

Article

# Bayesian Testing of a Point Null Hypothesis Based on the Latent Information Prior

Fumiyasu Komaki <sup>1,2</sup>

<sup>1</sup> Department of Mathematical Informatics, Graduate School of Information Science and Technology, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan;

E-Mail: komaki@mist.i.u-tokyo.ac.jp; Tel.: +81-3-5841-6941; Fax: +81-3-5841-8592

<sup>2</sup> RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

Received: 9 August 2013; in revised form: 16 September 2013 / Accepted: 10 October 2013 /

Published: 17 October 2013

---

**Abstract:** Bayesian testing of a point null hypothesis is considered. The null hypothesis is that an observation,  $x$ , is distributed according to the normal distribution with a mean of zero and known variance  $\sigma^2$ . The alternative hypothesis is that  $x$  is distributed according to a normal distribution with an unknown nonzero mean,  $\mu$ , and variance  $\sigma^2$ . The testing problem is formulated as a prediction problem. Bayesian testing based on priors constructed by using conditional mutual information is investigated.

**Keywords:** conditional mutual information; discrete prior; Kullback-Leibler divergence; prediction; reference prior; Jeffreys-Lindley paradox

---

## 1. Introduction

We investigate a problem of testing a point null hypothesis from the viewpoint of prediction. The null hypothesis,  $H_0$ , is that an observation,  $x$ , is distributed according to the normal distribution,  $N(0, \sigma^2)$ , with a mean of zero and variance  $\sigma^2$ , and the alternative hypothesis,  $H_1$ , is that  $x$  is distributed according to a normal distribution  $N(\mu, \sigma^2)$  with unknown nonzero mean  $\mu$  and variance  $\sigma^2$ . The variance,  $\sigma^2$ , is assumed to be known. This simple testing problem has various essential aspects in common with more general testing problems and has been discussed by many researchers. An essential part of our discussion in the present paper holds for other testing problems based on more general models.

The assumption that the sample size is one is not essential. When we have  $N$  observations  $x_1, x_2, \dots, x_N$  from  $N(0, \sigma^2)$  or  $N(\mu, \sigma^2)$ , then the sufficient statistic  $\bar{x} = \sum_{i=1}^N x_i / N$  is distributed according to  $N(0, \sigma^2 / N)$  under  $H_0$  or  $N(\mu / N, \sigma^2 / N)$  under  $H_1$ , respectively. Then, the null hypothesis is

that  $\bar{x}$  is distributed according to  $N(0, \tilde{\sigma}^2)$ , and the alternative hypothesis is that  $\bar{x}$  is distributed according to  $N(\tilde{\mu}, \tilde{\sigma}^2)$  ( $\tilde{\mu} \neq 0$ ), where  $\tilde{\sigma}^2 := \sigma^2/N$  and  $\tilde{\mu} := \mu/N$ . Thus, the testing problem with sample size  $N$  is essentially equal to that with the sample size one. From now on, the variance,  $\sigma^2$ , is set to be one without loss of generality.

We formulate the testing problem as a prediction problem. Let  $m = 0$  if  $H_0$  is true and  $m = 1$  if  $H_1$  is true. Let  $w$  be the probability that  $m = 0$ , and let  $\pi(d\mu)$  be the prior probability measure of  $\mu$ . The probability,  $w$ , is set to be  $1/2$  in many previous studies, and the choice of  $\pi(d\mu)$  is discussed; see, e.g., [1] and the references therein. The objective is to predict  $m$  by using a Bayesian predictive distribution,  $p_{w,\pi}(m | x)$ , depending on the prior  $\pi(d\mu)$  and the observation,  $x$ .

Common choices of  $\pi$  are the Normal prior  $(1/\sqrt{2\pi\tau^2}) \exp(-\mu^2/2\tau^2)d\mu$  and the Cauchy prior  $1/\{\pi\gamma(1 + m\mu^2/\gamma^2)\}d\mu$ , recommended by Jeffreys [2]. Sometimes, it is considered that large values of scale parameters  $\tau$  and  $\gamma$  represent “ignorance” about  $\mu$ . However, such a naive choice of scale parameter values could cause a serious problem known as the Jeffreys–Lindley paradox [3].

We choose  $\pi(d\mu)$  from the viewpoint of prediction and construct a Bayesian predictive distribution to predict  $m$  based on an objectively chosen prior. In the testing problem, the variable,  $m$ , is predicted, the variable,  $x$ , is observed and the parameter,  $\mu$ , is neither observed nor predicted. The latent information prior  $\pi^*$  [4] is defined as a prior maximizing the conditional mutual information:

$$I_{m;\mu|x}(w, \pi) = \sum_{m=0}^1 \iint p_{w,\pi}(x, \mu, m | w) \log \frac{p_{w,\pi}(m, \mu | x)}{p_{w,\pi}(m | x)p_{w,\pi}(\mu | x)} dx d\mu \quad (1)$$

between  $m$  and  $\mu$  given  $x$ .

The latent information prior introduced in [4] is an objective Bayes prior. An outline of the method based on it is as follows. First, a statistical problem is formulated as a prediction problem, in which  $x$  is the observed random variable,  $y$  is the random variable to be predicted and  $\theta$  is the unknown parameter. Then, a prior  $\pi(d\theta)$  that maximizes the conditional mutual information  $I_{y;\theta|x}(\pi)$  between  $y$  and  $\theta$  given  $x$  is adopted.

In Section 2, we consider for Kullback-Leibler loss for prediction corresponding to Bayesian testing. In Section 3, we obtain the latent information prior and discuss properties of Bayesian testing based on it. In Section 4, we compare the proposed testing based on the latent information prior with Bayesian testing based on the normal prior and the Cauchy prior.

## 2. Kullback-Leibler Loss of Predictive Densities

We consider Kullback-Leibler loss of prediction corresponding to Bayesian testing. The Bayesian predictive density with respect to  $w$  and  $\pi$  is given by:

$$p_{w,\pi}(m = 0 | x) = \frac{wp_0(x)}{wp_0(x) + (1 - w)p_\pi(x)} \quad (2)$$

and

$$p_{w,\pi}(m = 1 | x) = \frac{(1 - w)p_\pi(x)}{wp_0(x) + (1 - w)p_\pi(x)} \quad (3)$$

where:

$$p_0(x) = \phi(x; 0, 1) \quad \text{and} \quad p_\pi(x) = \int \phi(x; \mu, 1) \pi(d\mu) \quad (4)$$

and  $\phi(x; \mu, \sigma)$  is the density function of the normal distribution,  $N(\mu, \sigma^2)$ .

If the value of  $\mu$  is known, then the alternative hypothesis,  $H_1: N(\mu, 1)$ , becomes a simple hypothesis, and the predictive distribution is given by the posterior:

$$p_w(m = 0 | x, \mu) = \frac{w\phi(x; 0, 1)}{w\phi(x; 0, 1) + (1 - w)\phi(x; \mu, 1)} \quad (5)$$

and:

$$p_w(m = 1 | x, \mu) = \frac{(1 - w)\phi(x; \mu, 1)}{w\phi(x; 0, 1) + (1 - w)\phi(x; \mu, 1)} \quad (6)$$

To evaluate the performance of predictive densities, we adopt the Kullback-Leibler divergence:

$$\sum_{m=0}^1 p_w(m | x, \mu) \log \frac{p_w(m | x, \mu)}{p_{w,\pi}(m | x)} \quad (7)$$

from  $p_w(m | x, \mu)$  and to  $p_{w,\pi}(m | x)$  as a loss function.

