(x(1:k),\theta)$ does not depend on the initial post data density. Thus, from equation (21) it is can be seen that the newly-introduced criterion functional is defined as a functional of the initial post data density with a simple structure, so that it can be easy to be manipulated.

3.5 Special procedure for separable models

Let $U(1:n)=\{U_1,U_2,\ldots,U_n\}$ be a sample for a random variable U. Suppose $U(1:n)$ follows model density $f_{U(1:n)}(u(1:n)|\theta)$ with θ being a k-dimensional vector of parameters. We consider partition of the sample, $U(1:n)=\{U(1:m),U(m+1:n)\}$, and partition of the parameter vector, $\theta=\{\theta^{(1)},\theta^{(2)}\}$, with the dimension of $\theta^{(1)}$ being ℓ $(<k)$ for $\ell \le m < n$ and $k-\ell < n-m$. If the model density for $U(1:n)$ can be expressed by the form

$$f_{U(1:n)}(u(1:n)|\theta)=f_{U(1:m)}(u(1:m)|\theta^{(1)},\theta^{(2)})f_{U(m+1:n)}(u(m+1:n)|\theta^{(2)}), \qquad (26)$$

then we say that the model density $f_{U(1:n)}(u(1:n)|\theta)$ is separable. The feature of the model in equation (26) is that the model density $f_{U(m+1:n)}(u(m+1:n)|\theta^{(2)})$ for $U(m+1:n)$, depends only on $\theta^{(2)}$.

We can obtain the post data density $f_{\Theta^{(2)}}(\theta^{(2)}|u(m+1:n))$ for $\Theta^{(2)}$, given $u(m+1:n)$, and obtain the post data density $f_{\Theta^{(1)}}(\theta^{(1)}|u(1:m),\theta^{(2)})$ for $\Theta^{(1)}$, given $u(1:m)$ and $\theta^{(2)}$ by using the procedure of the prior-free inference separately. Then, the post data density for Θ can be obtained successively by

$$f_{\Theta}(\theta|u(1:n)) = f_{\Theta^{(1)}}(\theta^{(1)}|u(1:m),\theta^{(2)}) f_{\Theta^{(2)}}(\theta^{(2)}|u(m+1:n)).$$

Further, when the sample $U(1:n)$ for U is obtained from another sample, say $X(1:n)$, for random variable X through a one-to-one transformation

$$U(1:n) = \psi(X(1:n)), \tag{27}$$

the model density of $U(1:n)$ can be derived from that of $X(1:n)$ by

$$f_{U(1:n)}(u(1:n)|\theta) = f_{X(1:n)}(x(1:n)|\theta)|(\frac{\partial u_i}{\partial x_j})|^{-1}, \tag{28}$$

where $(\frac{\partial u_i}{\partial x_j})$ denotes the Jacobian matrix of the transformation (27). If the model density $f_{U(1:n)}(u(1:n)|\theta)$ in equation (28) can be expressed by the separable form expressed by equation (26), then we say the model density for $X(1:n)$ separable.

Sometimes, we can simplify the process of obtaining inferential results through a separated form for a separable model. For illustration we show the following example:

Example 3.1 Consider $X(1:n)$ as a sample that each X_i is independently distributed with the same normal density

$$f_{X_i}(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}, \quad -\infty < x_i < \infty \quad (i = 1, 2, \ldots, n),$$

where $\theta = (\mu, \sigma)^{\mathtt{t}}$ denotes the parameter vector with μ and σ being the mean and the standard deviation. We obtain the values $u(1:n)$ for $U(1:n)$ by using the transformation

$$(u_1, u_2, \ldots, u_n)^{\mathtt{t}} = H(x_1, x_2, \ldots, x_n)^{\mathtt{t}}, \tag{29}$$

where H denotes the Helmert matrix defined by

$$H = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{1\times 2}} & -\frac{1}{\sqrt{1\times 2}} & 0 & \cdots & \cdots & 0 \\ \frac{1}{\sqrt{2\times 3}} & \frac{1}{\sqrt{2\times 3}} & -\frac{2}{\sqrt{2\times 3}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & 0 \\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \cdots & \cdots & \frac{1}{\sqrt{(n-1)n}} & -\frac{n-1}{\sqrt{(n-1)n}} \end{bmatrix}.$$

It can be verified that from the model density of $X(1:n)$, the first part $U(1:1) = U_1$ of the sample $U(1:n)$ follows the model density

$$f_{U(1:1)}(u_1|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(u_1-\sqrt{n}\mu)^2}{2\sigma^2}\}, \quad -\infty < u_1 < \infty, \tag{30}$$

and for $i = 2, 3, \ldots, n$, each U_i follows the model density

$$f_{U_i}(u_i|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{u_i^2}{2\sigma^2}\}, \quad -\infty < u_i < \infty$$

independently. That is, $U(2:n)$ depends only on σ, its model density is given by

$$f_{U(2:n)}(u(2:n)|\sigma) = \frac{1}{(\sqrt{2\pi\sigma^2})^{n-1}} \exp\{-\frac{\sum_{i=2}^{n} u_i^2}{2\sigma^2}\}. \tag{31}$$

