J. Japan Statist. Soc.
Vol. 39 No. 1 2009 111–131

PREDICTIVE CREDIBLE REGION FOR BAYESIAN DIAGNOSIS OF A HYPOTHESIS

Takemi Yanagimoto* and Toshio Ohnishi**

A Bayesian method for diagnosing a hypothesis is proposed in terms of the optimum Bayesian predictor under the e-divergence loss. We introduce a predictive credible region as a modified version of a posterior credible region. The predictive credible region is closely related to the complement of the rejection region of the likelihood ratio test in the frequentist context. As an application we revisit the controversy regarding Lindley's paradox, and observe satisfactory performance of the proposed credible region in contrast to the Bayes factor. Another important application concerns a method for analyzing additional evidence when a hypothesis is once rejected.

Key words and phrases: Bayes factor, conjugate analysis, e-divergence loss, likelihood ratio test, Lindley's paradox, posterior credible region, statistical test.

1. Introduction

Consider a diagnostic problem for a null hypothesis $\mathrm{H}_0 : \theta = \theta_0$ against an alternative hypothesis $\mathrm{H}_1 : \theta \neq \theta_0$ or the one sided alternative, when a present data set and a prior information are available. This problem appears frequently in statistical applications, and various procedures for adjusting to the actual conditions are hoped to be developed further. In the frequentist approach, this problem is treated as the statistical test or the confidence interval without assuming any prior information. One of the standard methods in the frequentist context is to apply the likelihood ratio.

We attempt here to propose a novel credible region induced from the Bayesian optimum predictor under a relative entropy loss. The region can be useful for diagnosing a hypothesis under certain situations where existing methods do not perform satisfactorily. The present research is motivated by the controversy of so-called Lindley's paradox. This paradox has been referred to as an example that shows superiority of a procedure based on the Bayes factor over the statistical test, see Lindley (1957), Schwarz (1978), Shafer (1982), Kass and Raftery (1995) and Robert (2001). We wonder why the posterior density, or more specifically a posterior credible region, was not often discussed in relation to this controversy. In light of the most important position of the posterior density in Bayesian theory, one of the standard Bayesian procedures will be to calculate whether θ_0 is in the posterior credible region or not. It is known (Yanagimoto, 1999) that there is a large discrepancy between results induced from the Bayes factor and the posterior

Received October 8, 2008. Revised April 7, 2009. Accepted April 11, 2009.

*Department of Industrial and Systems Engineering, Chuo University, 1-13-27, Kasuga Bunkyo-ku, Tokyo 112-8551, Japan.

**The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.

density. We will see that the proposed credible region and the posterior one shed new light on this controversy regarding the fundamental inferential issue.

Because there are a wide variety of applications, it becomes necessary to develop various modifications of the posterior credible region, so as to adjust to the actual conditions we may encounter. Here we attempt to apply the recently developed prediction theory, which is a dual version of the classical Bayesian prediction theory, to construct a predictive credible region of θ. Following this, another important application claimed by Cornfield (1966) will be briefly discussed. The term "predictive credible region" is seen in the literature, see Bernardo and Smith (2000; p. 260) for example, where it is not a region of θ. To distinguish the proposed credible region from the existing one, we attach the words "of the θ" to emphasize this difference, when it is considered necessary.

Let $\{p(x;\theta) \mid \theta \in \Theta\}$ be a family of densities on $\mathcal{X} \subset R^1$, where Θ is an open interval. Let $\boldsymbol{x}$ be a sample vector of size n from a population having the sampling density $p(x;\theta)$. Write a prior density on Θ as $\pi(\theta)$, and a posterior density as $\pi(\theta \mid \boldsymbol{x})$. We will mostly assume that a sampling density function is in the exponential family (EF) of the form

$$p(\boldsymbol{x};\eta) = \exp[n\{\bar{t}\eta - M(\eta)\}]a(\boldsymbol{x}) \tag{1.1}$$

where $\bar{t} = \sum t_i/n$ is the sufficient statistic. The mean of $\bar{t}$ is written as μ, which is expressed as the first derivative of $M(\eta)$, $M_\eta(\eta)$. Recall that most sampling densities in the practical applications are in the EF. More importantly, this assumption makes the optimum predictor easier, as will be seen in the following section. The parameterization (1.1) is convenient for theoretical developments in this study, but we note here that our primary interest may be placed on the canonical parameter η, the mean parameter μ or another one $g(\mu)$ for an appropriate function $g(\cdot)$.

The conjugate analysis presents a simple and useful scheme in the fundamental Bayesian theory. Let $N(\mu)$ be the conjugate convex function of $M(\eta)$, which is written as $\mu\eta - M(\eta)$. Then it holds that $\eta = N_\mu(\mu)$. Write a member of the exponential dispersion model (EDM) with the position parameter m and the precision parameter δ_0 of the form

$$\pi(\eta; m, \delta_0) = \exp[\delta_0\{m\eta - M(\eta) - N(m)\}]b(\eta)\kappa(m, \delta_0) \tag{1.2}$$

as the EDM(m, δ_0). We assume that a prior density is the EDM(m, δ_0). This prior density is called to be conjugate, and yields that the posterior density

$$\pi(\eta \mid \boldsymbol{x}) = \exp[(n+\delta_0)\{\hat{\mu}_L\eta - M(\eta) - N(\hat{\mu}_L)\}]b(\eta)\kappa(\hat{\mu}_L, n+\delta_0)$$

where $\hat{\mu}_L = (n\bar{t} + \delta_0 m)/(n+\delta_0)$. This estimate of a weighted linear form is familiar in the conjugate analysis. Consequently, the posterior density becomes the EDM$(\hat{\mu}_L, n+\delta_0)$, and is a member of the same EDM as the prior density belongs to. This closure property is called *closed under sampling*.

The use of an appropriate function $b(\eta)$ is widely employed in the conjugate analysis in the existing Bayesian literature. It is regarded as a supporting measure, which corresponds to the function $a(\boldsymbol{x})$ of the EF in (1.1). When $b(\eta)$ is

constant, a conjugate prior density was called to be standard in Consonni and Veronese (1992). An advantage is that we can explain the prior density in a simple way; an example of the case of the exponential distribution will be given in Example 2.2. Another pertains to our flexible choice of the parameter of interest θ.

For national simplicity we will focus on the scalar case of θ; a vector parameter case will be discussed in the final section. An extension to the vector case is straightforward when the null hypothesis is simple.

2. Predictive credible region

The prediction problem is the key in the current statistics, as is seen in Aitchison and Dunsmore (1975) and Geisser (1993). When the present observation $\boldsymbol{x}$ is given, a predictor $p(\boldsymbol{y} \mid \boldsymbol{x})$ is used for estimating a future (or unobserved) observation $\boldsymbol{y} \in R^n$. It is regarded also as an estimate of the sampling density $p(\boldsymbol{y}; \theta)$ instead of that of the parameter θ. Write the expectation of a function $f(\theta)$ with respect to a probability measure $\pi(\theta)$ as $\mathrm{E}\{f(\theta) \mid \pi(\theta)\}$. Then the e-divergence loss $\mathrm{D}(p(\boldsymbol{y} \mid \boldsymbol{x}), p(\boldsymbol{y}; \theta))$ is defined as $\mathrm{E}[\log\{p(\boldsymbol{y} \mid \boldsymbol{x})/p(\boldsymbol{y}; \theta)\} \mid p(\boldsymbol{y} \mid \boldsymbol{x})]$. Note that this loss is to be distinguished from the dual loss $\mathrm{D}(p(\boldsymbol{y}; \theta), p(\boldsymbol{y} \mid \boldsymbol{x}))$, which is familiar in the existing prediction theory.

Consider the following predictor of the form

$$p_e(\boldsymbol{y} \mid \boldsymbol{x}) = \exp[\mathrm{E}\{\log p(\boldsymbol{y}; \theta) \mid \pi(\theta \mid \boldsymbol{x})\}]/c(\boldsymbol{x}) \tag{2.1}$$

with the normalizing constant $c(\boldsymbol{x})$. This predictor is shown to be optimum under the e-divergence loss. Corcuera and Giummole (1999) gave a unified view of the optimum predictors under general α-divergence losses for $-1 \leq \alpha \leq 1$. This general view owes to the differential geometric theory in statistics, see Nagaoka and Amari (2000). The case of $\alpha = 1$, which corresponds to the e-divergence loss, was extensively studied in Yanagimoto and Ohnishi (2009). Note that the existence of $p_e(\boldsymbol{y} \mid \boldsymbol{x})$ is valid under mild regularity conditions, and it will be assumed throughout.