The risk function is given by:

$$\begin{aligned} r_w(\mu, \pi) &= \int p_w(x | \mu) \sum_{m=0}^1 p_w(m | x, \mu) \log \frac{p_w(m | x, \mu)}{p_{w,\pi}(m | x)} dx \\ &= \sum_{m=0}^1 w(m) \int p(x | m, \mu) \log \frac{p_w(m | x, \mu)}{p_{w,\pi}(m | x)} dx \end{aligned} \quad (8)$$

where  $w(0) = w$  and  $w(1) = 1 - w$ . Here,  $p_{w,\pi}(m | x, \mu)$  and  $p_{w,\pi}(x | \mu)$  are denoted by  $p_w(m | x, \mu)$  and  $p_w(x | \mu)$ , respectively, because they do not depend on  $\pi$ . The distribution of  $x$  does not depend on  $\mu$  if  $m = 0$ , because  $p(x | m = 0, \mu) = \phi(x; 0, 1)$ .

It is not fruitful to discuss decision theoretic properties, such as the minimaxity of the risk defined by:

$$- \sum_{m=0}^1 w(m) \int p(x | m, \mu) \log p_{w,\pi}(m | x) dx \quad (9)$$

because it is easy to distinguish between  $H_0$  and  $H_1$  when  $|\mu|$  is very large.

The Kullback-Leibler risk in Equation (8) corresponds to the regret type quantity:

$$- \log p_{w,\pi}(m | x) + \log p_w(m | x, \mu) \quad (10)$$

which means the loss by not knowing the value of  $\mu$ . By considering the minimaxity of the regret type risk in Equation (8), several reasonable results are obtained.

**Lemma 1.** The risk of a Bayesian predictive density,  $p_{w,\pi}(m | x)$ , is given by:

$$r_w(\mu; \pi) = w \int p_0(x) \log \frac{1 + \frac{1-w}{w} \frac{p_\pi(x)}{p_0(x)}}{1 + \frac{1-w}{w} \frac{p_0(x-\mu)}{p_0(x)}} dx$$

$$+ (1-w) \int p_0(x) \log \frac{1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_\pi(x+\mu)}}{1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_0(x)}} dx \quad (11)$$

*Proof.* See the Appendix. □

The risk function in Equation (11) is a continuous function of  $\mu$  for every  $w$  and  $\pi$ .

The Bayes risk with respect to a prior  $\pi$  of a Bayesian predictive density based on  $\bar{\pi}$  is:

$$\begin{aligned} R_w(\pi; \bar{\pi}) &= \int r_w(\mu, \bar{\pi}) \pi(d\mu) \\ &= \sum_{m=0}^1 \iint w(m) p(x | m, \mu) \log \frac{p_w(m | x, \mu)}{p_{w, \bar{\pi}}(m | x)} \pi(d\mu) dx \\ &= \sum_{m=0}^1 \iint w(m) p(x | m, \mu) \log \frac{p_{w, \bar{\pi}}(m | x, \mu) p_{w, \bar{\pi}}(\mu | x)}{p_{w, \bar{\pi}}(m | x) p_{w, \bar{\pi}}(\mu | x)} \pi(d\mu) dx \\ &= \sum_{m=0}^1 \iint w(m) p(x | m, \mu) \log \frac{p_{\bar{\pi}}(\mu | m, x)}{p_{w, \bar{\pi}}(\mu | x)} \pi(d\mu) dx \end{aligned} \quad (12)$$

It is known that an important relation:

$$\inf_{\bar{\pi}} R_w(\pi; \bar{\pi}) = R_w(\pi; \pi) \quad (13)$$

holds; see [5]. Here,  $R_w(\pi; \pi)$  coincides with the conditional mutual information,  $I_{m; \mu | x}(w, \pi)$ , defined by Equation (1) between  $m$  and  $\mu$  given  $x$ .

### 3. Latent Information Priors

We obtain the latent information prior defined as a prior maximizing the conditional mutual information,  $I_{m; \mu | x}(w, \pi)$ . We restrict the original parameter space,  $\mathbb{R}$ , of  $\mu$  to a compact subset,  $K \subset \mathbb{R}$ , for mathematical convenience. A typical choice is a bounded closed interval  $K = [-b, b]$ . If  $b$  is large enough, the testing problem  $H_0 : N(0, \sigma^2)$  versus  $H_1 : N(\mu, \sigma^2)$ ,  $\mu \in [-b, b]$  is close to the original problem.

Let  $\mathcal{P}(K)$  and  $\mathcal{P}(\mathbb{R})$  be the spaces of all probability measures on  $K$  and  $\mathbb{R}$ , respectively, endowed with the weak convergence topology. Then,  $\mathcal{P}(K)$  is compact, since the  $K$  is compact. It is easy to verify that the conditional mutual information,  $I_{m; \mu | x}(w, \pi)$ , is a continuous function of  $w \in [0, 1]$  and  $\pi \in \mathcal{P}(K)$ . Therefore, there exists  $\pi_w^*$  that attains the maximum of Equation (1) for fixed  $w \in (0, 1)$ , since  $\mathcal{P}(K)$  is compact. In the following,  $\pi_w^*$  is denoted as  $\pi^*$  by omitting the subscript,  $w$ , when there is no confusion.

The Bayesian testing based on the latent information prior,  $\pi^* \in \mathcal{P}(K)$ , has the following minimax property.

**Theorem 1.** Let  $\pi^* \in \mathcal{P}(K)$  be the latent information prior. Then:

$$\inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\mu \in K} r_w(\mu, \pi) = \sup_{\mu \in K} r_w(\mu, \pi^*) = I_{m; \mu | x}(w, \pi^*) \quad (14)$$

*Proof.* It is sufficient to show the relations:

$$\begin{aligned} I_{\mu;m|x}(w, \pi^*) &= R_w(\pi^*, \pi^*) = \inf_{\pi \in \mathcal{P}(\mathbb{R})} R_w(\pi^*, \pi) \leq \sup_{\pi' \in \mathcal{P}(K)} \inf_{\pi \in \mathcal{P}(\mathbb{R})} R_w(\pi', \pi) \\ &\leq \inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\pi' \in \mathcal{P}(K)} R_w(\pi', \pi) = \inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\mu \in K} r_w(\mu, \pi) \leq \sup_{\mu \in K} r_w(\mu, \pi^*) \leq R_w(\pi^*, \pi^*) \end{aligned} \quad (15)$$

In the previous section, we have seen the equalities  $I_{\mu;m|x}(w, \pi) = R_w(\pi, \pi)$  and  $R_w(\pi', \pi') = \inf_{\pi} R_w(\pi', \pi)$ , corresponding to the first and second equalities in Equation (15). Thus, it is enough to show the last inequality,  $\sup_{\mu} r_w(\mu, \pi^*) \leq R_w(\pi^*, \pi^*)$ , since the relations, except for the first and second equalities and the last inequality, are obvious.