So, we can see that $U(1:1)$ and $U(2:n)$ are independent of each other, hence the model density for $U(1:n)$ can be expressed by the separated form

$$f_{U(1:n)}(u(1:n)|\theta) = f_{U(1:1)}(U(1:1)|\mu,\sigma) f_{U(2:n)}(u(2:n)|\sigma).$$

Since the transformation defined by equation (29) is an orthogonal transformation, we have

$$f_{U(1:n)}(u(1:n)|\theta) = f_{X(1:n)}(x(1:n)|\theta).$$

Therefore, the model density for $X(1:n)$ is separable.

4 Illustrations

Several examples are given in the present section in order to illustrate the procedure and performance of the prior-free inference.

4.1 Examples for single parameter case

In this subsection, we show three examples for the case that the model density is defined on a single parameter. In this case, we put $k = 1$, thus $\theta = \theta_1$, $z = z_1$ and so forth.

Example 4.1 Let $X(1:n)$ be a sample that each X_i is independently distributed with the same normal density

$$f_{X_i}(x_i|\theta) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(x_i-\theta)^2}{2}\}, \quad -\infty < x_i < \infty \quad (i = 1, 2, \ldots, n),$$

where $\theta \in (-\infty, \infty)$ is the mean as an unknown parameter. Given x_1, the inferential function is defined by

$$z = \varphi(x_1, \theta)|_{x_1} = \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(t-\theta)^2}{2}\} dt. \tag{32}$$

For a given value of $\theta \in (-\infty, \infty)$, $v = t - \theta \to -\infty$ as $t \to -\infty$. Thus, we have

$$z = \int_{-\infty}^{x_1-\theta} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{v^2}{2}\} dv.$$

Hence,

$$\begin{aligned} \frac{\partial z}{\partial \theta} &= -\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(\theta - x_1)^2}{2}\}, \\ \lambda &= \int_{-\infty}^{\infty} |\frac{\partial z}{\partial \theta}| d\theta = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(\theta - x_1)^2}{2}\} d\theta = 1. \end{aligned}$$

It shows that the inferential function defined by equation (32) is fully informative. Therefore, from equations (23) and (11) we obtain the initial post data density for Θ as

$$f_\Theta(\theta|x_1) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(\theta - x_1)^2}{2}\}, \quad -\infty < \theta < \infty.$$

Moreover, from equation (5) the final post data density for Θ is given by

$$f_\Theta(\theta|x(1:n)) = \sqrt{\frac{n}{2\pi}} \exp\{-\frac{n\ (\theta - \bar{x})^2}{2}\},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

Incidentally, for an unobserved random variable Y which follows the normal density

$$f_Y(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y - \theta)^2}{2}\}, \quad -\infty < y < \infty,$$

we obtain the predictive density as

$$\begin{aligned} f_Y(y|x(1:n)) &= \int_{-\infty}^{\infty} f_Y(y|\theta) f_\Theta(\theta|x(1:n)) d\theta \\ &= \sqrt{\frac{n}{2\pi(n+1)}} \exp\{-\frac{n(y - \bar{x})^2}{2(n+1)}\}, \quad -\infty < y < \infty. \end{aligned}$$

It shows that $Y \sim \mathrm{N}(\bar{x}, \sqrt{\frac{n+1}{n}})$ for given $x(1:n)$.

Example 4.2 Let $X(1:n)$ be a sample, and suppose each X_i is independently distributed with the same normal density

$$f_{X_i}(x_i|\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\{-\frac{x_i^2}{2\theta^2}\}, \quad -\infty < x_i < \infty \quad (i = 1, 2, \ldots, n),$$

where θ denotes the standard deviation as an unknown parameter. Assume that $x_1 \neq 0$, we define the inferential function as

$$z = \varphi(x_1, \theta)|_{x_1} = \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi\theta^2}} \exp\{-\frac{t^2}{2\theta^2}\} dt. \tag{33}$$

For a given value of $\theta \in (0, \infty)$, $\frac{t}{\theta} \to -\infty$ as $t \to -\infty$. So, we have

$$\frac{\partial z}{\partial \theta} = -\frac{|x_1|}{\sqrt{2\pi}\theta^2} \exp\{-\frac{x_1^2}{2\theta^2}\}.$$