When the sampling density is in the EF in (1.1) and we choose the canonical parameter η as a parameter of interest θ, the optimum predictor is written as $p(\boldsymbol{y}; \hat{\theta})$ with $\hat{\theta}$ being the posterior mean of θ. This means that the optimum predictor is obtained by plugging $\hat{\theta}$ into $p(\boldsymbol{y}; \theta)$. When a predictive density satisfies this property, we call it to be estimative. To avoid possible confusion, we will use throughout the symbol $\hat{\eta}$ without any suffix to denote the posterior mean of the canonical parameter η, and the corresponding estimate of a transformed parameter $\theta = g(\eta)$ will be written as $\hat{\theta}$ $(= g(\hat{\eta}))$.

A prior density is assumed to be of the form

$$\pi(\theta; c, \delta_0) = \exp\{-\delta_0 d(\theta, c)\} b(\theta) \kappa(c, \delta_0) \tag{2.2}$$

for appropriate nonnegative functions $d(\cdot, \cdot)$ and $b(\theta)$, and constants $\delta_0 (> 0)$ and $c (\in \Theta)$, where $\kappa(c, \delta_0)$ is the normalizing constant. This informative prior density is a generalization of a conjugate prior density in (1.2). The function $d(\theta, c)$

denotes a distance between θ and c, and $\exp\{-\delta_0 d(\theta, c)\}$ in the right-hand side of (2.2) is the main factor of this informative prior density. A larger value of δ_0 represents our stronger belief in the values near the position parameter c of the prior density. The function $b(\theta)$ is a supporting factor of the density, which corresponds with the supporting measure $a(\boldsymbol{x})$ of the sampling density in the EF in (1.1). In a practical example, it appears as a non-informative prior density such as Jeffreys' prior. We will assume that the probability of a non-empty open subset of Θ with respect to a prior density is positive, and also that $b(\theta)$ is continuous. This prior density is generalized in two different ways. One is to allow the main factor to be decomposed into two terms, $q\exp\{-\delta_1 d_1(\theta, c_1)\}\kappa(c_1, \delta_1) + (1-q)\exp\{-\delta_2 d_2(\theta, c_2)\}\kappa(c_2, \delta_2)$, for a positive mixing probability q. The other is to allow the limit as δ_0 tends to zero or infinity. These extensions make the prior density flexible; a non-informative prior is derived as a limit at $\delta_0 = 0$ in (2.2). Taking the limit of an extended prior density at $\delta_1 = \infty$, we obtain the following mixed prior density

$$\pi(\theta; c_1, c_2, q) = q\delta_D(\theta - c_1) + (1-q)\exp\{-\delta_2 d_2(\theta, c_2)\}b(\theta)\kappa(c_2, \delta_2) \tag{2.3}$$

where $\delta_D(\theta - c)$ is Dirac's δ function. This type of a mixed prior density will appear in Lindley's paradox.

When we consider a parameter transformation $\theta = g(\xi)$ with $g(\cdot)$ being a differentiable function, we set the supporting factor as $b(g(\xi))g'(\xi)$, whilst we set $d(\theta, c)$ as $d(g(\xi), c)$. Recall that a usual non-informative prior density is adjusted by the Jacobian $g'(\xi)$. On the other hand, the main factor is independent of the Jacobian.

Note that the e-divergence loss is the key to the present approach because of its relation with the optimality of the predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$. Thus we write $d(\theta \mid \boldsymbol{x}) = \mathrm{D}(p_e(\boldsymbol{y} \mid \boldsymbol{x}), p(\boldsymbol{y}; \theta))$, and define a density function of θ induced from this distance as follows.

Definition 2.1. *Let the optimum predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$ be defined in* (2.1). *The predictor-based density of θ is defined by*

$$\pi_e(\theta \mid \boldsymbol{x}) = \exp\{-d(\theta \mid \boldsymbol{x})\}b(\theta)K(\boldsymbol{x}) \tag{2.4}$$

where $b(\theta)$ is given in a prior density (2.2) *and $K(\boldsymbol{x})$ is the normalizing constant.*

When the optimum predictor is not estimative, the minimum of the distance $d(\theta \mid \boldsymbol{x})$ is positive. Then we can replace $d(\theta \mid \boldsymbol{x})$ in (2.4) by $d(\theta \mid \boldsymbol{x}) - d(\check{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})$ with $\check{\theta}(\boldsymbol{x}) = \operatorname{Argmin} d(\theta \mid \boldsymbol{x})$. The minimizer $\check{\theta}(\boldsymbol{x})$ is an estimate of θ induced from the optimum predictor. A regularity condition on the existence of the density in (2.4) is satisfied in usual Bayesian models, and it will be assumed throughout. We note that the predictor-based density of θ can be regarded as a modification of the posterior density from the viewpoint of the optimum predictor. However, the difference between them becomes large, when a mixed prior density of the form in (2.3) is assumed. In this case the predictor-based

density is a usual density function of the form in (2.2), whilst the posterior density is still a mixed density of the form in (2.3).

We restricted our attention only on a prior density of the form in (2.2) and its direct extensions. This restriction, however, is not severe in practice, since our interest focuses on constructing a credible region in order to examine whether it includes a specific value θ_0 or not. On the other hand, the form in (2.2) may be too general, since $\delta_0 d(\theta, c)$ can be regarded as another distance $\tilde{d}(\theta, c)$. A reason why we employ this form is that the choice of a value of δ_0 is usually very important in defining a prior density. For example, we can regard δ_0 in a conjugate prior density in (1.2) as representing the strength of our belief compared with the sample size.

Our assumption on the factorization of a prior density into two factors in (2.2) may be less familiar in the Bayesian literature. Note, however, a familiar conjugate prior and its modifications often have this form, as will be found in the subsequent examples.

The predictor-based density of θ yields a predictive credible region of θ.

DEFINITION 2.2. *For a given significance level α the predictive credible region is given by*

$$C_\alpha = \{\theta \mid 2\{d(\theta \mid \boldsymbol{x}) - d(\breve{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})\} \leq c_\alpha(\boldsymbol{x})\} \tag{2.5}$$

where $c_\alpha(\boldsymbol{x})$ satisfies $\Pr\{C_\alpha \mid \pi_e(\theta \mid \boldsymbol{x})\} = 1 - \alpha$. The complement of the predictive credible region will be called the predictor incredible region, and is written as I_α.

The multiplyer 2 is attached to yield a χ^2-approximation, $c_\alpha \doteq \chi^2_{1-\alpha}$. A directed version of the predictive credible region is defined by using the statistic $\mathrm{sgn}(\theta - \theta_0)\sqrt{2d(\theta \mid \boldsymbol{x})}$, when a one sided alternative is of interest. Recall that this treatment was employed also in the likelihood ratio test (LRT) statistic.

We emphasize here that the predictive credible region is a truly Bayesian one. In fact, the components necessary for deriving the credible region consist of the posterior density, a Bayesian predictor and the e-divergence loss. These three components are familiar in Bayesian literature; recall that the e-divergence loss appears frequently in Bernardo and Smith (2000) and Robert (2001). A familiar posterior credible region is the highest posterior density (HPD) credible region. Note that we do not use $\pi_e(\theta \mid \boldsymbol{x})$ but the exponent of $\pi_e(\theta \mid \boldsymbol{x})$ to define the credible region in (2.5). This treatment allows the predictive credible region to be invariant under the parameter transformation $\theta = g(\xi)$ with $g(\cdot)$ being a strictly monotone and differentiable function. In other words, for the predictive credible region $\tilde{C}_\alpha$ of ξ, it holds that θ is in C_α if and only if ξ is in $\tilde{C}_\alpha$.

Recall that an alternative hypothesis is affirmatively claimed by rejecting the null hypothesis in the context of the statistical test. This is why our primary attention is placed on the rejection region in the statistical test. In this concern the incredible region I_α will play an important role in diagnostic problems. We will suppress the suffix α to C and I, when no confusion will be anticipated.

In the following, we present three familiar examples. We will find that necessary calculations are straightforward.

Example 2.1 (Normal distribution). Let $\boldsymbol{x}$ be a sample vector of size n from a normal population with mean μ and variance 1. Assume a normal prior density with mean m and variance $1/\delta_0$, $p_N(\mu; m, 1/\delta_0)$. Then the optimum predictor is written as $p(\boldsymbol{y}; \hat{\mu}_L, 1) = \prod p_N(y_i; \hat{\mu}_L, 1)$ with $\hat{\mu}_L (= \hat{\mu}) = (n\bar{x} + \delta_0 m)/(n + \delta_0)$. It follows that

$$\pi_e(\mu \mid \boldsymbol{x}) = \sqrt{\frac{n}{2\pi}} \exp\left\{-\frac{n}{2}(\hat{\mu}_L - \mu)^2\right\},$$

that is, the predictor-based density of μ is $p_N(\mu; \hat{\mu}_L, 1/n)$. Consequently, the predictor incredible region is expressed as

$$I = \{\mu \mid n(\hat{\mu}_L - \mu)^2 > \chi^2_{1-\alpha}\}.$$

Note that the critical value is independent of $\boldsymbol{x}$. Recall that the rejection region of the standard statistical test for $\mathrm{H}_0 : \mu = \mu_0$ against $\mathrm{H}_1 : \mu \neq \mu_0$ is written as $R = \{\boldsymbol{x} \mid n(\bar{x} - \mu_0)^2 > \chi^2_{1-\alpha}\}$. Thus the difference between them is in the use of $\hat{\mu}_L$ in I and that of $\bar{x}$ in R.

