NONINFORMATIVE PRIOR FOR MODEL SELECTION

by

Katsuaki Iwaki
Purdue University

Technical Report #96-34C

Department of Statistics
Purdue University
West Lafayette, IN USA

July 1996

◇

Noninformative Prior for Model Selection

Katsuaki Iwaki

Abstract

In this article, we introduce the noninformative prior for the model selection. The noninformative prior developed for each model separately is determined only up to a multiplicative constant, so it can not be used for the model selection. However, starting from this prior, we obtain the completely determined noninformative priors for the models simultaneously. This prior is based on the standard experiment, which is a subjective element. However, we can apply the widely accepted notion of the minimal sample principle. Three related topics with this noninformative prior will also be discussed.

1 INTRODUCTION

In this article, we discuss the model selection problem from Bayesian point of view, using the noninformative prior.

Let $\mathcal{M} = \{M_1, \cdots, M_m\}$ be the set of statistical models about the observable random variable $\tilde{x}$. In this article, all random variables are shown to be with tilde. Each model M_j accompanies its own parameter space Θ_j which is a singleton or an open subspace of $\Re^{d_j}$. The model M_j is called simple when Θ_j is a singleton and it is called composite otherwise. Under $M_j \in \mathcal{M}$ and $\theta_j \in \Theta_j$, $\tilde{x}$ has a probability density $p_j(x|\theta_j)$ w.r.t. a common underlying measure $m(\cdot)$ on $rng(\tilde{x}_n)$ (range of $\tilde{x}$).

Let $p_j(\theta_j)$ be the usual noninformative prior on Θ_j w.r.t. Lebesgue measure on $\Re^{d_j}$ when only M_j is considered. This prior is determined only up to a multiplicative constant. Starting from $p_j(\theta_j)$, we derive the noninformative prior for the model selection $\pi_j(\theta_j)$, which

is completely determined. It is not necessarily $\pi_j(\theta_j) \propto p_j(\theta_j)$. In this article, $p_j(\theta_j)$ is called the initial (noninformative) prior and $\pi_j(\theta_j)$ is called the noninformative prior for the model selection.

It is impossible to determine the multiplicative constants directly. Thus some intermediate concepts have been proposed to obtain the posterior probability. See Spiegelhalter and Smith (1982), Suzuki (1983) and Klein and Brown (1984), although these ideas are not satisfactory. See Iwaki (1996) for these methods and other modified Bayesian approaches.

In this article, we introduce the standard experiment as an intermediate concept to determine the prior completely. The choice of the standard experiment is basically subjective but the experimenters are supposed to have some intuition about it. Also the widely accepted notion of the minimal sample principle can be applied.

In Section 2.1, the motivation and the basic notion are introduced. In Section 2.2, examples are given. In Section 2.3, it is discussed how to choose the standard experiment. In Section 3.1, the problem of robustness to the choice of the standard experiment is discussed using the linear regression model as an example. In Section 3.2, it is discussed how to incorporate prior partial knowledge into the prior distribution. In Section 3.3, Lindley's paradox is discussed.

2 THE CONCEPT OF NONINFORMATIVE PRIOR FOR MODEL SELECTION

2.1 Definition and its Motivation

Let $\tilde{x}^I$ be a random variable whose distribution is determined by $M_j \in \mathcal{M}$ and $\theta_j \in \Theta_j$ and is conditionally independent of $\tilde{x}$ under M_j and θ_j for $j = 1, \cdots, m$. This random variable $\tilde{x}^I$ is called the standard experiment. The choice of the standard experiment is subjective but we discuss how to work intuition for this choice later in Section 2.3.