Hence,

$$\lambda = \int_0^{\infty} |\frac{\partial z}{\partial \theta}| d\theta = \int_0^{\infty} \frac{|x_1|}{\sqrt{2\pi}\theta^2} \exp\{-\frac{x_1^2}{2\theta^2}\} d\theta = \frac{1}{2}.$$

It shows that the inferential function defined by equation (33) is partially informative. From equations (23) and (11) we obtain the initial post data density of Θ as

$$f_\Theta(\theta|x_1) = \sqrt{\frac{2}{\pi}} \frac{|x_1|}{\theta^2} \exp\{-\frac{x_1^2}{2\theta^2}\}, \quad 0 < \theta < \infty.$$

Further, from equation (5) the final post data density for Θ is obtained as

$$f_\Theta(\theta|x(1:n)) = \frac{(\sum_{i=1}^n x_i^2)^{n/2}}{2^{(n-2)/2}\Gamma(n/2)} \frac{1}{\theta^{n+1}} \ \exp\{-\frac{\sum_{i=1}^n x_i^2}{2\theta^2}\}.$$

For an unobserved random variable Y which follows the normal density

$$f_Y(y|\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\{-\frac{y^2}{2\theta^2}\}, \quad -\infty < y < \infty,$$

we obtain the predictive density as

$$\begin{aligned} f_Y(y|x(1:n)) &= \int_{-\infty}^{\infty} f_Y(y|\theta) f_\Theta(\theta|x(1:n)) d\theta \\ &= \frac{4}{\sqrt{\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 + y^2} \right)^{n/2}, \quad -\infty < y < \infty. \end{aligned}$$

Example 4.3 Assume that $X(1:n)$ is a sample which each X_i is independently distributed with the same uniform density

$$f_{X_i}(x_i|\theta) = \frac{1}{\theta}, \quad 0 \le x_i < \theta \quad (i = 1, 2, \ldots, n),$$

where $\theta \in (0, \infty)$ denotes the upper limit which is regarded as an unknown parameter. Given x_1, the inferential function is defined by

$$z = \varphi(x_1, \theta)|_{x_1} = \int_0^{x_1} f_X(t|\theta) dt = \int_0^{x_1} \frac{1}{\theta} \, dt = \frac{x_1}{\theta}, \quad \theta \in (x_1, \infty). \tag{34}$$

Then, we have

$$\frac{\partial z}{\partial \theta} = -\frac{x_1}{\theta^2},$$

hence,

$$\lambda = \int_{x_1}^{\infty} |\frac{\partial z}{\partial \theta}| d\theta = \int_{x_1}^{\infty} \frac{x_1}{\theta^2} d\theta = 1.$$

Thus, the inferential function defined by equation (34) is fully informative. Further, from equations (23) and (11) we obtain the initial post data density as

$$f_\Theta(\theta|x_1) = \frac{x_1}{\theta^2}, \quad \theta \in (x_1, \infty).$$

Moreover, from equation (5) the final post data density of Θ is given by

$$f_\Theta(\theta|x(1,n)) = \frac{n x_{max}^n}{\theta^{n+1}}, \quad \theta \in [x_{max}, \infty),$$

where $x_{max} = \max\{x_1, x_2, \ldots, x_n\}$.

For an unobserved random variable Y which follows the uniform density

$$f_Y(y|\theta) = \frac{1}{\theta}, \quad 0 \le y < \theta,$$

we obtain the predictive density as follows:

$$\begin{aligned} f_Y(y|x(1:n)) &= \int_{x_{max}}^{\infty} f_Y(y|\theta) f_\Theta(\theta|x(1:n))d\theta = \int_{x_{max}}^{\infty} \frac{nx_{max}^n}{\theta^{n+2}}d\theta \\ &= \frac{n}{(n+1)x_{max}}, \quad 0 \le y < \frac{n+1}{n}x_{max}. \end{aligned}$$

It can be seen that the above results in Examples 4.1 and 4.2 agree with the results obtained by using Jeffreys priors. Example 4.3 is very simple and the result can be easily obtained by using the proposed approach. However, it may be difficult when some traditional Bayesian approaches are applied because the model density is not defined by an explicit function of the parameter.

4.2 Example for multivariate parameter case

In the following example, we continue Example 3.1 and show how to utilize the procedure of prior-free inference to obtain the post data density for the parameter vector $\theta = (\mu, \sigma)^{\mathtt{t}}$.