Example 2.2 (Exponential distribution). Consider the exponential population with mean λ, and assume that a parameter of interest is $\theta = 1/\lambda$. A familiar conjugate prior density on θ is a gamma prior density with mean $1/m$ and the squared coefficient of variation $1/\delta_0$, which will be written as $p_G(\theta; 1/m, \delta_0)$. Then the parameter θ denotes the canonical parameter, and the supporting factor in (2.2) is $b(\theta) = 1/\theta$. This supporting factor is regarded as Jeffreys' (non-informative) prior density. It follows that the optimum predictor is written as $\prod p_G(y_i; 1/\hat{\theta}, 1)$ with $\hat{\theta} = (n + \delta_0)/(n\bar{x} + \delta_0 m)$. Note that $\hat{\theta}$ is the posterior mean of θ and that $1/\hat{\theta}$ is not that of $1/\theta$. Setting $d(\theta \mid \boldsymbol{x}) = n\{\theta/\hat{\theta} - \log(\theta/\hat{\theta}) - 1\}$, we obtain the predictor-based density of θ

$$\pi_e(\theta \mid \boldsymbol{x}) = \frac{n^n}{\Gamma(n)\exp(n)} \exp\{-d(\theta \mid \boldsymbol{x})\}\frac{1}{\theta}.$$

Consequently, the predictive credible region is written as in the form in (2.5). The critical value c_α is independent of $\boldsymbol{x}$, and therefore is independent of $\hat{\theta}$. This property comes from our concentration only on the exponent of $\pi_e(\theta \mid \boldsymbol{x})$. A similar problem arises also in eliciting a conjugate prior density, see Yanagimoto and Ohnishi (2005).

It requires some computations to obtain the exact value of c_α, but a familiar asymptotic approximation leads us to $c_\alpha \doteq \chi^2_{1-\alpha}$ for a fairly large n. Recall that the common difficulty arises in determining the critical value of the LRT statistic. In fact, the maximum likelihood estimate of θ is $1/\bar{x}$, and the LRT statistic is expressed as $2n\{\theta\bar{x} - \log(\theta\bar{x}) - 1\}$, which is equivalent with $2d(\theta \mid \boldsymbol{x})$ when $\delta_0 = 0$. The rejection region is written as $R = \{\boldsymbol{x} \mid 2n\{\theta\bar{x} - \log(\theta\bar{x}) - 1\} > c_\alpha\}$. Note that this critical value is similar to that in the predictive incredible region I_α, as

is the normal case. Thus it is expected that necessary sample sizes to assure a good χ^2-approximation are common in our approach and the LRT.

Example 2.3 (Binomial distribution). The third example pertains to the binomial sampling distribution $\mathrm{Bi}(n,p)$. Let x be an observation. The familiar conjugate prior density in this case is expressed by setting $\theta = p$ as

$$\pi(p; m, \delta_0) = \frac{1}{B(\delta_0 m, \delta_0(1-m))} p^{\delta_0 m - 1}(1-p)^{\delta_0(1-m)-1} \tag{2.6}$$

where $B(\cdot,\cdot)$ is the beta function. This beta density will be written as the $\mathrm{Be}(\delta_0 m, \delta_0(1-m))$, which is regarded also as the $\mathrm{EDM}(m,\delta_0)$ in (1.2) by setting $\eta = \log\{p/(1-p)\}$, $\kappa(m,\delta_0) = m^{\delta_0 m}(1-m)^{\delta_0(1-m)}/B(\delta_0 m, \delta_0(1-m))$ and $b(p) = 1/\{p(1-p)\}$. Then the posterior density is the $\mathrm{EDM}(\hat{p}_L, n+\delta_0)$ with $\hat{p}_L = (x+\delta_0 m)/(n+\delta_0)$, which is the posterior mean of the mean parameter p. This familiar supporting measure is not Jeffreys' prior, but is still a non-informative one; see Takeuchi and Amari (2005; Example 2) and Yanagimoto and Ohnishi (2009; Section 6) for recent discussions.

A familiar parameter transformation in this model is the logit one, $\theta(=\eta) = \log\{p/(1-p)\}$. Recall that p is the mean parameter, and that η is the canonical parameter. The equivalent prior density is given by replacing p and $b(p)$ by η and 1 in (2.6), respectively. This conjugate prior density is standard, and the above supporting factor is regarded as the Jacobian. The posterior mean of η, $\hat{\eta}$, is expressed as $\psi(x+\delta_0 m) - \psi(n-x+\delta_0(1-m))$ in terms of the digamma function $\psi(\cdot)$. The predictor-based density of θ becomes the $\mathrm{EDM}(\hat{\mu}, n)$, as will be stated in Proposition 3.1.

The estimate $\hat{\mu} = \exp(\hat{\eta})/\{1+\exp(\hat{\eta})\}$ is induced from $\hat{\eta}$, and is not equal to the posterior mean of the mean parameter $\hat{p}_L$. Superficially, $\hat{p}_L$ may be more appealing than $\hat{\mu}$, since this form is simple and is very familiar in the conjugate analysis. However, it looks difficult to construct an optimal predictor based on the posterior mean of the mean parameter.

3. Some elementary properties

Some elementary properties and two additional examples are given here to aid our understanding of the proposed credible region. Suppose that the sampling density $p(\boldsymbol{x};\theta)$ is in the EF of the form in (1.1) with $\eta = \theta$. We assume that a prior density is the $\mathrm{EDM}(m,\delta_0)$, which is conjugate. A notable property is that an induced density of θ is in the same family as a prior density, which may be called that it satisfies the closure property under sampling.

Proposition 3.1. *Suppose that the sampling density is in the EF in* (1.1) *and also that a prior density is the* $\mathrm{EDM}(m,\delta_0)$ *in* (1.2). *Then the predictor-based density of* θ, $\pi_e(\theta \mid \boldsymbol{x})$, *is the* $\mathrm{EDM}(\hat{\mu}, n)$ *with* $\hat{\mu} = M_\eta(\hat{\eta})$. *Thus it satisfies the closure property under sampling.*

Recall that the predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$ is estimative under the same conditions in the above proposition. This property can be regarded as a type of the closure property under sampling. Consequently, all the five densities necessary to define the predictive credible region are included in the two families to define a Bayesian model.

There are two differences between the posterior and the predictor-based densities of θ. One concerns the position parameters of the densities, and the other pertains to the dispersion parameters. The latter difference appeared in all the examples in the previous section, but the former one appeared only in the binomial case (Example 2.3). One of the reasons is because the posterior means of the canonical and the mean parameters are not equivalent in the binomial case. A notion of a family of prior densities, proposed by Yanagimoto and Ohnishi (2005), is useful to elucidate this point. A family of prior densities of the form $\pi(\eta; m, \delta_0)$ is called to be c-preserving, if

$$\mathrm{E}\{\eta \mid \pi(\eta; m, \delta_0)\} = N_\mu(m)$$

for every m and δ_0. To obtain a sufficient condition, consider the proper $\mathrm{EDM}(m, \delta_0)$ of the form

$$\pi(\eta; m, \delta_0) = \exp[\delta_0\{m\eta - M(\eta) - N(m)\}]b(\eta)\kappa(\delta_0). \tag{3.1}$$

Note that the normalizing is independent of m.

Proposition 3.2. *Suppose that a prior density is the proper* $\mathrm{EDM}(m, \delta_0)$, *and set* $\hat{\mu}_L = (n\bar{t} + \delta_0 m)/(n + \delta_0)$. *Then it holds that* $\hat{\mu}(= M_\eta(\hat{\eta})) = \hat{\mu}_L$. *As a result, the posterior and the predictor-based densities of* θ *become the* $\mathrm{EDM}(\hat{\mu}, n + \delta_0)$ *and the* $\mathrm{EDM}(\hat{\mu}, n)$, *respectively.*

The invariance property under the parameter transformation is important for a credible region, but the HPD credible region is not invariant. Thus we modify it and introduce another posterior credible region, the highest standardized posterior density (HsPD) credible region of the form $C_S = \{\theta \mid p(\boldsymbol{x}; \theta)\exp\{-\delta_0 d(\theta, c)\} \geq \tilde{c}_\alpha\}$. This treatment appeared in Yanagimoto and Ohnishi (2005), where the standardized posterior mode was introduced in the conjugate analysis. Here we formally state that the HsPD credible region is invariant.