We prove the inequality by contradiction. Assume that there exists a value,  $\xi \in K$ , such that:

$$r_w(\xi, \pi^*) > R_w(\pi^*, \pi^*) \quad (16)$$

Let  $\pi_t = (1-t)\pi^* + t\delta_\xi$  ( $0 \leq t \leq 1$ ), where  $\delta_\xi$  is the delta measure concentrated at  $\xi$ . Then,  $\pi_t \in \mathcal{P}(K)$ . From Equations (12) and (16):

$$\begin{aligned} \left. \frac{\partial}{\partial t} R_w(\pi_t; \pi_t) \right|_{t=0} &= \frac{\partial}{\partial t} \left\{ w \int p_0(x) \log \frac{\frac{wp_0(x)}{wp_0(x) + (1-w)p_1(x|\mu)}}{\frac{wp_0(x)}{wp_0(x) + (1-w)p_{\pi_t}(x)}} dx \pi_t(d\mu) \right. \\ &\quad \left. + (1-w) \int p_1(x|\mu) \log \frac{\frac{(1-w)p_1(x|\mu)}{wp_0(x) + (1-w)p_1(x|\mu)}}{\frac{(1-w)p_{\pi_t}(x)}{wp_0(x) + (1-w)p_{\pi_t}(x)}} dx \pi_t(d\mu) \right\} \Big|_{t=0} \\ &= w \int p_0(x) \frac{(1-w)\{p_{\delta_\xi}(x) - p_{\pi^*}(x)\}}{wp_0(x) + (1-w)p_{\pi_t}(x)} dx \pi_t(d\mu) \Big|_{t=0} \\ &\quad + w \int p_0(x) \log \frac{\frac{wp_0(x)}{wp_0(x) + (1-w)p_1(x|\mu)}}{\frac{wp_0(x)}{wp_0(x) + (1-w)p_{\pi_t}(x)}} dx \{-\pi^*(d\mu) + \delta_\xi(d\mu)\} \Big|_{t=0} \\ &\quad - (1-w) \int p_1(x|\mu) \frac{(1-w)\{p_{\delta_\xi}(x) - p_{\pi^*}(x)\}}{(1-w)p_{\pi_t}(x)} dx \pi_t(d\mu) \Big|_{t=0} \\ &\quad + (1-w) \int p_1(x|\mu) \frac{(1-w)\{p_{\delta_\xi}(x) - p_{\pi^*}(x)\}}{wp_0(x) + (1-w)p_{\pi_t}(x)} dx \pi_t(d\mu) \Big|_{t=0} \\ &\quad + (1-w) \int p_1(x|\mu) \log \frac{\frac{(1-w)p_1(x|\mu)}{wp_0(x) + (1-w)p_1(x|\mu)}}{\frac{(1-w)p_{\pi_t}(x)}{wp_0(x) + (1-w)p_{\pi_t}(x)}} dx \{-\pi^*(d\mu) + \delta_\xi(d\mu)\} \Big|_{t=0} \\ &= w \int p_0(x) \log \frac{\frac{wp_0(x)}{wp_0(x) + (1-w)p_1(x|\mu)}}{\frac{wp_0(x)}{wp_0(x) + (1-w)p_{\pi^*}(x)}} dx \{-\pi^*(d\mu) + \delta_\xi(d\mu)\} \end{aligned}$$

$$\begin{aligned}
& + (1-w) \int p_1(x | \mu) \log \frac{\frac{(1-w)p_1(x | \mu)}{wp_0(x) + (1-w)p_1(x | \mu)}}{\frac{(1-w)p_{\pi^*}(x)}{wp_0(x) + (1-w)p_{\pi^*}(x)}} dx \{-\pi^*(d\mu) + \delta_\xi(d\mu)\} \\
& = -R_w(\pi^*, \pi^*) + r_w(\xi, \pi^*) > 0
\end{aligned} \tag{17}$$

where we put  $p_1(x | \mu) := p(x | m = 1, \mu)$ . However,  $\max_{t \in [0,1]} R_w(\pi_t; \pi_t) = R_w(\pi_0; \pi_0) = R_w(\pi^*; \pi^*)$ , because of the definition of  $\pi^*$  and the fact that  $\pi_t \in \mathcal{P}(K)$ . This is a contradiction. Thus, we have proven the desired result.  $\square$

The discussion in the proof is parallel to that for submodels of multinomial models in [4], although the testing problem is not included in the class considered there. Closely related discussion on the unconditional mutual information is given in Csiszár [6]. See also, [7,8].

We set  $K = [-b, b]$  with  $b = 7$  and consider two values, 0.5 and 0.355, of  $w$ . The latent information priors,  $\pi_w^*$ , for two values  $w = 0.5$  and  $w = 0.355$  are numerically obtained by using a generalized Arimoto-Blahut algorithm, the details of which will be discussed in another place. Here,  $w = 0.5$  is the setting adopted in many previous studies, and  $w = 0.355$  is the value maximizing  $I_{m;\mu|x}(w, \pi_w^*)$ .

The Arimoto-Blahut algorithm [9,10] is widely used in information theory to obtain the capacity of channels. A channel is defined to be a conditional distribution,  $p(y | \theta)$ , of  $y$  given  $\theta$ , where  $y$  and  $\theta$  are random variables taking values in finite sets,  $\mathcal{Y}$  and  $\Theta$ , respectively. If a channel,  $p(y | \theta)$ , is given, then the mutual information,  $I_{y;\theta}(\pi)$ , between  $y$  and  $\theta$  is a function of the distribution,  $\pi(\theta)$ , of  $\theta$ . The maximum value,  $\max_{\pi} I_{y;\theta}(\pi)$ , of the mutual information as a function of  $\pi$  is called the capacity of the channel  $p(y | \theta)$ . The Arimoto-Blahut algorithm is an iterative algorithm to obtain the capacity  $\max_{\pi} I_{y;\theta}(\pi)$  and the corresponding distribution  $\pi(\theta)$ , attaining the maximum value. The original Arimoto-Blahut algorithm cannot be directly applied to our problem, since we need to maximize the conditional mutual information,  $I_{m;\theta|x}$ , where  $x$  and  $\theta$  are not discrete random variables, to obtain the latent information prior.