Example 4.4 From the model density of $U(1:1)$ expressed by equation (30), we obtain the post data density for μ, given u_1 and σ, as

$$f_\mu(\mu|u_1, \sigma) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\{-\frac{(\sqrt{n}\mu - u_1)^2}{2\sigma^2}\}, \quad -\infty < \mu < \infty.$$

The results in Example 4.1 imply that given u_1 and σ, $\mu \sim \mathrm{N}(\frac{u_1}{\sqrt{n}}, \frac{\sigma^2}{n})$. Moreover, from the model density of $U(2:n)$ expressed by equation (31), we obtain the post data density for σ, given $u(2:n)$, as

$$f_\sigma(\sigma|u(2:n)) = \frac{(\sum_{i=2}^n u_i^2)^{(n-1)/2}}{2^{(n-3)/2}\Gamma((n-1)/2)} \frac{1}{\sigma^n} \exp\{-\frac{\sum_{i=2}^n u_i^2}{2\sigma^2}\},$$

by applying the results in Example 4.3. Thus, the post data density of $\theta = (\mu, \sigma)^{\mathtt{t}}$ is given by

$$\begin{aligned} f_\Theta(\theta|u(1:n)) &= f_\mu(\mu|u(1:1), \sigma) f_\sigma(\sigma|u(2:n)) \\ &= \frac{n^{1/2}(\sum_{i=2}^n u_i^2)^{(n-1)/2}}{2^{(n-2)/2}\pi^{1/2}\Gamma((n-1)/2)} \frac{1}{\sigma^{n+1}} \exp\{-\frac{\sum_{i=2}^n u_i^2 + (\sqrt{n}\mu - u_1)^2}{2\sigma^2}\}. \end{aligned}$$

Since $U(1:n)$ is obtained from $X(1:n)$ by equation (29) which is an one-to-one transformation, the post data density given $x(1:n)$ is the same as that given $u(1:n)$, i.e., $f_\Theta(\theta|x(1:n)) = f_\Theta(\theta|u(1:n))$. Finally, for an unobserved random variable Y which follows the normal density

$$f_Y(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}, \quad -\infty < y < \infty,$$

we obtain the predictive density of Y based on $x(1:n)$ as

$$f_Y(y|x(1:n)) = c \left(\frac{(n+1)\sum_{i=2}^n u_i^2}{(n+1)\sum_{i=2}^n u_i^2 + (\sqrt{n}y - u_1)^2} \right)^{n/2},$$

where $c = (\frac{n}{(n+1)\pi\sum_{i=2}^n u_i^2})^{1/2} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$.

It can be verified that the results in Example 4.4 agree with the results obtained by using Jeffreys priors. It should be noted that our procedure is well systematized and the procedure using Jeffreys priors is somewhat ad hoc.

5 Methodology for model selection

When a number of contending models are constructed, we have to select one as the best among them. Recently, many information criteria are introduced for statistical model selection (for example, see [23]). A well-known and widely applied information criterion is Akaike information criterion, AIC (see [1, 21, 29]). The definition of AIC is really simple but it can be applied only to the case that each contending model density is defined by a specific function. For a case that the likelihood can not be defined, Konishi and Kitagawa developed generalized information criterion, GIC [22]. In this section, we introduce a new information criterion by extending the procedure of the prior-free inference.

5.1 Prior-free information criterion

Consider here $X(1:n)$ as a random simple that each X_i follows the same model density $f_X(x_i|\theta)$ with θ being a k-dimensional parameter vector. Let $\widetilde{X}$ be a set of m values in $X(1:n)$ for $k \leq m < n$. The model density for $\widetilde{X}$ can be defined based on the model density f_X, then the post data density $f_\Theta(\theta|\widetilde{x})$, that is regarded as a functional of f_X for given $\widetilde{x}$, can be obtained by using the procedure of the prior-free inference. For an unobserved random quantity, Y, which follows the model density $f_X(y|\theta)$, the predictive density is given by

$$p(y|\widetilde{x}) = \int f_X(y|\theta) f_\Theta(\theta|\widetilde{x}) d\theta.$$

We attempt to evaluate the model density f_X through evaluating the predictive density $p(y|\widetilde{x})$ because $p(y|\widetilde{x})$ can also be regarded as a functional of f_X.

Let $g_X(y)$ and $g_{\widetilde{X}}(\widetilde{x})$ denote the true densities of Y and $\widetilde{X}$, respectively. For given $\widetilde{x}$, the Kullback-Leibler information between $g_X(y)$ and $p(y|\widetilde{x})$ is as

$$I_K^C(g_X; p|\widetilde{x}) = \int \ln\{\frac{g_X(y)}{p(y|\widetilde{x})}\} g_X(y) dy.$$

Then, the expectation of $I_K^C(g_X; p|\widetilde{x})$ with respect to $g_{\widetilde{X}}(\widetilde{x})$ is given by

$$\mathrm{E}\{I_K^C(g_X; p|\widetilde{x})\} = \int I_K^C(g_X; p|\widetilde{x}) g_{\widetilde{X}}(\widetilde{x}) d\widetilde{x} = c + EIP,$$

where $c = \int \ln\{g_X(y)\} g_X(y) dy$ is a quantity that does not depend on $p(y|\widetilde{x})$, and EIP is the expected information for prediction defined by

$$EIP = -\int \ln\{p(y|\widetilde{x})\} g_X(y) g_{\widetilde{X}}(\widetilde{x}) dy d\widetilde{x}. \tag{35}$$

It is advisable to obtain a predictive density leading to a smaller value of $\mathrm{E}\{I_K^C(g_X; p|\widetilde{x})\}$, or equivalently a smaller value of EIP.