Proposition 3.3. *Suppose that a prior density is expressed as* $\pi(\theta) \propto \exp\{-d(\theta, c)\}b(\theta)$ *for an appropriate distance* $d(\cdot, \cdot)$ *and a constant* c, *and also that our interest is in* ξ *with* $\theta = g(\xi)$. *Then an equivalent prior is given by* $\tilde{\pi}(\xi) = \exp\{-d(g(\xi), c)\}\tilde{b}(\xi)$ *with* $\tilde{b}(\xi) = b(g(\xi))g'(\xi)$. *Thus the HsPD credible region is invariant under the parameter transformation.*

As discussed in the previous section, an explicit form of the critical value c_α is possible in various familiar cases. However, similarly to obtaining a critical value of the LRT, an asymptotic formula can be applied to yielding an approximate of

c_α under mild regularity conditions. The following proposition states our claim in a formal way.

PROPOSITION 3.4. *Suppose that $d(\theta \mid \boldsymbol{x})$ is in the C^2 class, that is, it has the continuous second derivative. Then the critical value c_α is approximated by $\chi^2_{1-\alpha}$ when the sample size n is large.*

PROOF. Let $p_e(y_i \mid \boldsymbol{x})$ be the optimum predictor of the i-th component under the e-divergence loss, which is proportional to $\exp[\mathrm{E}\{\log p(y_i;\theta) \mid \pi(\theta \mid \boldsymbol{x})\}]$. Then it holds that

$$p_e(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} p_e(y_i \mid \boldsymbol{x}). \tag{3.2}$$

Note that this factorization property implies the following expression:

$$d(\theta \mid \boldsymbol{x}) = n\mathrm{D}(p_e(y_1 \mid \boldsymbol{x}), p(y_1;\theta)).$$

Set $g(\theta \mid \boldsymbol{x}) = (1/n)\{\mathrm{D}(p_e(\boldsymbol{y} \mid \boldsymbol{x}), p(\boldsymbol{y};\theta)) - \mathrm{D}(p_e(\boldsymbol{y} \mid \boldsymbol{x}), p(\boldsymbol{y};\check{\theta}(\boldsymbol{x}))\}$ with $\check{\theta}(\boldsymbol{x})$ being the minimizer of $d(\theta \mid \boldsymbol{x})$. We calculate the moment generating function of $T = 2ng(\theta \mid \boldsymbol{x})$, and show that it converges to that of the χ^2 distribution with one degree of freedom as n tends to infinity. Since it holds that $d(\theta \mid \boldsymbol{x}) = ng(\theta \mid \boldsymbol{x}) + d(\check{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})$, we have

$$\mathrm{E}\{\exp(sT) \mid \pi_e(\theta \mid \boldsymbol{x})\} = \frac{\int_\Theta \exp\{-(1-2s)ng(\theta \mid \boldsymbol{x})\}b(\theta)d\theta}{\int_\Theta \exp\{-ng(\theta \mid \boldsymbol{x})\}b(\theta)d\theta}.$$

Note that the probability of an open interval including $\check{\theta}(\boldsymbol{x})$ is close to 1 when n is large. Taylor's expansion gives $g(\theta \mid \boldsymbol{x}) \doteq (1/2)g_{\theta\theta}(\check{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})(\theta - \check{\theta}(\boldsymbol{x}))^2$, and $b(\theta)$ is a positive and continuous function. Therefore, we can apply Laplace approximations to both the numerator and the denominator, see Robert (2001; p. 298). Then it yields

$$\begin{aligned}\mathrm{E}\{\exp(sT) \mid \pi_e(\theta \mid \boldsymbol{x})\} &\doteq \sqrt{\frac{2\pi}{(1-2s)ng_{\theta\theta}(\check{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})}} \Bigg/ \sqrt{\frac{2\pi}{ng_{\theta\theta}(\check{\theta}(\boldsymbol{x}) \mid \boldsymbol{x})}} \\ &= 1/\sqrt{1-2s}\end{aligned}$$

which is the moment generating function of the χ^2 distribution with one degree of freedom. Thus we complete the proof.

Though we assumed the existence of the moment generating function in an implicit way, the assumption can be avoided by applying the characteristic function instead of the moment generating function. The key property in the above proof is the factorization one in (3.2), which comes from the use of the logarithmic transformation in defining the optimum predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$.

We close this section by giving two examples supplementing those in the previous section.

Example 3.1 (von Mises distribution). Consider the von Mises sampling density $\mathrm{vM}(\theta, \tau_0)$ of the form

$$p(x;\theta,\tau_0) = \exp\{\tau_0 \cos(x-\theta)\}/\{2\pi I_0(\tau_0)\} \qquad (0 \le x < 2\pi) \tag{3.3}$$

where θ is the mean direction $(0 \le \theta < 2\pi)$, τ_0 (> 0) is the known concentration parameter and $I_0(\cdot)$ denotes the modified Bessel function of the first kind and order zero. Note that the density (3.3) is not in the EF but in the curved EF included in the EF having the canonical parameter $\boldsymbol{\eta} = (\tau_0 \cos(\theta), \tau_0 \sin(\theta))$. Let $\boldsymbol{x}$ be a sample vector of size n, and define $\bar{x}$ and $R(\boldsymbol{x})$ by the equalities, $R(\boldsymbol{x})\cos(\bar{x}) = \sum \cos(x_i)$ and $R(\boldsymbol{x})\sin(\bar{x}) = \sum \sin(x_i)$. Then $(\bar{x}, R(\boldsymbol{x}))$ is a sufficient statistic.

Assume a prior density $\mathrm{vM}(c, \delta_0)$, which is regarded as a conjugate prior, see Rodrigues *et al.* (2000). To describe the posterior density, define $\check{\theta}$ and δ^* by the equalities, $\delta^* \cos(\check{\theta}) = \tau_0 R(\boldsymbol{x}) \cos(\bar{x}) + \delta_0 \cos(c)$ and $\delta^* \sin(\check{\theta}) = \tau_0 R(\boldsymbol{x}) \sin(\bar{x}) + \delta_0 \sin(c)$. Then it follows that the posterior density is $\mathrm{vM}(\check{\theta}, \delta^*)$, and that the optimum predictor is written as

$$p_e(\boldsymbol{y} \mid \boldsymbol{x}) = \exp\{\tau_0 A(\delta^*) R(\boldsymbol{y}) \cos(\bar{y} - \check{\theta})\}/\{(2\pi)^n I_0^n(\tau_0 A(\delta^*))\} \qquad (\boldsymbol{y} \in [0, 2\pi)^n)$$

with $A(\tau) = I_0'(\tau)/I_0(\tau)$. Thus this predictor is not estimative, though components of $\boldsymbol{y}$ follow mutually independently the von Mises density $\mathrm{vM}(\check{\theta}, \tau_0 A(\delta^*))$. Then the predictor-based density of θ becomes $\mathrm{vM}(\check{\theta}, n\tau_0 A(\tau_0 A(\delta^*)))$. As a result, the predictor-based density satisfies the closure property under sampling.

Example 3.2 (Uniform distribution). The last example treats the uniform distribution $U(0, \theta)$, which is neither in the EF nor in the curved EF. Assume a Pareto prior density on θ, $\pi(\theta; a, b)$ $(= p_P(\theta; a, b)) = ab^a/\theta^{1+a}$ for $\theta > b$, where a and b are positive. This prior density is of the form in (2.2) by setting that $\delta_0 = a$, $d(\theta, b) = \log(\theta/b)$ and $b(\theta) = 1/\theta$. Write the largest order statistic from a sample of size n as $x_{(n)}$, and set $z = \mathrm{Max}\{x_{(n)}, b\}$. It follows that the posterior density is $p_P(\theta; n + a, z)$, and that the optimum predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$ follows $U^n(0, z)$. Since the distance $d(\theta \mid \boldsymbol{x})$ is expressed as $n \log(\theta/z)$, the predictor-based density of θ is $p_P(\theta; n, z)$. Thus the predictive incredible region is expressed as

$$I = \{\theta \mid \theta^n > \alpha z^n\}. \tag{3.4}$$

In this case the predictive incredible region becomes an upper half line.

The LRT statistic is written as $2n \log(\theta/x_{(n)})$, and the rejection region is given by $R = \{\boldsymbol{x} \mid \theta^n > \alpha x_{(n)}^n\}$. In other words, the rejection region R is formally obtained by replacing the Bayesian estimate z in (3.4) by the maximum likelihood estimate $x_{(n)}$.