**Figure 1.** Latent information priors for (a)  $w = 0.5$  and for (b)  $w = 0.355$ .

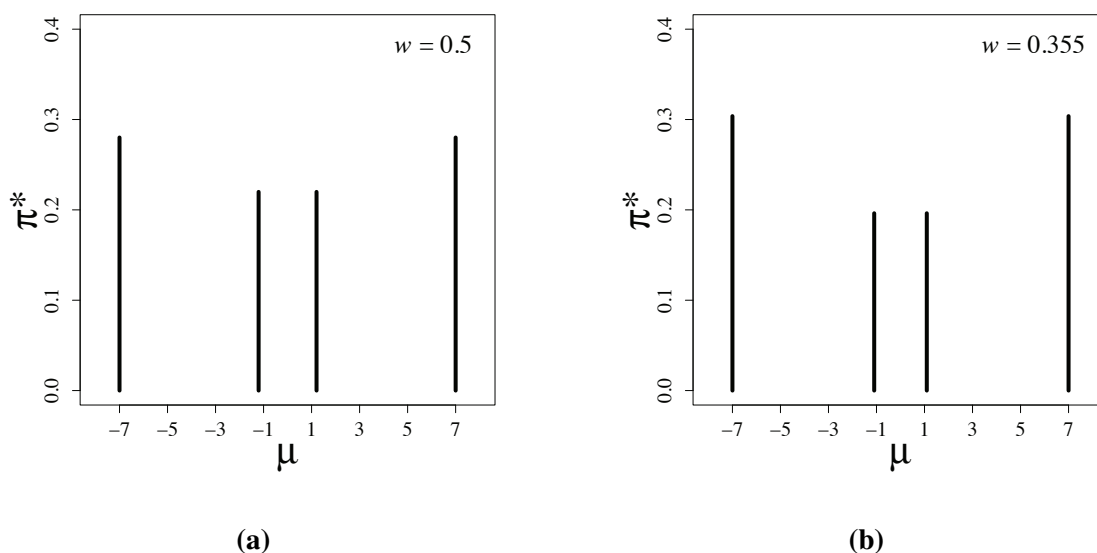


Figure 1 shows the numerically-obtained latent information priors. The priors have the form:

$$\pi_w^* = \frac{u}{2}(\delta_{-a} + \delta_a) + \frac{1-u}{2}(\delta_{-b} + \delta_b) \quad (18)$$

The parameter values are  $a = 1.21$ ,  $b = 7$  and  $u = 0.440$ , when  $w = 0.5$ , and  $a = 1.10$ ,  $b = 7$  and  $u = 0.393$ , when  $w = 0.355$ .

Lemma 2 below gives the risk of Bayesian testing based on the prior in Equation (18).

**Lemma 2.** *Let:*

$$\pi_{a,b,u} = \frac{u}{2}(\delta_{-a} + \delta_a) + \frac{1-u}{2}(\delta_{-b} + \delta_b) \quad (19)$$

where  $a, b > 0$  and  $0 \leq u \leq 1$ . Then, the risk in Equation (8) is given by:

$$\begin{aligned} r_w(\mu; \pi_{a,b,u}) = & -w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}\mu^2 + \mu x \right) \right\} dx \\ & - (1-w) \int \phi(x) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}\mu^2 - \mu x \right) \right\} dx \\ & + w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} (1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx) + \frac{1-w}{w} u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) \right\} dx \\ & + (1-w) \int \phi(x-\mu) \log \left\{ 1 + \frac{w}{1-w} \frac{1}{(1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx) + u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax)} \right\} dx \end{aligned} \quad (20)$$

and the conditional mutual information in Equation (1) is given by:

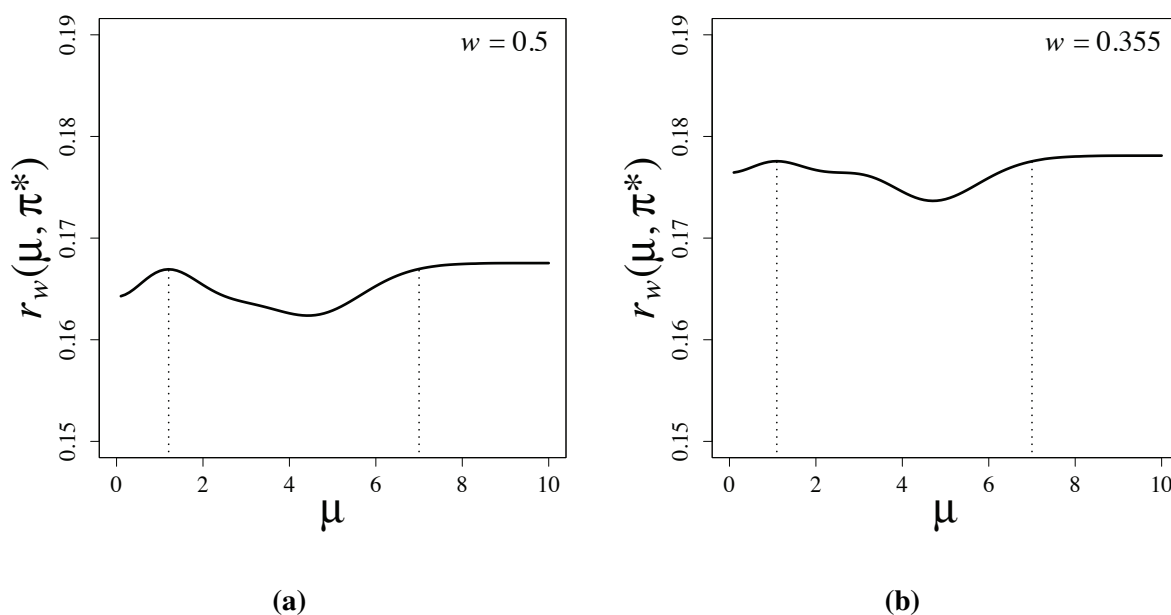
$$\begin{aligned} I_{m;\mu|x}(w, \pi_{a,b,u}) = & u \left[ -w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}a^2 - ax \right) \right\} dx \right. \\ & \left. - (1-w) \int \phi(x) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}a^2 - ax \right) \right\} dx \right] \\ & + (1-u) \left[ -w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}b^2 - bx \right) \right\} dx \right. \\ & \left. - (1-w) \int \phi(x) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}b^2 - bx \right) \right\} dx \right] \\ & + w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) + \frac{1-w}{w} (1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx) \right\} dx \\ & + (1-w)u \int \phi(x-a) \log \left\{ 1 + \frac{w}{1-w} \frac{1}{(1-u) \exp \left( -\frac{b^2}{2} \right) \cosh(bx) + u \exp \left( -\frac{a^2}{2} \right) \cosh(ax)} \right\} dx \\ & + (1-w)(1-u) \int \phi(x-b) \log \left\{ 1 + \frac{w}{1-w} \frac{1}{(1-u) \exp \left( -\frac{b^2}{2} \right) \cosh(bx) + u \exp \left( -\frac{a^2}{2} \right) \cosh(ax)} \right\} dx \end{aligned} \quad (21)$$

The first and second terms in Equation (20) do not depend on  $\pi$ . The third term in Equation (20) does not depend on  $\mu$ .

Figure 2 shows the risk functions of the latent information priors when  $w = 0.5$  and  $w = 0.355$ , respectively. Note that  $\max_{\mu \in [-b,b]} r_w(\mu, \pi^*)$  is attained at  $\mu = a$  and  $b$  in both

examples. This is consistent with the proof of Theorem 1, and it is numerically verified that the prior maximizes the conditional mutual information. Furthermore, we observe that the supremum value,  $\sup_{\mu \in \mathbb{R}} r_w(\mu, \pi^*)$ , of the risk without restriction  $\mu \in [-b, b]$  is only slightly larger than the maximum value,  $\max_{\mu \in [-b, b]} r_w(\mu, \pi^*)$ , with the restriction  $\mu \in [-b, b]$ . The risk functions rapidly converge as  $\mu$  exceeds seven.