In order to estimate the value of EIP in equation (35), we draw a random sample (called the re-sample) of size m from $X(1:n)$ without replacement in once re-sampling

and repeat such re-sampling N times. Let $\widetilde{X}^{(i)} = \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}$ be the i-th re-sample, and let $\{X_{m+1}^{(i)}, X_{m+2}^{(i)}, \ldots, X_n^{(i)}\}$ be the elements of $X(1:n)$ except $\widetilde{X}^{(i)}$. From the law of large numbers, an estimate for twice EIP, which is called prior-free information criterion (PFIC), is obtained as

$$PFIC = -\frac{2}{N(n-m)} \sum_{i=1}^{N} \sum_{j=m+1}^{n} \ln\{p(x_j^{(i)}|\widetilde{x}^{(i)})\}, \tag{36}$$

where $\widetilde{x}^{(i)}$ and $x_j^{(i)}$ denote the observations for $\widetilde{X}^{(i)}$ and $X_j^{(i)}$, respectively. Obviously, $PFIC$ defined by equation (36) is as a functional of the model density f_X. Thus, we can use PFIC as a criterion for evaluating the model density for $X(1:n)$. It can be seen that a model is better than the others if it leads to a smaller value of PFIC. Such rule of model selection is called minimum PFIC method.

Note that we only give here a formula of PFIC for a random sample. The formula of PFIC may depend on a sample scheme, but the basic consideration may be eternal.

5.2 Selection of regression models

Consider a linear regression model as

$$x_j^{(i)} = \sum_{\ell=1}^{L} w_{j\ell}^{(i)} \beta_\ell + e_j^{(i)} \quad (j = 1, 2, \ldots, m), \tag{37}$$

for the observations $\widetilde{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}\}$ of the i-th re-sample $\widetilde{X}^{(i)}$ with $m \geq L+1$. Here, $w_{j\ell}^{(i)}$ is a given regressor, β_ℓ is an unknown regression coefficient, $e_j^{(i)}$ is an error term. As the usual case, we assume that the error terms are uncorrelated normal random variables distributed with zero mean and unknown variance σ^2. Redefining by $\widetilde{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})^{\mathtt{t}}$ a vector of the observations for ith re-sample, the regression model (37) can be expressed as

$$\widetilde{x}^{(i)} = W^{(i)}\beta + \varepsilon^{(i)}, \tag{38}$$

where $W^{(i)}$ is an $m \times L$ matrix with rank L, $\beta = (\beta_1, \beta_2, \cdots, \beta_L)^{\mathtt{t}}$ is a vector of the regression coefficients, $\varepsilon^{(i)} = (e_1^{(i)}, e_2^{(i)}, \ldots, e_m^{(i)})^{\mathtt{t}}$ is an random vector distributed with $\mathrm{N}(0, \sigma^2 I_m)$.

In order to simplify the procedure, we find an orthogonal matrix $H^{(i)} = ((H_1^{(i)})^{\mathtt{t}}|(H_2^{(i)})^{\mathtt{t}})^{\mathtt{t}}$ to reduce the regression model (38) into a separated form:

$$\begin{aligned} H_1^{(i)}\widetilde{x}^{(i)} &= R^{(i)}\beta + H_1^{(i)}\varepsilon^{(i)}, \\ H_2^{(i)}\widetilde{x}^{(i)} &= H_2^{(i)}\varepsilon^{(i)}, \end{aligned}$$

where $R^{(i)}$ is an $L \times L$ right-trigonometric matrix. Thus, from the properties of orthogonal matrix, we have $H^{(i)}\varepsilon^{(i)} \sim \mathrm{N}(0, \sigma^2 I_m)$, and we can see also that $H_1^{(i)}\varepsilon^{(i)} \sim \mathrm{N}(0, \sigma^2 I_L)$ and $H_2^{(i)}\varepsilon^{(i)} \sim \mathrm{N}(0, \sigma^2 I_{m-L})$ are independent of each other.