4. Competitors

To examine the role of the predictive credible region, we compare it with other procedures applicable to the problem of diagnosing a hypothesis. The

present comparison study involves various difficulties, since the procedures in this comparison study are derived under largely different disciplines. Thus when no confusion is anticipated, we do not distinguish the following three judgments; 1) a hypothesis $\mathrm{H}_0 : \theta = \theta_0$ is rejected, 2) θ_0 is not in the credible (confidence) region, and 3) an alternative model is chosen.

We will consider a simple case where the sampling density $p(\boldsymbol{x}; \eta)$ is in the EF with $\eta = \theta$ in (1.1), and assume a prior density $\pi(\eta; m, \delta_0)$ is in the proper EDM in (3.1). A mixed prior density can be assumed in the Bayesian model selection, which is expressed as

$$\pi(\eta; q, \mu_0, \delta_0) = q\delta_D(\eta - \eta_0) + (1-q)\pi(\eta; \mu_0, \delta_0) \tag{4.1}$$

with $\mu_0 = M_\eta(\eta_0)$. Recall that this is a specific form of (2.3). Following Chacon *et al.* (2007), we will call a prior density to be precise when $q = 0$.

4.1. HPD credible region

The HPD credible region is familiar in the Bayesian literature, see Bernardo and Smith (2000) for example. This credible region is regarded as a Bayesian version of the confidence region in the frequentist context. A formal application of the HPD credible region allows us to diagnose a hypothesis. The HPD method gives a plausible region by $\pi(\eta \mid \boldsymbol{x}) > c$ for an appropriate value c. We choose the value $c = c_\alpha$ such that $\Pr\{C_H \mid \pi(\eta \mid \boldsymbol{x})\} \geq 1-\alpha$ with $C_H = \{\eta \mid \pi(\eta \mid \boldsymbol{x}) > c_\alpha\}$. A sufficient condition for this credible region to attain the exact level $1-\alpha$ is that a prior density is precise. The hypothesis H_1 is chosen when $\eta_0 \notin C_H$.

When a prior density is precise and is in the proper EDM, Proposition 3.2 states that the predictor-based and the posterior densities of η share the common position parameter $\hat{\mu}$ but have the different dispersion parameters n and $n + \delta_0$, respectively. The HsPD credible region results in the form $C_S = \{\eta \mid (n + \delta_0)\{M(\eta) + N(\hat{\mu}) - \hat{\mu}\eta\} \geq \tilde{c}_\alpha\}$ for a suitable $\tilde{c}_\alpha$. The HsPD credible region C_S is of the same form as the predictive credible region C, but they have different critical values. When the χ^2-approximation can be applied, we may set $\tilde{c}_\alpha \doteq \{(n + \delta_0)/n\}\chi^2_{1-\alpha}$.

When a mixed prior density (4.1) is assumed, the difference becomes sharp. The posterior density is a mixed density with a posterior mixing probability at η_0. Since the posterior mixing probability is positive for every $\boldsymbol{x}$, η_0 is always included in the HPD credible region for every $\alpha < 1$. On the other hand, $\pi_e(\eta \mid \boldsymbol{x})$ is expressed as $\pi(\eta; \hat{\mu}, n)$, which is of the same form as in the case of a precise prior density.

4.2. Likelihood ratio test

The LRT statistic is written as

$$\mathrm{LRT} = 2\log\left\{\frac{p(\boldsymbol{x}; \hat{\eta}_M)}{p(\boldsymbol{x}; \eta_0)}\right\}.$$

The rejection region is expressed as $R = \{\boldsymbol{x} \mid \mathrm{LRT} > c_\alpha\}$ for a critical value c_α. The χ^2-approximation is widely employed in obtaining the critical value, and the null hypothesis H_0 is rejected when $\mathrm{LRT} > \chi^2_{1-\alpha}$.

We can find superficial but close correspondences between derivations of the predictive incredible region I and of the rejection region R. In fact, we consider the density of the form

$$p(\boldsymbol{x};\eta_0) = \exp\{-\mathrm{LRT}/2\}p(\boldsymbol{x};\hat{\eta}_M).$$

Then we learn that $p(\boldsymbol{x};\eta_0)$ and LRT/2 correspond to $\pi_e(\eta \mid \boldsymbol{x})$ and $d(\eta \mid \boldsymbol{x})$, respectively. More interestingly, the term $p(\boldsymbol{x};\hat{\eta}_M)$ corresponds formally to the supporting factor $b(\eta)$. Note also that the forms of $d(\eta_0 \mid \boldsymbol{x})$ and LRT/2 are common, when $\hat{\eta}$ and $\hat{\eta}_M$ are equivalent.

The LRT statistic owes heavily on the MLE of η, though there is no general optimality property of the MLE under a finite sample size. This is to be compared with the fact that the posterior mean of η and the predictor $p_e(\boldsymbol{y} \mid \boldsymbol{x})$ satisfy their own optimalities.

4.3. Bayes factor

The Bayes factor is familiar in Bayesian theory, see Kass and Raftery (1995), Chacon *et al.* (2007) and Johnson (2008). Write the prior density under a hypothesis H_i as $\pi(\eta \mid \mathrm{H}_i)$ for $i = 1, 2$. A prior density corresponding to the null hypothesis $\pi(\eta \mid \mathrm{H}_0)$ is written as $\delta_D(\eta - \eta_0)$, and that corresponding to the alternative hypothesis $\pi(\eta \mid \mathrm{H}_1)$ is given by an appropriate density. Then the Bayes factor is expressed as

$$\mathrm{BF} = \frac{\mathrm{E}\{p(\boldsymbol{x};\eta) \mid \pi(\eta \mid \mathrm{H}_0)\}}{\mathrm{E}\{p(\boldsymbol{x};\eta) \mid \pi(\eta \mid \mathrm{H}_1)\}}$$

and the alternative hypothesis H_1 is chosen when $\mathrm{BF} < 1$. Both the numerator and the denominator become the marginal densities of the sample under different priors.

In spite of the importance of the Bayes factor, it looks necessary to examine conditions of a situation where the Bayes factor can be suitably applied to diagnosing a hypothesis. To discuss this, we recall the fundamental relation of the theory of Bayesian statistics: $p(\boldsymbol{x};\eta)\pi(\eta \mid \mathrm{H}_i) = \pi(\eta \mid \boldsymbol{x}, \mathrm{H}_i)p_m(\boldsymbol{x} \mid \mathrm{H}_i)$ for $i = 1$ or 2. Then the posterior densities $\pi(\eta \mid \boldsymbol{x}, \mathrm{H}_i)$ for $i = 1, 2$ are essential in inferential procedures of η, whilst the roles of $p_m(\boldsymbol{x} \mid \mathrm{H}_i)$ for $i = 1, 2$ are usually considered to be marginal. Thus it is necessary to pursue behaviors of the ratio of the two marginal densities.

5. An application: Lindley's paradox

Suppose that $\boldsymbol{x}$ is a sample vector of size n from a normal population $\mathrm{N}(\mu, 1/\tau_0)$. Let $\mathrm{H}_0 : \mu = \mu_0$ be the null hypothesis and $H_1 : \mu \neq \mu_0$ be the alternative hypothesis. It is believed through Lindley's paradox that the performance of the Bayes factor is superior to a standard statistical test when n is large. To avoid potential incommutability between largely different contexts, we make the procedures as simple as possible so far as we maintain their characteristics and advantages.

Our aim here is to show that the predictive incredible region provides us with a novel view of this controversial issue. We begin with our discussions under the conditions often employed in existing works, and develop them under other conditions where our comparison studies are simplified or generalized. Speculations of the following comparison study will be summarized in the Subsection 5.4.

5.1. Traditional case

We assume a mixed prior density of the following form:

$$\pi(\mu; q, \mu_0, \delta_0) = q\delta_D(\mu - \mu_0) + (1-q)p_N(\mu; \mu_0, 1/\delta_0)$$

where q $(0 < q < 1)$ is the mass probability at μ_0. Then the posterior density is given by

$$\pi(\mu; q, \mu_0, \delta_0 \mid \boldsymbol{x}) = Q_0\delta_D(\mu - \mu_0) + (1-Q_0)p_N(\mu; \hat{\mu}_L, 1/(n\tau_0 + \delta_0))$$

where $\hat{\mu}_L = (n\tau_0\bar{x} + \delta_0\mu_0)/(n\tau_0 + \delta_0)$ and

$$Q_0 = \frac{qp_N(\bar{x}; \mu_0, 1/n\tau_0)}{qp_N(\bar{x}; \mu_0, 1/n\tau_0) + (1-q)p_N(\bar{x}; \mu_0, (n\tau_0 + \delta_0)/(n\tau_0\delta_0))}.$$

Then it follows that the HPD credible region is expressed as

$$C_H = \{\mu_0\} \cup \{\mu \mid (n\tau_0 + \delta_0)(\hat{\mu}_L - \mu)^2 < \chi^2_{(1-Q_0-\alpha)/(1-Q_0)}\} \tag{5.1}$$

when Q_0 is less than $1 - \alpha$. Otherwise, it becomes a singleton $\{\mu_0\}$.