For simplicity, we assume that all models are composite for the time being. If the realization of the standard experiment $\tilde{x}^I = x^I$ were obtained before the real data $\tilde{x}$ is observed, it

would be used to obtain the prior density under M_j and x^I by

$$p(\theta_j|M_j, x^I) = \frac{p(\theta_j)p_j(x^I|\theta_j)}{\int_{\Theta_j} p(x^I|\theta_j)p_j(\theta_j)d\theta_j}, \tag{2.1}$$

provided that

$$\int_{\Theta_j} p(x^I|\theta_j)p_j(\theta_j)d\theta_j < \infty. \tag{2.2}$$

Then, using this density as the prior together with $p(M_j)$, the probabilities of the models are obtained by

$$p(M_j|x; x^I) \propto p(M_j) \int_{\Theta_j} p_j(x|\theta_j)p(\theta_j|M_j, x^I)d\theta_j. \tag{2.3}$$

Usually, $p(M_j) = 1/m$.

Since the standard experiment $\tilde{x}^I$ is not observable, this probability can not be obtained. However, we introduce the noninformative prior the result of which agrees with this imaginary solution in the following sense.

Definition Let

$$\bar{p}_j(\theta_j|\theta_j^*) = E[p_j(\theta_j|\tilde{x}^I)|M_j, \theta_j^*]. \tag{2.4}$$

We call

$$\pi_j(\theta_j) = \bar{p}_j(\theta_j|\theta_j), \tag{2.5}$$

the noninformative prior for the model selection.

Note that $\bar{p}_j(\theta_j|\theta_j^*)$ is a proper density.

Example Suppose that $\tilde{x}^I|M_j, \theta_j \sim Exp(\theta_j^{-1})$ and $p_j(\theta_j) \propto \theta_j^{-1}$. Then

$$\begin{aligned}
\bar{p}_j(\theta_j|\theta_j^*) &= E\left[\tilde{x}^I \exp\left(-\theta\tilde{x}^I\right)|M_j, \theta_j^*\right] \\
&= \int_0^\infty x^I \exp\left(-\theta_j x^I\right) \theta_j^* \exp\left(-\theta_j^* x^I\right) dx^I \\
&= \frac{\theta_j^*}{(\theta_j^* + \theta_j)^2}.
\end{aligned}$$

Thus

$$\pi_j(\theta_j) = \frac{1}{4\theta_j}. \tag{2.6}$$

This prior has the following asymptotic property. We denote the observable random variable with the sample size n by $\tilde{x}_{(n)} = (\tilde{x}_1, \cdots, \tilde{x}_n)$. The probability distribution of $(\tilde{x}_1, \cdots, \tilde{x}_n, \cdots)$ given $M_j \in \mathcal{M}$ and $\theta_j \in \Theta_j$ is denoted by $P_j(\cdot|\theta_j)$. In the following theorem, the suffix j is omitted for simplicity.

Theorem Assume that for any $\theta^* \in \Theta$ and $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{\int_{U(\theta^*,\varepsilon)^c} p(\tilde{x}_n|\theta)\pi(\theta)d\theta}{\int_{U(\theta^*,\varepsilon)} p(\tilde{x}_n|\theta)\pi(\theta)d\theta} = 0, \qquad (P(\cdot|\theta^*) - a.e.). \tag{2.7}$$

where $U(\theta^*, \varepsilon) = \{\theta \in \Theta : \|\theta - \theta^*\| < \varepsilon\}$. Also $\bar{p}(\theta|\theta^*)/\pi(\theta)$ is assumed to be bounded and continuous. Then

$$\begin{aligned} &\int_\Theta p(\tilde{x}_{(n)}|\theta)\bar{p}(\theta|\theta^*)d\theta \\ &\quad = \int_\Theta p(\tilde{x}_{(n)}|\theta)\pi(\theta)d\theta(1 + \tilde{o}(1)), \qquad (P_j(\cdot|\theta^*) - a.e.), \end{aligned} \tag{2.8}$$

for any $\theta_j^* \in \Theta_j$.