**Figure 2.** Risk functions of Bayesian testing based on latent information priors for (a)  $w = 0.5$  and for (b)  $w = 0.355$ . When  $w = 0.5$ ,  $a = 1.21$  and  $b = 7$ . When  $w = 0.355$ ,  $a = 1.10$  and  $b = 7$ . The vertical dotted lines indicate the locations of  $a$  and  $b$ .



Since:

$$\begin{aligned} \sup_{\mu \in K} r(\mu, \pi^*) &= \sup_{\pi' \in \mathcal{P}(K)} \inf_{\pi \in \mathcal{P}(K)} R(\pi', \pi) = \sup_{\pi' \in \mathcal{P}(K)} \inf_{\pi \in \mathcal{P}(\mathbb{R})} R(\pi', \pi) \\ &\leq \sup_{\pi' \in \mathcal{P}(\mathbb{R})} \inf_{\pi \in \mathcal{P}(\mathbb{R})} R(\pi', \pi) \leq \inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\pi' \in \mathcal{P}(\mathbb{R})} R(\pi', \pi) = \inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\mu \in \mathbb{R}} r(\mu, \pi) \leq \sup_{\mu \in \mathbb{R}} r(\mu, \pi^*) \end{aligned} \quad (22)$$

and  $\sup_{\mu \in \mathbb{R}} r(\mu, \pi^*) - \sup_{\mu \in K} r(\mu, \pi^*)$  is small in our problem when  $K = [-b, b]$  ( $b = 7$ ), the supremum value,  $\sup_{\mu \in \mathbb{R}} r(\mu, \pi^*)$ , of the risk function of the latent information prior,  $\pi^*$ , under the parameter restriction,  $\mu \in [-7, 7]$ , is only slightly larger than the minimax value,  $\inf_{\pi \in \mathcal{P}(\mathbb{R})} \sup_{\mu \in \mathbb{R}} r(\mu, \pi)$  without the restriction. We see in the next section that the supremum,  $\sup_{\mu \in \mathbb{R}} r(\mu, \pi)$ , of the risk functions of commonly used priors are much larger than those of  $\pi^*$ .

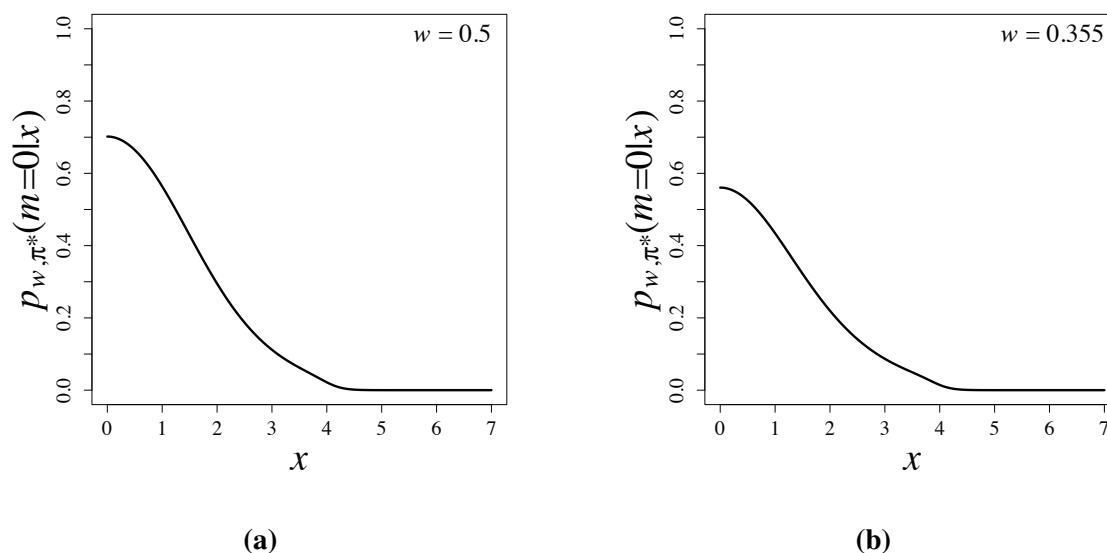
The discreteness of latent information priors shown in Figure 1 is a remarkable feature. In Bayesian statistics,  $k$ -reference priors have been known to be discrete measures in many examples; see [11–13]. The  $k$ -reference prior is defined to be a prior maximizing the mutual information between  $x^k$  and  $\theta$  when we have a set,  $x^k$ , of  $k$ -independent observations,  $x_1, \dots, x_k$ , from  $p(x | \theta)$  in a parametric model,  $\{p(x | \theta) | \theta \in \Theta \subset \mathbb{R}^d\}$ . However, such discrete priors have not been widely used. Instead of  $k$ -reference priors, reference priors introduced by Bernardo [14] have been used for many problems. Reference priors are not discrete and are defined by considering the limit that the sample size  $k$  goes

to infinity. One main reason why discrete priors are not popular is that discrete priors are totally unacceptable from the viewpoint of subjective Bayes in which priors are considered to represent prior belief on parameters.

Although they have not been widely used, discrete priors, such as latent information priors, are reasonable from the viewpoint of prediction and objective Bayes. Various statistical problems, including estimation and testing, can be formulated from the viewpoint of prediction, and priors can be constructed by considering the conditional mutual information. Thus, latent information priors depending on the choice of variables to be predicted could play important roles in many statistical applications. Conditional mutual information is essential in information theory and naturally appeared in several studies in statistics; see e.g., [15,16]. Priors based on conditional mutual information and those based on unconditional mutual information are often quite different; see [4].

Bayesian testing based on latent information priors is free from the Jeffreys-Lindley paradox [3], since the priors are constructed by using conditional mutual information and depend properly on sample sizes. Posterior probabilities,  $p_{w,\pi^*}(m = 0 | x)$ , are shown in Figure 3 and are compared with  $p$ -values of the two-sided test in Table 1. When  $x = 2, 3$  and 4, posterior probabilities are much smaller than  $p$ -values of the two-sided test. Large differences of posterior probabilities and  $p$ -values have been widely observed and discussed in [1,17,18].

**Figure 3.** Posterior probabilities  $p_{w,\pi^*}(m = 0 | x)$  based on latent information priors for (a)  $w = 0.5$  and for (b)  $w = 0.355$ .



**Table 1.** Comparison of posterior probabilities and  $p$ -values.

| $x$                         | 0     | 1     | 2      | 3       | 4                     |
|-----------------------------|-------|-------|--------|---------|-----------------------|
| $p_{w=0.5}(m = 0   x)$      | 0.702 | 0.564 | 0.295  | 0.112   | 0.0217                |
| $p_{w=0.355}(m = 0   x)$    | 0.560 | 0.434 | 0.220  | 0.0867  | 0.0145                |
| $p$ -value (two-sided test) | 1     | 0.317 | 0.0455 | 0.00267 | $6.33 \times 10^{-5}$ |

#### 4. Other Common Priors

Discrete priors, including latent information priors discussed in the previous section, have not been widely used in Bayesian statistics. Common priors for the testing are the normal prior and the Cauchy prior. It seems to have been believed by many statisticians that the Cauchy prior is slightly better than the normal prior; see, e.g., [1,2]. In this section, we evaluate the conditional mutual information for the priors and compare the performance of them to that of the latent information prior.