To obtain simple forms of the incredible and the rejection regions under study, set $\mathrm{Z} = \sqrt{n\tau_0}(\bar{x} - \mu_0)$, which follows the standard normal distribution under the hypothesis H_0. The four procedures for rejecting the null hypothesis, including the proposed one, are expressed as follows.

Proposed method:

$$\mathrm{Z}^2 > \frac{(n\tau_0 + \delta_0)^2}{(1-Q_0)(n\tau_0)^2}\chi^2_{1-\alpha}.$$

Bayes factor (BF) method:

$$\mathrm{Z}^2 > \frac{n\tau_0 + \delta_0}{n\tau_0}\log\left(\frac{n\tau_0 + \delta_0}{\delta_0}\right).$$

Highest posterior density (HPD) method:

$$\mathrm{Z}^2 < 0.$$

Likelihood ration test (LRT) method:

$$\mathrm{Z}^2 > \chi^2_{1-\alpha}.$$

We will denote the "rejection regions" of the BF and the LRT methods by $R(\mathrm{BF})$ and $R(\mathrm{LRT})$, respectively. Then the claim of Lindley's paradox is

reviewed as follows. When the null hypothesis is true, the probability $\Pr\{R(\mathrm{BF})\}$ tends to 0 as n tends to infinity, whilst $\Pr\{R(\mathrm{LRT})\}$ takes the value α for every n. In addition, when an alternative hypothesis is true, both the probabilities tend to 1 as n tends to infinity. This consistency property of the BF method has been claimed as a desirable one.

We can find a large difference between the HPD and the BF methods under certain conditions. Write the equality (5.1) as $C_H = \{\mu_0\} \cup C_R$, and consider a case where the BF takes the value 1 and the LRT takes a large value. Then it is likely to hold that $\mu_0 \notin C_R$, as was pointed out in Yanagimoto (1999). A defect of the HPD method is that the null hypothesis is never rejected, when q is positive. Therefore, the HPD credible region performs less satisfactorily, when a prior density is not precise.

Next, we consider cases where the mixing probability q varies. A notable fact is that "the critical value" of the BF method is independent of q. On the contrary, the critical value of the proposed method increases, and therefore the predictive credible region C becomes wider as q increases.

Though it is difficult to compare the proposed method with the LRT one, we can state that the proposed one is better than the LRT one, when a prior density is dependable.

5.2. Precise case

Suppose that a prior density is $p_N(\mu; \mu_0, 1/\delta_0)$. Then the incredible and rejection regions of the two methods are simplified as follows.
Proposed method:

$$\mathrm{Z}^2 > \frac{(n\tau_0 + \delta_0)^2}{(n\tau_0)^2}\chi^2_{1-\alpha}.$$

HPD method:

$$\mathrm{Z}^2 > \frac{n\tau_0 + \delta_0}{n\tau_0}\chi^2_{1-\alpha}.$$

A notable fact is that the proposed method behaves more closely to the HPD method than the other two methods. Although these similar forms of the inequalities come partly from the normality assumption, this fact shows a close relationship between the proposed and the HPD methods; both the methods rely heavily on the posterior density.

On the other hand, we find large differences among these two methods and the BF method. In fact, the critical values for the former two are decreasing in n and increasing in δ_0, whilst the reverse properties hold for the BF method. Note that a larger value of δ_0 represents our belief in a smaller variance of a prior density whose position parameter is located at μ_0.

The critical values of the proposed and the BF methods are functions of $n\tau_0$ and δ_0 only through the ratio $r = n\tau_0/\delta_0$. The function of the proposed method is strictly decreasing in r; it takes the values ∞, $4\chi^2_{1-\alpha}$ and $\chi^2_{1-\alpha}$ at 0, 1, and ∞, respectively. On the other hand, that of the BF method is strictly increasing in r; it takes the values 1, $2\log 2$ and ∞ at 0, 1, and ∞, respectively. They cross at $r = 49.36$. Consider an asymptotic case where r is fixed and n tends to infinity.

Then the BF method becomes inconsistent, as was pointed out in Yanagimoto (1999) and Johnson (2008).

Though the consistency property was traditionally discussed under a true model where $\mu_a = \mu_0 + a$ for an arbitrary value of a and a large number n, we may consider the true model where $\mu_a = \mu_0 + a/\sqrt{n}$. Then the asymptotic coverage probability of I is greater than α, but the probability that the BF method rejects the null hypothesis is asymptotically 0 for every fixed a.

5.3. Case of a general prior mean

Next, we assume that the mean of a prior density $p_N(\mu; \mu_1, 1/\delta_0)$ is not necessarily equal to μ_0. This case is not discussed in depth in the controversy regarding Lindley's paradox, though such a prior density is realistic. By avoiding any unnecessary restriction on a prior density, we can expect a different view of this controversy.

Set $\tilde{W}^2 = 2\mathrm{D}(p(\boldsymbol{y}; \hat{\mu}), p(\boldsymbol{y}; \mu_0)) = n\tau_0(\hat{\mu} - \mu_0)^2$ with $\hat{\mu}(= \hat{\mu}_L) = (n\tau_0\bar{x} + \delta_0\mu_1)/(n\tau_0 + \delta_0)$. This statistic is expressed in terms of Z as

$$\tilde{W}^2 = \frac{(n\tau_0)^2}{(n\tau_0 + \delta_0)^2} \left\{ Z + \frac{\delta_0}{\sqrt{n\tau_0}}(\mu_1 - \mu_0) \right\}^2 .$$

To obtain simple expressions of the incredible and rejection regions, we set $W^2 = \{(n\tau_0 + \delta_0)/(n\tau_0)\}^2 \tilde{W}^2$. Then the parallel calculations to the previous subsection yield the incredible and the rejection regions.

Proposed method:

$$W^2 > \frac{(n\tau_0 + \delta_0)^2}{(n\tau_0)^2} \chi^2_{1-\alpha}.$$

BF method:

$$W^2 > \frac{n\tau_0 + \delta_0}{n\tau_0} \left\{ \log\left(\frac{n\tau_0 + \delta_0}{\delta_0} \right) + \delta_0(\mu_1 - \mu_0)^2 \right\}.$$

HPD method:

$$W^2 > \frac{n\tau_0 + \delta_0}{n\tau_0} \chi^2_{1-\alpha}.$$

We observe that the right-hand sides of the inequalities in the cases of the proposed and the HPD methods remain unchanged though a new prior density is employed. The differences are in the left-hand sides, where Z^2 is replaced by W^2 in each inequality. The sampling distribution of W^2 under the null hypothesis is the χ^2 distribution with the non-centrality parameter $\{\delta_0(\mu_1 - \mu_0)\}^2/n\tau_0$, and that under an alternative hypothesis $\mu = \mu_a$ has the non-centrality parameter $\{n\tau_0(\mu_a - \mu_0) + \delta_0(\mu_1 - \mu_0)\}^2/n\tau_0$. Thus W^2 is stochastically larger than Z^2 under the null hypothesis. Similarly, this property holds under an alternative hypothesis when $(\mu_a - \mu_0)\{n\tau_0(\mu_a - \mu_0) + 2\delta_0(\mu_1 - \mu_0)\} \geq 0$, but the reverse property holds otherwise.

In contrast, a positive term is added in the right-hand side in the case of the BF method. Consider a case where δ_0 is not small but $\delta_0/\sqrt{n\tau_0}$ is small. Then these conditions yield that our belief in the position parameter μ is in favor of the alternative hypothesis, but that the null hypothesis is less likely to be rejected.

5.4. Speculations

Based on the findings obtained in the previous subsections, we summarize speculations on the present comparison study, primarily on that between the proposed and the BF methods.

So far as our present study is concerned, we do not find any unexpected behavior of the proposed method. The present study covers cases of a mixed prior density, various values of the sample size n, δ_0 and the ratio $r = n\tau_0/\delta_0$.

On the contrary, we observed that the critical value of the BF method is independent of the mass probability q of a prior density, and that it is decreasing in δ_0. These findings contradict the fact that larger values of q or δ_0 represent our stronger belief in the null hypothesis. Next, for a fixed δ_0 our belief in the null hypothesis becomes weaker, when μ_1 becomes farther from μ_0. However, the critical value of the BF method unexpectedly increases. In summary, our stronger belief in a hypothesis does not yield our preference of choosing the hypothesis in the BF method. Further, a smaller value of n is unexpectedly associated with a larger power of the BF method.