By using the procedure of the prior-free inference, we obtain the post data density for β given $\widetilde{x}^{(i)}$ and σ as follows:

$$f_\beta(\beta|\sigma, \widetilde{x}^{(i)}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^L |\det(R^{(i)})| \exp\{-\frac{a_i(\beta)}{2\sigma^2}\},$$

and the post data density for σ given $\widetilde{x}^{(i)}$ is

$$f_\sigma(\sigma|\widetilde{x}^{(i)}) = \frac{(b_i)^{(m-L)/2}}{2^{(m-L-2)/2}\Gamma((m-L)/2)\sigma^{m-L+1}} \exp\{-\frac{b_i}{2\sigma^2}\}.$$

Moreover, for $j = m+1, m+2, \ldots, n$, the predictive distribution density of $X_j^{(i)}$ is given by

$$\begin{aligned} p(x_j^{(i)}|\widetilde{x}^{(i)}) &= (\frac{d_{ij}}{\pi b_i})^{1/2}\frac{\Gamma((m-L+1)/2)}{\Gamma((m-L)/2)} \\ &\quad \times(1+\frac{d_{ij}(x_j^{(i)}-c_{ij})^2}{b_i})^{-(m-L+1)/2} \end{aligned} \tag{39}$$

In the above equations, $a_i(\beta)$, b_i, c_{ij}, and d_{ij} are defined, respectively, as the follows:

$$\begin{aligned} a_i(\beta) &= (R^{(i)}\beta - H_1^{(i)}\widetilde{x}^{(i)})^{\mathtt{t}}(R^{(i)}\beta - H_1^{(i)}\widetilde{x}^{(i)}), \\ b_i &= (H_2^{(i)}\widetilde{x}^{(i)})^{\mathtt{t}}H_2^{(i)}\widetilde{x}^{(i)}, \\ c_{ij} &= (w_j^{(i)})^{\mathtt{t}}(R^{(i)})^{-1}H_1^{(i)}\widetilde{x}^{(i)}, \\ d_{ij} &= (1+(w_j^{(i)})^{\mathtt{t}}((R^{(i)})^{\mathtt{t}}(R^{(i)}))^{-1}w_j^{(i)})^{-1}, \end{aligned}$$

where $w_j^{(i)} = (w_{j1}^{(i)}, \ldots, w_{jL}^{(i)})^{\mathtt{t}}$ is the vector of the regressors corresponding to $\widetilde{x}^{(i)}$. Thus, PFIC for the model can be obtained by using equations (36) and (39).

5.3 Simulation study

In order to examine the performance of the minimum PFIC method, we carried out a simulation study. The data used here are generated by using the polynomial of degree three:

$$x_t = -10 + 0.2t - 0.09t^2 + 0.002t^3 + r_t, \quad (t = 1, 2, \ldots, n), \tag{40}$$

which can be regarded as the true distribution, where r_t is generated by using the standard normal random numbers. We fit the polynomial regression model

$$x_t = \sum_{\ell=0}^{L} t^\ell \beta_\ell + e_t, \quad (t = 1, 2, \ldots, n)$$

to the data generated by equation (40), where L denotes the degree of the model, and e_t is a random error term. The probability distribution for the error terms in this model is the same as that in the model (37).

Here, we compare our minimum PFIC method with other methods such as the minimum AIC method (see [21] and [29]) and the minimum BIC method (see [30]). The values of PFIC, AIC and BIC are calculated respectively for $L = 0, 1, \ldots, 5$. Then, we can estimate the model degrees of by using the minimum PFIC, AIC and BIC methods. Such experiment was repeated 1000 times with the size of re-sample being $m = 6$ and the times of re-sampling for each experiment being $N = 1000$.

Table 5.1 and Table 5.2 show the frequencies of the model degrees determined by using each method for $n = 30$ and $n = 60$, respectively. As shown in the tables, the model degrees determined by using the minimum PFIC method agree with the true model degree perfectly but the others are not. The result shows that the performance of minimum PFIC method is obviously better than that of the others, and it can be seen that the minimum PFIC method works well even for a small sample.

Table 5.1: Frequencies of estimated model degree ($n = 30$).

model order	0	1	2	3	4	5
PFIC	0	0	0	1000	0	0
AIC	0	0	0	725	166	109
BIC	0	0	0	880	86	34

Table 5.2: Frequencies of estimated model degree ($n = 60$).

model order	0	1	2	3	4	5
PFIC	0	0	0	1000	0	0
AIC	0	0	0	752	142	106
BIC	0	0	0	934	46	20

6 Concluding remarks

A new procedure of statistical inference, named by the prior-free inference, was introduced for developing objective Bayesian analysis based on an information-theoretic approach. 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. Thus, this approach does not only have the advantages of the Bayesian approach but also can avoid the difficulty of the traditional Bayesian approach encounters due to a lack of prior information. A methodology, named by the minimum PFIC method, for model selection was also developed by utilizing a newly-introduced information criterion, PFIC, based on the extension of the procedure for prior-free inference. The result of simulation study shown that the performance of minimum PFIC method is very good.