The proof of the theorem is in the Appendix. See Berk(1966,1970), Sono(1986) and Dmochowski (1995) for the assumption (2.7) , which implies that the posterior probability mass on Θ accumulates in the neighborhood of the true value. This result shows that the method proposed by Iwaki (1996) agrees with the method here asymptotically. See Proposition 1 in Iwaki (1996).

Once $\pi_j(\theta_j)$ is obtained, the posterior probability is obtained by

$$p(M_j|x) \propto \int_{\Theta_j} p_j(x|\theta_j)\pi_j(\theta_j)d\theta_j, \tag{2.9}$$

for the composite model and

$$p(M_j|x) \propto p_j(x|\theta_0), \tag{2.10}$$

for the simple model M_j with $\Theta_j = \{\theta_0\}$.

If the imaginary prior $p_j(\theta_j|\tilde{x}^I)$ is reasonable, then its expectation under the true value, $\bar{p}_j(\theta_j|\theta_j^*)$, is reasonable. The noninformative prior $\pi_j(\theta_j)$ is desirable, because its marginal likelihood asymptotically agrees with the marginal likelihood based on $\bar{p}_j(\theta_j|\theta_j^*)$ under the suitable conditions.

2.2 Examples

Example 1 This is an example from Akaike (1991). Let $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ be an i.i.d. sample from $M_j : N(\theta, \sigma_j^2)$, $j = 1, 2$, where $\theta \in \Re$ is unknown while σ_1 and σ_2 are known. If the standard experiment is $\tilde{x}^I|M_j, \theta \sim N(\theta, \sigma_0)$ and the initial prior is $p_j(\theta) \propto 1$, then the noninformative prior is

$$\pi_j(\theta) = \frac{1}{2\sqrt{\pi}\sigma_0}, \qquad j = 1, 2, \tag{2.11}$$

and the posterior odds ratio in favor of M_1 is

$$\left(\frac{\sigma_2}{\sigma_1}\right)^{n-1} \exp\left(-\frac{1}{2}(\sigma_1^{-2} - \sigma_2^{-2})\sum_{i=1}^{n}(x_i - \bar{x})^2\right). \tag{2.12}$$

where $\bar{x}$ is the sample mean. This result does not depend on the choice of σ_0 and the posterior odds ratio is always 1 when $n = 1$. A Comment is found in Section 2.3.

Example 2: Normal distribution Assume that $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ is an i.i.d. sample from $N(\mu, \tau^{-1})$. The models are $M_1 : (\mu, \tau) \in \{(0, 1)\}$, $M_2 : (\mu, \tau) \in \Re \times \{1\}$, $M_3 : (\mu, \tau) \in \{0\} \times (0, \infty)$ and $M_4 : (\mu, \tau) \in \Re \times (0, \infty)$. The initial noninformative prior is supposed to be $p_2(\mu) \propto 1$, $p_3(\tau) \propto \tau^{-1}$ and $p_4(\mu, \tau) \propto \tau^{\alpha-1}$. The standard experiment $\tilde{x}^I = (\tilde{x}_1^I, \tilde{x}_2^I)$ is supposed to be a two dimensional random vector with distribution $\tilde{x}^I|\mu, \tau \sim N_2(\mu\iota_2, \tau^{-1}I_2)$ where $\iota_n = (1, \cdots, 1)' \in \Re^n$ and I_n is the identity matrix of order n.

Then the noninformative priors for the model selection are

$$\pi_2(\mu) = \frac{1}{\sqrt{2\pi}}, \tag{2.13}$$

$$\pi_3(\tau) = \frac{1}{4\tau}, \tag{2.14}$$

and

$$\pi_4(\mu, \tau) = \frac{\Gamma(\alpha+1)}{2^{\alpha+3/2}\pi\Gamma(\alpha+1/2)}\frac{1}{\sqrt{\tau}}. \tag{2.15}$$