The consistency property is probably a key reason why the BF method is widely accepted by researchers. However, it should be noted that the property can cause the separation of the HPD credible region, as we noted. It looks that behaviors of the BF and the HPD methods can contradict each other. Another consequence of the consistent property is a small power, when a true mean μ_a is $\mu_0 + a/\sqrt{n}$ for a large n.

Examining cases of various values of the ratio r provides us with another view. When the position parameter of a prior density is μ_0, we can say that the ratio represents the amount of information from the observation $\boldsymbol{x}$ to those corresponding to the strength of our belief in the null hypothesis. Unexpectedly, the power of the BF method decreases under the null hypothesis, as r increases. Further, the BF method is not consistent when r is bounded.

Finally, we discuss the HPD method. The fact that the null hypothesis μ_0 is always included in the C_H for a mixed prior density is an undesirable property as a diagnostic method. It looks as through the assumption on the positive mass at μ_0 is too strong to reject the null hypothesis based on a sample of finite size. A naive way to dissolve this difficulty is to avoid a mixed prior density, but it is undesirable because it places an unnecessary restriction on our choice of a prior density. Another way will be discussed later in the Subsection 7.1.

6. Another application: Additional evidence

To strengthen our claim that the proposed credible region can be useful under various conditions where the existing method does not perform satisfactorily, we discuss another possible application, which pertains to the analysis of an experiment followed by a non-significant result.

This problem involves fundamental issues of statistical inference. In fact, one of the accepted rules is that once a significance level is fixed, it should be kept throughout the research. Under this rule many sophisticated methods have been

developed. Cornfield (1966) however criticized this rule, and stated "no amount of additional evidence can be collected", if an experiment was once finished without showing significant difference. Thus a lot of researchers completely neglect the principle of statistical testing because of too restricted formal rules. It looks that these two rules are obviously unsatisfactory in various practical situations, and further detailed studies are desired. We do not attempt here to reach a definite conclusion, but to point out that our approach can shed new light on this challenging problem. Thus our attention here focuses on the performance of the proposed method under a simple situation.

Consider the statistical test problem for the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ against the one-sided alternative hypothesis $\mathrm{H}_1 : \mu > \mu_0$ in the normal population with variance 1. The situation is outlined as follows: An experimenter already has a previous observation of size n, $\boldsymbol{y} = (y_1, \ldots, y_n)$, which showed a non-significant result, that is, $\boldsymbol{y} \notin R_0 = \{\boldsymbol{y} \mid \sqrt{n}(\bar{y} - \mu_0) > z_\alpha\}$ with $\Phi(z_\alpha) = 1 - \alpha$. However, he/she still hopes to analyze a present observation $\boldsymbol{x}$ of size n.

Set α and β ($\alpha + \beta < 1$) to be the type I and the type II errors, respectively, and set $\sqrt{n}(\mu_T - \mu_0) = z_\alpha + z_\beta$. The mean μ_T was used for determining the sample size allowing the power of the previous experiment $1 - \beta$. Therefore, it is regarded as a representative value of the alternative hypothesis. We consider the two rejection regions: $R_1 = \{\boldsymbol{x} \mid \sqrt{n}(\bar{x} - \mu_0) > z_\alpha\}$ and $R_2 = \{\boldsymbol{x} \mid \sqrt{n}(\bar{z} - \mu_0) > z_\alpha\}$ with $\bar{z} = (\bar{x} + \bar{y})/2$, which are derived from neglecting the result of the previous experiment and from the proposed method, respectively. To apply the proposed method to this problem, we elicit a practical prior density, in terms of the likelihood function of the previous experiment, as

$$\pi(\mu) \propto \prod p(y_i; \mu). \tag{6.1}$$

This prior is conjugate, and allows us an explicit expression of the rejection region. Since the alternative hypothesis is one-sided, the directed version of the incredible region I will be used.

Write the regions $\{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \in R_i \text{ and } \boldsymbol{y} \in R_0^c\}$ with R_0^c being the complement of R_0 as $R_i \otimes R_0^c$ for $i = 1$ and 2. We calculate the probabilities $\mathrm{P}(R_1 \otimes R_0^c)$ and $\mathrm{P}(R_2 \otimes R_0^c)$ under the null and the selected alternative hypotheses. It follows that $\mathrm{P}(R_1 \otimes R_0^c \mid \mathrm{H}_0) = \alpha(1-\alpha)$ and $\mathrm{P}(R_1 \otimes R_0^c \mid \mathrm{H}_1) = (1-\beta)\beta$. The two other probabilities are represented explicitly as

$$\mathrm{P}(R_2 \otimes R_0^c \mid \mathrm{H}_0) = 1 - \alpha - \frac{(1-\alpha)^2 + \Phi(\sqrt{2}z_\alpha)}{2}$$

and

$$\mathrm{P}(R_2 \otimes R_0^c \mid \mathrm{H}_1) = \beta - \frac{\beta^2 + \Phi(\sqrt{2}z_{1-\beta})}{2}.$$

Note that these two formulas are mathematically common.

Consider the case that $\alpha = 0.05$ and $\beta = 0.2$, where the above four probabilities become .0475, .16, .00375 and .122, respectively. A notable observation is that the actual type I error of the proposed method is slightly greater than

the nominal one, whilst that of the former procedure is nearly twice the nominal one. On the other hand, the powers of these two methods are fairly close. We calculate the ratios, $r_i = \mathrm{P}(R_i \otimes R_0^c \mid \mathrm{H}_1)/\mathrm{P}(R_i \otimes R_0^c \mid \mathrm{H}_0)$ for $i = 1$ and 2. This ratio is associated with the false discovery rate in Benjamini and Hochberg (1995), which provides us with a view different from a statistical test. For reference we also calculate $r_0 = \mathrm{P}(R_0 \mid \mathrm{H}_1)/\mathrm{P}(R_0 \mid \mathrm{H}_0)$. The ratios r_i for $i = 0$, 1 and 2 become 16, 3.37 and 32.4, respectively. It looks amazing that the proposed method takes a large number of r_2. This means that the alternative hypothesis is more likely to be true when the null hypothesis is rejected in the analysis of the present experiment. On the contrary, the fact that the former procedure takes a small number of r_1 is discouraging, and resultantly makes us hesitate to use this procedure. In fact, a superficially large power is simply a consequence of the unexpectedly large excess of the significance level. Note that the difference among ratios becomes sharper when smaller significance errors α and β are selected.

Next, we discuss behaviors of the two procedures given the value of $T_y = \sqrt{n}(\bar{y}-\mu_0)$. When the value is very close to the critical value z_α, the probabilities $\mathrm{P}(R_1 \mid \mathrm{H}_1, T_y)$ and $\mathrm{P}(R_2 \mid \mathrm{H}_1, T_y)$ are close. However, as the value decreases, the former probability remains constant and the latter decreases rapidly. This behavior of the proposed procedure is reasonable, since our belief in the alternative hypothesis based on the former experiment decreases as T_y decreases. So far as our studies are concerned, the performance of the proposed procedure is always reasonable under various situations.

Two other standard procedures provide the rejection regions: $\sqrt{2n}(\bar{z}-\mu_0) > z_\alpha$ and the empty set. The former is derived by the HPD method by assuming a prior density of the form in (6.1). Both regions attract our attention, but we do not further discuss here any more due to existing controversies relating to basic scientific inference.

We complete this example by referring to the BF method. A prior density corresponding to that based on the previous experiment in (6.1) in the case of the one sided alternative hypothesis $\mathrm{H}_1 : \mu > \mu_0$ is written as $p_N(\mu; \bar{y}, 1/n)/(1 - \Psi(n(\mu_0 - \bar{y}))$ on $[\mu_0, \infty)$. Then the Bayes factor is written as

$$\log(\mathrm{BF}) = \log\left\{\frac{\sqrt{2}(1 - \Psi(n(\mu_0 - \bar{y})))}{1 - \Psi(2n(\mu_0 - \bar{z}))}\right\} - \frac{n}{4}\{(\bar{x} - \mu_0)^2 - (\bar{y} - \mu_0)^2 + 2(\bar{x} - \mu_0)(\bar{y} - \mu_0)\}.$$

The null hypothesis is rejected when log(BF) takes a negative value. The region is largely different from others discussed above, and does not look attractive. For example, log(BF) is not necessarily decreasing in $\bar{y}$.

7. Discussions

To supplement the role of the predictive credible region, we discuss its potential modification, where we apply the posterior density to determine a critical value c_α in Definition 2.2. Then, some remarks on the case where the parameter is a vector and the null hypothesis is composite are given.