##### 4.1. The Normal Prior

The normal prior,  $\phi(\mu; 0, \tau^2)$ , is denoted by  $N_\tau$ . From Lemma 1, we have:

$$\begin{aligned} r_w(\mu, N_\tau) = & -w \int \phi(x; 0, 1) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}\mu^2 + \mu x \right) \right\} dx \\ & - (1-w) \int \phi(x; 0, 1) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}\mu^2 - \mu x \right) \right\} dx \\ & + w \int \phi(x; 0, 1) \log \left\{ 1 + \frac{1-w}{w} \frac{\phi(x; 0, \tau^2 + 1)}{\phi(x; 0, 1)} \right\} dx \\ & + (1-w) \int \phi(x; \mu, 1) \log \left\{ 1 + \frac{w}{1-w} \frac{\phi(x; 0, 1)}{\phi(x; 0, \tau^2 + 1)} \right\} dx \end{aligned} \quad (23)$$

Thus, the conditional mutual information is given by:

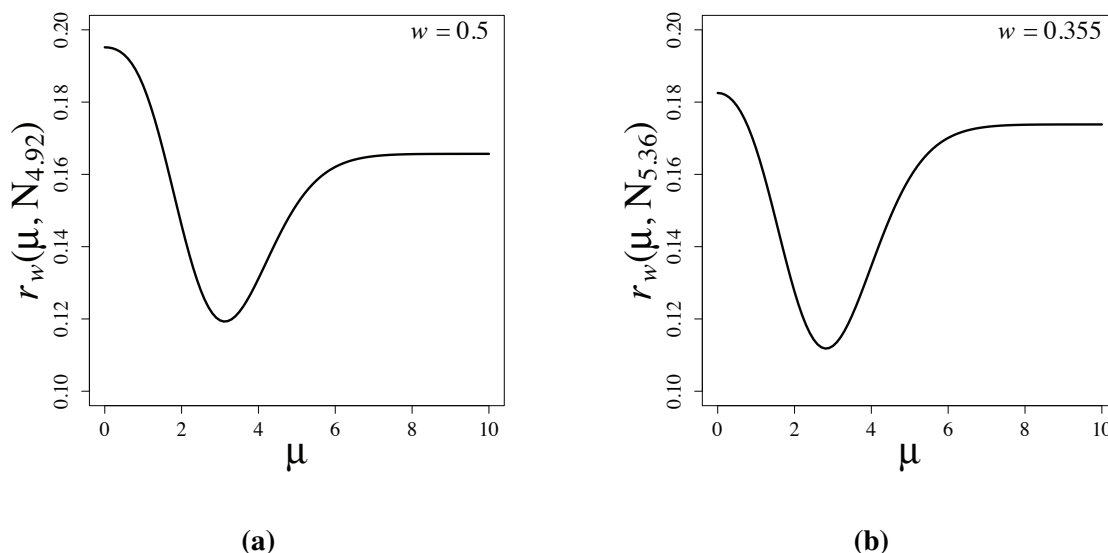
$$\begin{aligned} I_{m;\mu|x}(w, N_\tau) = & \int r_w(\mu, N_\tau) \phi(\mu; 0, \tau^2) d\mu \\ = & -w \int \phi(x; 0, 1) \phi(\mu; 0, \tau^2) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}\mu^2 + \mu x \right) \right\} d\mu dx \\ & - (1-w) \int \phi(x; 0, 1) \phi(\mu; 0, \tau^2) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}\mu^2 - \mu x \right) \right\} d\mu dx \\ & + w \int \phi(x; 0, 1) \log \left\{ 1 + \frac{1-w}{w} \frac{\phi(x; 0, \tau^2 + 1)}{\phi(x; 0, 1)} \right\} dx \\ & + (1-w) \int \phi(x; 0, \tau^2 + 1) \log \left\{ 1 + \frac{w}{1-w} \frac{\phi(x; 0, 1)}{\phi(x; 0, \tau^2 + 1)} \right\} dx \end{aligned} \quad (24)$$

The conditional mutual information is evaluated by numerical integration. When  $w = 0.5$  and  $w = 0.355$ , the maximum values:

$$\max_{\tau} I_{m;\mu|x}(w = 0.5, N_\tau) = 0.156 \quad \text{and} \quad \max_{\tau} I_{m;\mu|x}(w = 0.355, N_\tau) = 0.166 \quad (25)$$

of Equation (24) are attained at  $\tau = 4.92$  and  $\tau = 5.36$ , respectively. The variation of the risk functions,  $r_{w=0.5}(\mu, N_{\tau=4.92})$  and  $r_{w=0.355}(\mu, N_{\tau=5.36})$ , shown in Figure 4 are much larger than those of the risk functions of the latent information priors shown in Figure 2. Thus, the performance of the Bayesian testing based on the normal prior is worse than that based on the latent information prior if we adopt the Kullback-Leibler loss.

**Figure 4.** Risk functions of Bayesian testing based on normal priors for (a)  $w = 0.5$  and  $\tau = 4.92$ ; and for (b)  $w = 0.355$  and  $\tau = 5.36$ . The functions have symmetry  $r_w(-\mu, N_\tau) = r_w(\mu, N_\tau)$  about the origin.



#### 4.2. The Cauchy Prior

The Cauchy prior,  $1/\{\gamma\pi(\mu^2/\gamma^2 - 1)\}$ , is denoted by  $C_\gamma$ . Since the characteristic functions of  $N(0, \sigma^2)$  and  $C_\gamma$  are  $\exp(-\frac{1}{2}\sigma^2 t^2)$  and  $\exp(-\gamma|t|)$ , respectively, the characteristic function of the marginal density:

$$p_C(x | \gamma) = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \frac{1}{\pi(\mu^2/\gamma^2 - 1)} \frac{1}{\gamma} d\mu \quad (26)$$

with respect to the Cauchy prior,  $C_\gamma$ , is given by:

$$\exp\left(-\gamma|t| - \frac{1}{2}t^2\right) \quad (27)$$

The expression:

$$\begin{aligned} p_C(x | \gamma) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} \exp\left(-\gamma|t| - \frac{1}{2}t^2\right) dt \\ &= \frac{1}{\sqrt{2\pi}\sigma} \operatorname{Re} \left[ \exp\left\{\frac{(ix - \gamma)^2}{2}\right\} \operatorname{erfc}\left(-i\frac{x + i\gamma}{\sqrt{2}}\right) \right] \end{aligned} \quad (28)$$

where  $\operatorname{erfc}$  is the complementary error function defined by:

$$\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} e^{-t^2} dt \quad (29)$$

obtained by the inverse transform of Equation (27) is useful for numerical computation; see [19] (p. 183) and [20]. From Lemma 1, we have:

$$r_w(\mu; C_\gamma) = -w \int \phi(x; 0, 1) \log \left\{ 1 + \frac{1-w}{w} \exp\left(-\frac{1}{2}\mu^2 + \mu x\right) \right\} dx$$

$$\begin{aligned}
& - (1 - w) \int \phi(x; 0, 1) \log \left\{ 1 + \frac{w}{1 - w} \exp \left( -\frac{1}{2} \mu^2 - \mu x \right) \right\} dx \\
& + w \int \phi(x; 0, 1) \log \left\{ 1 + \frac{1 - w}{w} \frac{p_C(x | \gamma)}{\phi(x; 0, 1)} \right\} d\varepsilon \\
& + (1 - w) \int \phi(x; 0, 1) \log \left\{ 1 + \frac{w}{1 - w} \frac{\phi(x + \mu; 0, 1)}{p_C(x + \mu | \gamma)} \right\} dx
\end{aligned} \quad (30)$$