An important problem is the relation between our prior-free inference and Fisher's fiducial approach. It can be verified that for models with a single parameter that has a sufficient statistic, these two approaches 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 maybe difficult for multivariate parameter cases.

Nowadays many objective Bayesian approaches use Jeffreys priors. Sometimes, the procedure of the prior-free inference and that using Jeffreys priors may lead to a same result. However, it can be seen that the procedure of prior-free inference is well systematized and the procedure using Jeffreys is somewhat ad hoc. Moreover, a number of objections can be made to the Bayesian procedure using Jeffreys priors, the most important of which is that it depends on the values 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. Sometimes, the Bayesian procedure using Jeffreys priors will violate the likelihood principle, and it is difficult to apply the procedure to multivariate parameter cases. Also, there are diffi-

culties in Bernardo's reference priors approach using the Lindley's criterion functional. Such difficulties can be overcome by the use of the procedure of prior-free inference.

References

[1] Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC19** (1974) 716–723.

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

[3] Akaike, H. On minimum information prior distributions. *Ann. Inst. Statist. Math.* **35** (1983) 139–149.

[4] Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc.* **53** (1763) 370–418.

[5] Berger, J.O. An overview of robust Bayesian analysis. *Test* **3** (1994) 5–124.

[6] Berger, J.O., Bernardo, J.M. On the development of reference priors (with discussion). In: *Bayesian Statistics 4*. (Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., eds). Oxford University Press, Oxford, 1992, 35–60.

[7] Bernardo, J.M. Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.* **B41** (1979) 113–147.

[8] Bernardo, J.M. Nested hypothesis testing: the Bayesian reference criterion (with discussion). In: *Bayesian Statistics 6*. (Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., eds). Oxford University Press, Oxford, 1999, 101–130.

[9] Box, G.E.P., Tiao, G.C. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Massachusetts, 1973.

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

[11] Fine, T.L. Foundations of probability (update). In: *Encyclopedia of Statistical Sciences, Up. Vol. 3* (Kotz, S., ed.). John Wiley and Sons, New York, 1999, 246–254.

[12] Fisher, R.A. Inverse probability. *Proc. Camb. Phil. Soc.* **26** (1930) 528–535.

[13] Fisher, R.A. The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. Roy. Soc.* **A139** (1933) 343–348.

[14] Fisher, R.A. The fiducial argument in statistical inference. *Ann. Eugenics* **6** (1935) 391–398.

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

[16] Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc.* **A186** (1946) 453–461.

[17] Jiang, X.Q. A general procedure of statistical inference based on information theory, In: *Statistical Physics* (Tokuyama, M., Stanley, H.E., eds.). American Institute of Physics, Melville, 2000, 642–644.

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

[19] Jiang, X.Q. Prior-free Bayesian inference based on information-theoretic approach. *The Journal of Asahikawa University* No. 57·58 (2004) 1–17.

[20] Jiang, X.Q. Estimation and identification of regression models via prior-free Bayesian inference. In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (Fischer, R., Preuss, R., Toussaint, U., eds). American Institute of Physics, Melville, 2004, 501–508.

[21] Kitagawa, G., Gersch, W. *Smoothness Prior Analysis of Time Series*. Springer-Verlag, New York, 1996.

[22] Konishi, S., Kitagawa, G. Generalized information criterion in model selection. *Biometrika* **83** (1996) 875–890.

[23] Konishi, S., Kitagawa, G. *Information Criteria*. Asakura Syoten, Tokyo, 2004. [in Japanese]

[24] Kullback, S. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

[25] Laplace, P.S. *Théorie Analytique des Probabilités*. Courcier, Paris, 1812.

[26] Lehmann, E.L., Casellab, G. *Theory of Point Estimation* (2-nd edn). Springer-Verlag, New York, 1998.

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

[28] Lindley, D.V. On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** (1956) 986–1005.

[29] Sakamoto, Y., Ishiguro, M., Kitagawa, G. *Akaike Information Criterion*. KTK Scientific Publishers, Tokyo, 1986.

[30] Schwarz, G. Estimation the dimension of a model. *Ann. Statist.* **6** (1978) 461–464.

[31] Stone, M. Strong inconsistency from uniform priors (with discussion). *J. Amer. Statist. Assoc.* **71** (1976) 114–125.

[32] Zacks, S. *The Theory of Statistical Inference*. John Wiley and Sons, New York, 1971.

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

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

[35] Zellner, A. Optimal information processing and Bayes's theorem (with discussion). *Amer. Statist.* **42** (1988) 278–294.