The posterior probabilities of the models are

$$p(M_1|x) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right), \tag{2.16}$$

$$p(M_2|x) \propto \frac{1}{\sqrt{n}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right), \tag{2.17}$$

$$p(M_3|x) \propto 2^{(n-4)/2}\Gamma(n/2)\left(\sum_{i_1}^{n} x_i^2\right)^{-n/2}, \tag{2.18}$$

and

$$p(M_4|x) \propto \frac{\Gamma(\alpha+1)\Gamma(n/2)}{\Gamma(\alpha+1/2)\sqrt{\pi}2^{\alpha+2}\sqrt{n}}2^{n/2-\alpha-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{-n/2}. \tag{2.19}$$

What is interesting is that $\pi_4(\mu, \tau) \propto \tau^{-1/2}$ for all α.

Example 3: Exponential, Lognormal and Weibull We assume that $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ is an i.i.d. sample. It is also assumed that the distribution of $\tilde{x}_i$ is exponential with mean θ^{-1} under M_1, the distribution of $\log \tilde{x}_i$ is normal with mean μ and variance τ^{-1} under M_2 and $\tilde{x}^\alpha$ is exponential with mean β^{-1} under M_3. The initial priors are assumed to be $p_1(\theta) \propto \theta^{-1}$, $p_2(\mu, \tau) \propto \tau^{-1}$ and $p_3(\alpha, \beta) \propto \alpha^{-1}\beta^{-1}$. It is supposed that the standard experiment is the two dimensional vector $\tilde{x}^I = (\tilde{x}_1^I, \tilde{x}_2^I)$ and its distribution is the same as $(\tilde{x}_1, \tilde{x}_2)$ under each model and each parameter.

Then the noninformative priors for model selection are

$$\pi_1(\theta) = \frac{3}{8\theta}, \tag{2.20}$$

$$\pi_2(\mu,\tau) = \frac{1}{2^{3/2}\pi^{3/2}}\frac{1}{\sqrt{\tau}}. \tag{2.21}$$

and

$$\pi_3(\alpha,\beta) = K\frac{1}{\beta}, \tag{2.22}$$

where

$$\begin{aligned} K &= \frac{1}{8}\int_{(0,\infty)^2} z_1 z_2 \exp\left(-(z_1+z_2)\right)|\log z_1 - \log z_2|\, dz_1 dz_2 \\ &\approx 0.111. \end{aligned}$$

The probabilities of the models are

$$p(M_1|x) \propto \frac{3\Gamma(n)}{8t^n}, \tag{2.23}$$

$$p(M_2|x) \propto \frac{\Gamma(n/2)}{(\prod_{i=1}^n x_i)2\pi^{n/2+1}\sqrt{n}}(\sum_{i=1}^n \log x_i - \hat{\mu})^2)^{n/2}, \tag{2.24}$$

and

$$p(M_3|x) \propto K\Gamma(n)\int_0^\infty \left(\prod_{i=1}^n x_i\right)^{\alpha-1}\left(\sum_{i=1}^n x_i^\alpha\right)^{-\alpha}\alpha^{n-1}d\alpha. \tag{2.25}$$

where $\hat{\mu}$ is the sample mean of $\log x_i$s.

Proschan (1963) presents the following 30 time intervals (in hours) between failures of the air conditioning system of an airplane: 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95. For this data, we obtain the posterior probabilities: $p(M_1|x) = 0.663$, $p(M_2|x) = 0.200$ and $p(M_3|x) = 0.138$.

2.3 Choice of the Standard Experiment

The noninformative prior for the model selection $\pi_j(\theta_j)$ is defined by relating it to the imaginary probability $p(\theta_j|M_j, x^I)$ through (2.4) and (2.8) . Thus it may be justified only if the imaginary analysis is rational.

In the imaginary analysis, it should be noted that the information $\tilde{x}^I = x^I$ is neglected for the model probability $p(M_j)$, while the conditional distribution, $p_j(\theta_j|x^I)$, is affected by $\tilde{x}^I = x^I$. In order to form the prior conditional distribution, it is desirable that information

from the standard experiment is large. However, it is ridiculous that the model probabilities are not affected by the large information, while the extreme results tend to be avoided under the small information. Experimenters are supposed to have their own personal judgement about whether a statistical experiment is trivial or not. Thus the standard experiment should contain as much information as possible, subject to the information being small enough that it would not affect the assessment of the model prior probabilities.