We numerically evaluate the conditional mutual information:

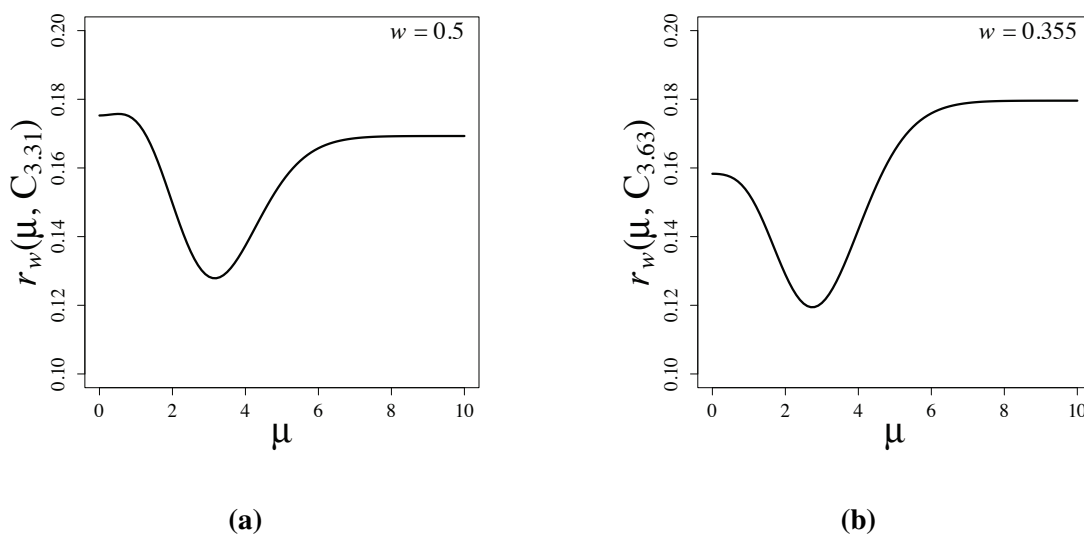
$$I_{m;\mu|x}(w, C_\gamma) = \int r_w(\mu; C_\gamma) \frac{1}{\pi(\mu^2/\gamma^2 - 1)} \frac{1}{\gamma} d\mu \quad (31)$$

by the Monte-Carlo method. When  $w = 0.5$  and  $w = 0.355$ , the maximum values:

$$\max_{\gamma} I_{m;\mu|x}(w = 0.5, C_\gamma) = 0.161 \quad \text{and} \quad \max_{\gamma} I_{m;\mu|x}(w = 0.355, C_\gamma) = 0.170$$

of Equation (31) are attained at  $\gamma = 3.31$  and  $\gamma = 3.63$ , respectively. The risk functions  $r_{w=0.5}(\mu, C_{\gamma=3.31})$  and  $r_{w=0.355}(\mu, C_{\gamma=3.63})$  are shown in Figure 5. The variation of the risk function  $r_{w=0.5}(\mu, C_{\gamma=3.31})$  is milder than that of the risk function  $r_{w=0.5}(\mu, N_{\tau=4.92})$  based on the normal prior, and the inequality  $\sup_{\mu} r_{w=0.5}(\mu, C_{\gamma=3.31}) < \sup_{\mu} r_{w=0.5}(\mu, N_{\tau=4.92})$  holds. Thus, the Cauchy prior is preferable to the normal prior from the viewpoint of the Kullback-Leibler loss. However, the variation of the risk function shown in Figure 2 based on the latent information prior is much smaller than that of  $r_{w=0.5}(\mu, C_{\gamma=3.31})$ . Similar relations also hold when  $w = 0.355$ .

**Figure 5.** Risk functions of Bayesian testing based on Cauchy priors for (a)  $w = 0.5$  and  $\gamma = 3.31$ ; and for (b)  $w = 0.355$  and  $\gamma = 3.63$ . The functions have symmetry  $r_w(-\mu, C_\gamma) = r_w(\mu, C_\gamma)$  about the origin.



## 5. Conclusions

We discussed the use of latent information priors for Bayesian testing of a point null hypothesis. The testing problem was formulated as a prediction problem, and latent information priors were numerically

obtained. The variations of the risk functions of latent information priors are much smaller than those of normal and Cauchy priors. Although the testing problem treated in the present paper is simple, the results may indicate that latent information priors could be useful for various problems, since many statistical problems can be formulated from the viewpoint of prediction.

When the parameter space is multidimensional, it becomes difficult to numerically obtain latent information priors, and some approximations need to be used. One possible approach is to use asymptotic methods, and another possible approach is to choose an approximating prior from a tractable subset of the set of all probability measures on the parameter space. These approaches require further investigation.

## Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research (23300104, 23650144) and by the Aihara Innovative Mathematical Modelling Project, the Japan Society for the Promotion of Science (JSPS) through the “Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program),” initiated by the Council for Science and Technology Policy (CSTP).

## Conflicts of Interest

The author declares no conflict of interest.

## References

1. Berger, J.O.; Sellke, T. Testing a point null hypothesis: The irreconcilability of  $p$  values and evidence. *J. Am. Stat. Assoc.* **1987**, *82*, 112–122.
2. Jeffreys, H. *Theory of Probability*, 3rd ed.; Oxford University Press: Oxford, UK, 1961.
3. Lindley, D.V. A statistical paradox. *Biometrika* **1957**, *44*, 187–192.
4. Komaki, F. Bayesian predictive densities based on latent information priors. *J. Stat. Plan. Inference* **2011**, *141*, 3705–3715.
5. Aitchison, J. Goodness of prediction fit. *Biometrika* **1975**, *62*, 547–554.
6. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.
7. Haussler, D. A general minimax result for relative entropy. *IEEE Trans. Inf. Theory* **1997**, *43*, 1276–1280.
8. Grünwald, P.D.; Dawid, A.P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Stat.* **2004**, *32*, 1367–1433.
9. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20.
10. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473.
11. Hartigan, J.A. *Bayes Theory*; Springer: New York, NY, USA, 1983.