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

Appendix A: proof of Lemma 3.1

Under the conditions C1 and C2, we have $\lambda = \int_{\boldsymbol{\mathcal{S}}(f_{Z|x(1:k)})} dz$ from equations (10) and (11). Thus, the proof is completed from the fact that $\boldsymbol{\mathcal{S}}(f_Z|x(1:k)) \subseteq [0,1] \times [0,1] \times \cdots \times [0,1]$.

Appendix B: proof of Theorem 3.1

From equation (20), we have

$$I_K^C(\pi; f_\Theta|x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))}\}\frac{\pi(\theta)}{f_\Theta(\theta|x(1:k))} f_\Theta(\theta|x(1:k))d\theta.$$

By applying equation (4), the above equation can be rewritten as

$$I_K^C(\pi; f_\Theta|x(1:k)) = \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_\Theta(\theta|x(1:k))d\theta, \tag{41}$$

where

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

For a given model density $f_{X(1:k)}(x(1:k)|\theta)$, if $\pi(\theta)$ is fixed in a proper density, then the marginal density $h(x(1:k))$ is fixed by equation (3). Hence, the function $\phi(x(1:k),\theta)$ in equation (41) cannot be changed through $f_\Theta(\theta|x(1:k))$. Thus, we have to maximize $I_K^C(\pi; f_\Theta|x(1:k))$ with respect to $f_\Theta(\theta|x(1:k))$. It is well-known that a solution to maximizing $I_K^C(\pi; f_\Theta|x(1:k))$ can be obtained when we put

$$f_\Theta^{(1)}(\theta|x(1:k)) = c_1 \tag{42}$$

with $c_1 = 1/\psi$ being a constant. Thus, the solution given by equation (22) is obtained.

On the other hand, by applying equation (12) to equation (41), we have

$$\begin{aligned} I_K^C(\pi; f_\Theta|x(1:k)) &= \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_Z(z|x(1:k))|\det(J)|d\theta \\ &= \int_{\boldsymbol{\mathcal{S}}(f_Z|x(1:k))} \ln\{\phi(x(1:k),\theta)\}\phi(x(1:k),\theta) f_Z(z|x(1:k))dz. \end{aligned}$$

It is obvious that $I_K^C(\pi; f_\Theta|x(1:k))$ can also be maximized when we put $f_Z(z|x(1:k)) = c_2$ or equivalently

$$f_\Theta^{(2)}(\theta|x(1:k)) = c_2\ |\det(J)| \tag{43}$$

from equation (12) with $c_2 = 1/\lambda$ being a constant. Then, the solution given by equation (23) is obtained from equation (43). Moreover, from equation (19), we can see that $F(f_\Theta, \pi|f_{X(1:k)})$ is maximized as long as $I_K^C(\pi; f_\Theta|x(1:k))$ is maximized. Thus, Theorem 3.1 is proved.

Appendix C: proof of Theorem 3.2

If the value of $\psi = \int_{\boldsymbol{\mathcal{S}}(\pi)} d\theta$ is finite, then the value of $I(f_\Theta|x(1:k))$ is given by

$$I(f_\Theta^{(1)}|x(1:k)) = \ln\{c_1\}, \tag{44}$$

for the solution given by equation (42). On the other hand, the value of $I(f_\Theta|x(1:k))$ is as

$$\begin{aligned} I(f_\Theta^{(2)}|x(1:k)) &= c_2 \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{c_2|\det(J)|\}|\det(J)|d\theta \\ &= c_2 \int_{\boldsymbol{\mathcal{S}}(f_Z|x(1:k))} \ln\{c_2|\det(J)|\}dz \end{aligned} \tag{45}$$

for the solution given by equation (23). It is obvious that

$$I(f_\Theta^{(2)}|x(1:k)) - I(f_\Theta^{(1)}|x(1:k)) = c_2 \int_{\boldsymbol{\mathcal{S}}(\pi)} \ln\{\frac{c_2|\det(J)|}{c_1}\}|\det(J)|d\theta > 0.$$

from equations (44) and (45).

Further, if the value of $\psi = \int_{\boldsymbol{\mathcal{S}}(\pi)} d\theta$ is infinite, then $I(f_\Theta^{(1)}|x(1:k)) = -\infty$ as $c_1 \to 0$ under the assumption that $f_\Theta(\theta|x(1:k))$ is proper. On the other hand, from the conditions C1 and C2, equations (11) and (45), we can see that $I(f_\Theta^{(2)}|x(1:k))$ must be finite. Thus, we have $I(f_\Theta^{(2)}|x(1:k)) - I(f_\Theta^{(1)}|x(1:k)) = \infty$. So, Theorem 3.2 is proved.