7.1. Use of the posterior density

Though a critical value c_α was determined in terms of the predictor-based density of θ, we can apply the posterior density as an alternative. A formal definition is given as follows.

DEFINITION 7.1. *For a given significance level* α, *a modified predictive credible region is given by*

$$C_M = \{\theta \mid 2\{d(\theta \mid \boldsymbol{x}) - d(\check{\theta} \mid \boldsymbol{x})\} \le c_\alpha(\boldsymbol{x})\} \tag{7.1}$$

where $c_\alpha(\boldsymbol{x})$ *satisfies* $\Pr\{C_M \mid \pi(\theta \mid \boldsymbol{x})\} \ge 1 - \alpha$.

Note that this credible region does not always attain the exact level $1 - \alpha$, since the posterior density can be a mixed one of the form in (2.3). Obviously, it is expected that this credible region is close to the HsPD credible region. The following proposition presents a sufficient condition of equivalency for them.

PROPOSITION 7.1. *Consider a conjugate analysis model given by* (1.1) *and* (1.2), *and suppose that* $\hat{\mu}_L = M_\eta(\hat{\eta})$. *Then the modified posterior credible region* C_M *is equivalent with the HsPD credible region* C_S.

The modified posterior credible region behaves satisfactorily under the traditional case in the Subsection 5.1, though an explicit form of the incredible region under the case becomes rather complicated. First, it is consistent, since the posterior distribution has a positive mass at μ_0. Next, we evaluate the asymptotic coverage probability of the modified posterior incredible region I_M when a true mean is $\mu_a = \mu_0 + a/\sqrt{n}$, as in the Subsection 5.2. It is shown that the probability tends to 1 as a tends to infinity under the traditional case. This means that the null hypothesis can be rejected. In the precise cases in the Subsections 5.2 and 5.3, the modified posterior credible region becomes equivalent with the HPD credible region.

A reason why we do not employ this credible region is because of its restricted applications, due to its very close relation with the HsPD credible region as in the above proposition. Another reason pertains to the logical relation of the definition of the credible region of the form in (7.1) and the posterior density. Though the optimum predictor is induced from the posterior density, the relation of the distance $d(\theta \mid \boldsymbol{x})$ and the posterior density looks indirect.

7.2. The case of the vector parameter $\boldsymbol{\theta}$

An extension to the case of the vector parameter is straightforward except for the notational complexity, when the null hypothesis is simple. When it is composite, careful treatments are necessary to construct a credible region. One of important aspects regarding a composite null hypothesis is discussed below.

Recall that a credible region is constructed only in terms of a prior density representing an alternative hypothesis or a general model. Any explicit prior density is not assumed to represent a null hypothesis or a restricted model. This is to be compared with the Bayes factor, where a prior density $\delta_D(\mu - \mu_0)$

is assumed. This difference requires much to be done in order to extend the proposed method to a case of a composite null hypothesis. To explain more, we revisit the Student t-test statistic, and examine frequentist and Bayesian derivations.

We assume that the sampling density is the normal density $p_N(x;\mu,\sigma^2)$. Then the canonical parameter $\boldsymbol{\eta}=(\eta_1,\eta_2)$ becomes $(\mu/\sigma^2, 1/\sigma^2)$. We write a density of a sample vector $\boldsymbol{x}$ of size n as $p_N(\boldsymbol{x};\mu,\sigma^2)$, which may be written also as $p_N(\boldsymbol{x};\boldsymbol{\eta})$. The MLE of μ is $\bar{x}$, and the conditional MLE given $\bar{x}$ is expressed as $\hat{\sigma}^2_C=\sum(x_i-\bar{x})^2/(n-1)$. The resultant estimates are $\hat{\boldsymbol{\eta}}_A=(\bar{x}/\hat{\sigma}^2_C, 1/\hat{\sigma}^2_C)$ under the alternative hypothesis, and $\hat{\boldsymbol{\eta}}_N=(\mu_0/\hat{\sigma}^2_C, 1/\hat{\sigma}^2_C)$ under the null hypothesis. The MLE $\hat{\sigma}^2_M$ is often employed to obtain the maximized likelihood. However, this estimate is less satisfactory, since the estimator $(\bar{x},\hat{\sigma}^2_C)$ has less risk than the MLE $(\bar{x},\hat{\sigma}^2_M)$ for every μ and σ^2 under the e-divergence loss for $n\geq 4$. Then the test statistic becomes $2D(p_N(\boldsymbol{y};\bar{x},\hat{\sigma}^2_C), p_N(\boldsymbol{y};\mu_0,\hat{\sigma}^2_C))=n(\bar{x}-\mu_0)^2/\hat{\sigma}^2_C$, which is the square of the well known Student t-test statistic.

A naive Bayesian approach to this problem is to assume an improper prior density $\pi(\boldsymbol{\eta})=b(\boldsymbol{\eta})=1/\eta_2^2$, which yields the posterior density as

$$\pi(\boldsymbol{\eta}\mid\boldsymbol{x})\propto \eta_2^{n/2-2}\exp\left\{-\frac{\eta_2}{2}\sum\left(x_i-\frac{\eta_1}{\eta_2}\right)^2\right\}.$$

Then its conditional density function of η_1 given a fixed η_2 is $p_N(\eta_1;\eta_2\bar{x},\eta_2)$, and its marginal density of η_2 is written as the gamma density $p_G(\eta_2;1/\hat{\sigma}^2_C,(n-1)/2)$. It follows that the posterior means of $\eta_2-(1/\hat{\sigma}^2_C)$ and $\eta_1-\eta_2\bar{x}$ vanish, and therefore the resultant estimate becomes the same as $\hat{\boldsymbol{\eta}}_A$. When μ is known as μ_0, we estimate σ^2 by the (marginal) posterior mean of η_2, which yields $\hat{\boldsymbol{\eta}}_N$. Consequently, the optimum predictors in (2.1) are obtained by plugging these estimates in $p_N(\boldsymbol{y};\boldsymbol{\eta})$. Setting $d(\mu\mid\boldsymbol{x})=D(p_N(\boldsymbol{y};\bar{x},\hat{\sigma}^2_C), p_N(\boldsymbol{y};\mu,\hat{\sigma}^2_C))$, we obtain the predictive credible region of μ. The conditions of the two regions have similar forms of inequalities, and the difference is in the critical values. The proposed predictive credible region uses the upper α-point of the χ^2 distribution.

Acknowledgements

The two anonymous reviewers pointed out various places to be clarified, which were useful for improving the presentation.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*, Cambridge University Press, London.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. B*, **57**, 289–300.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Wiley, Chichester.

Chacon, J. E., Montanero, J., Nogales, A. G. and Perez, P. (2007). On the use of Bayes factor in the frequentist testing of a precise hypothesis, *Comm. In Statist.-Theory and Methods*, **36**, 2251–2261.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions, *J. Am. Statist. Assoc.*, **87**, 1123–1127.

Corcuera, J. M. and Giummole, F. (1999). A generalized Bayes rule for prediction, *Scand. J. Statist.*, **26**, 265–279.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle, *Am. Statistician*, **20**, 18–23.

Geisser, S. (1993). *Predictive Inference*: *An Introduction*, Chapman & Hall, New York.

Johnson, V. E. (2008). Properties of Bayes factors based on test statistics, *Scad. J. Statist.*, **35**, 354–368.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *J. Am. Statist. Assoc.*, **90**, 773–795.

Lindley, D. V. (1957). A statistical paradox, *Biometrika*, **44**, 187–192.

Nagaoka, H. and Amari, S.-I. (2000). *Methods of Information Geometry*, AMS, Load Island.

Robert, C. P. (2001). *The Bayesian Choice*, Second ed., Springer, New York.

Rodrigues, J., Leite, J. G. and Milan, L. A. (2000). An empirical Bayes inference for the von Mises distribution, *Austral. New Zealand J. Statist.*, **42**, 433–440.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Shafer, G. (1982). Lindley's paradox (with discussions), *J. Am. Statist. Assoc.*, **71**, 325–351.

Takeuchi, J. and Amari, S. (2005). α-parallel priors and its properties, *IEEE Trans. Inf. Theory*, **51**, 1011–1023.

Yanagimoto, T. (1999). Limitations on the use of Bayesian test under a vague prior distribution, *Proc. Inst. Statist. Math.*, **47**, 81–90 (in Japanese with an English abstract).

Yanagimoto, T. and Ohnishi, T. (2005). Standardized posterior mode for the flexible use of a conjugate prior, *J. Statist. Plann. Inf.*, **131**, 253–269.

Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture, *J. Statist. Plann. Inf.*, **139**, 3064–3075.