12. Berger, J.; Bernardo, J.M.; Mendoza, M. On Priors that Maximize Expected Information. In *Recent Developments in Statistics and Their Applications*; Klein, J.P., Lee, J.C., Eds.; Freedom Press: Seoul, Korea, 1989; pp. 1–20.
13. Zhang, Z. Discrete Noninformative Priors. Ph.D. Dissertation, Department of Statistics, Yale University, New Haven, CT, USA, 1994.
14. Bernardo, J.M. Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. B* **1979**, *41*, 113–147.
15. Clarke, B.; Yuan, A. Partial information reference priors: Derivation and interpretations. *J. Stat. Plan. Inference* **2004**, *123*, 313–345.
16. Ebrahimi, N.; Soofi, E.S.; Soyer, R. On the sample information about parameter and prediction. *Stat. Sci.* **2010**, *25*, 348–367.
17. Edwards, W.; Lindman, H.; Savage, L.J. Bayesian statistical inference for psychological research. *Psychol. Rev.* **1963**, *70*, 193–242.
18. Dickey, J.M. Is the tail area useful as an approximate Bayes factor? *J. Am. Stat. Assoc.* **1977**, *72*, 138–142.
19. Temme, N.M. Error Functions, Dawson's and Fresnel Integrals. In *NIST Handbook of Mathematical Functions*; Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 159–171.
20. Poppe, G.P.M.; Wijers, C.M.J. Algorithm 680: Evaluation of the complex error function. *ACM Trans. Math. Softw. (TOMS)* **1990**, *16*, doi:10.1145/77626.77630.

## Appendix. Proofs of Lemmas

*Proof of Lemma 1.* From Equation (8), we have:

$$r_w(\mu; \pi) = w \int p(x | m = 0) \log \frac{p_w(m = 0 | \mu, x)}{p_{w,\pi}(m = 0 | x)} dx + (1 - w) \int p(x | m = 1, \mu) \log \frac{p_w(m = 1 | \mu, x)}{p_{w,\pi}(m = 1 | x)} dx \quad (32)$$

because  $m$  and  $\mu$  are independent. Since:

$$\frac{p_w(m = 0 | \mu, x)}{p_{w,\pi}(m = 0 | x)} = \frac{\frac{wp(x | m = 0)}{wp(x | m = 0) + (1 - w)p(x | m = 1, \mu)}}{\frac{wp(x | m = 0)}{wp(x | m = 0) + (1 - w)p_\pi(x | m = 1)}} = \frac{1 + \frac{1 - w}{w} \frac{p_\pi(x | m = 1)}{p(x | m = 0)}}{1 + \frac{1 - w}{w} \frac{p(x | m = 1, \mu)}{p(x | m = 0)}} \quad (33)$$

and:

$$\frac{p_w(m = 1 | \mu, x)}{p_{w,\pi}(m = 1 | x)} = \frac{\frac{(1 - w)p(x | m = 1, \mu)}{wp(x | m = 0) + (1 - w)p(x | m = 1, \mu)}}{\frac{(1 - w)p_\pi(x | m = 1)}{wp(x | m = 0) + (1 - w)p_\pi(x | m = 1)}} = \frac{\frac{w}{1 - w} \frac{p(x | m = 0)}{p_\pi(x | m = 1)} + 1}{\frac{w}{1 - w} \frac{p(x | m = 0)}{p(x | m = 1, \mu)} + 1} \quad (34)$$

we have:

$$\begin{aligned}
 r_w(\mu; \pi) &= w \int p(x | m = 0) \frac{1 + \frac{1-w}{w} \frac{p_\pi(x | m = 1)}{p(x | m = 0)}}{1 + \frac{1-w}{w} \frac{p(x | m = 1, \mu)}{p(x | m = 0)}} dx \\
 &\quad + (1-w) \int p(x | m = 1, \mu) \log \frac{1 + \frac{w}{1-w} \frac{p(x | m = 0)}{p_\pi(x | m = 1)}}{1 + \frac{w}{1-w} \frac{p(x | m = 0)}{p(x | m = 1, \mu)}} dx \\
 &= w \int p_0(x) \log \frac{1 + \frac{1-w}{w} \frac{p_\pi(x)}{p_0(x)}}{1 + \frac{1-w}{w} \frac{p_0(x-\mu)}{p_0(x)}} dx + (1-w) \int p_0(x) \log \frac{1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_\pi(x+\mu)}}{1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_0(x)}} dx \quad (35)
 \end{aligned}$$

□

*Proof of Lemma 2.* Since:

$$\begin{aligned}
 \phi(x-a) + \phi(x+a) &= \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x^2 - 2ax + a^2) \right\} + \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x^2 + 2ax + a^2) \right\} \\
 &= 2\phi(x) \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) \quad (36)
 \end{aligned}$$

we have:

$$\begin{aligned}
 \frac{p_\pi(x)}{p_0(x)} &= \frac{\frac{1}{2}u \{ \phi(x+a) + \phi(x-a) \} + \frac{1}{2}(1-u) \{ \phi(x+b) + \phi(x-b) \}}{\phi(x)} \\
 &= u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) + (1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx) \quad (37)
 \end{aligned}$$

From Lemma 1, we have:

$$\begin{aligned}
 r_w(\mu; \pi) &= -w \int p_0(x) \log \left\{ 1 + \frac{1-w}{w} \frac{p_0(x-\mu)}{p_0(x)} \right\} dx - (1-w) \int p_0(x) \log \left\{ 1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_0(x)} \right\} dx \\
 &\quad + w \int p_0(x) \log \left\{ 1 + \frac{1-w}{w} \frac{p_\pi(x)}{p_0(x)} \right\} dx + (1-w) \int p_0(x) \log \left\{ 1 + \frac{w}{1-w} \frac{p_0(x+\mu)}{p_\pi(x+\mu)} \right\} dx \\
 &= -w \int \phi(x) \log \left\{ 1 + \frac{1-w}{w} \exp \left( -\frac{1}{2}\mu^2 + \mu x \right) \right\} dx \\
 &\quad - (1-w) \int \phi(x) \log \left\{ 1 + \frac{w}{1-w} \exp \left( -\frac{1}{2}\mu^2 - \mu x \right) \right\} dx \\
 &\quad + w \int \phi(x) \\
 &\quad \quad \times \log \left\{ 1 + \frac{1-w}{w} u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) + \frac{1-w}{w} (1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx) \right\} dx \\
 &\quad + (1-w) \int \phi(x-\mu)
 \end{aligned}$$

$$\times \log \left\{ 1 + \frac{w}{1-w} \frac{1}{u \exp \left( -\frac{1}{2}a^2 \right) \cosh(ax) + (1-u) \exp \left( -\frac{1}{2}b^2 \right) \cosh(bx)} \right\} dx \quad (38)$$

The conditional mutual information is:

$$\begin{aligned} I_{m;\mu|x}(w, \pi) &= \frac{u}{2} \{r_w(-a; \pi) + r_w(a; \pi)\} + \frac{1-u}{2} \{r_w(-b; \pi) + r_w(b; \pi)\} \\ &= ur_w(a; \pi) + (1-u)r_w(b; \pi) \end{aligned} \quad (39)$$

From Equations (38) and (39), we obtain the desired result.  $\square$

© 2013 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).