Widely accepted criterion is the "minimal sample principle", though it is differently applied to different methods. See Berger and Pericchi (1995, 1996), Iwaki(1996), Jeffreys (1961), Kass and Wasserman (1995), Klein and Brown (1984), O'Hagan(1995) and Spiegelhalter and Smith (1982).

Definition Let $p_j(\theta_j)$ be an initial noninformative prior. The sample x is called to be proper if

$$\int_{\Theta_j} p_j(x|\theta_j)p_j(\theta_j)d\theta_j < \infty, \tag{2.26}$$

for any $M_j \in \mathcal{M}$ and minimal if it is proper and no subset is proper.

For our method, the minimal samples are assumed to be the realizations of the same distribution. Then this principle may be applied so that the standard experiment have the same distribution as the minimal samples under each model and each parameter. Even if the assumption does not hold, the principle may provide guidelines for the choice of the standard experiment. This approach has been illustrated in the example 2 and the example 3 in Section 2.2. Note that the same model may have the different standard experiments depending on the whole structure in which it is contained. If M_1 in Example 3 in Section 2.2 is compared with, e.g., simple models, then $\pi_1(\theta_1)$ is given by (2.6) instead of (2.20) .

This idea should not be applied mechanically. In Example 1 in Section 2.2, if it is applied to the problem, the posterior odds ratio is

$$\left(\frac{\sigma_2}{\sigma_1}\right)^n \exp\left(-\frac{1}{2}\left(\sigma_1^{-2} - \sigma_2^{-2}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2\right). \tag{2.27}$$

Thus if σ_2/σ_1 is very large, the posterior odds ratio is also very large for any data even if $n = 1$. Thus the standard experiment should be chosen as in Section 2.2.

3 FURTHER DISCUSSIONS

3.1 Robustness to the Standard Experiment

Let $\tilde{y}$ be an observable n-dimensional random vector and assume that

$$\tilde{y}|M_j, \beta_j, \tau \sim N_n\left(X_j\beta_j, \tau^{-1}I_n\right), \tag{3.1}$$

where X_j is an $n \times k_j$ matrix of rank k_j. The standard experiment is chosen to be

$$\tilde{y}^I|M_j, \beta_j, \tau \sim N_{n_0}\left(X_j^I\beta_j, \tau^{-1}I_{n_0}\right), \tag{3.2}$$

where $n_0 = \max\{k_j : j = 1, \cdots, m\} + 1$ and X_j^I is an $n_0 \times k_j$ matrix of rank k_j.

One may choose X_j^I to be

$$X_j^{I_1} = \begin{bmatrix} I_{k_j} \\ 0 \end{bmatrix}. \tag{3.3}$$

However, one may choose X_j^I to satisfy the relation that

$$nX_j^{I_2\prime}X_j^{I_2} = n_0X_j'X_j. \tag{3.4}$$

In terms of the Fisher information, the information of the standard experiment is n_0/n times as much as the information of the actual data under (3.4) . In the i.i.d. case, the standard experiment based on the minimal sample principle satisfies this relation. Thus (3.4) is thought to be generalization of this principle.

Once X_j^I is chosen, the noninformative prior for the model selection is

$$\pi_j(\beta_j, \tau) = \frac{\tau^{k_j/2-1}}{2^{n_0}\pi^{k_j/2}B((n_0-k_j)/2, (n_0-k_j)/2)}\left|X_j^{I\prime}X_j^I\right|^{1/2}. \tag{3.5}$$

The posterior probability is given by

$$p(M_j|y, X) \propto \frac{2^{k_j/2}}{B((n_0-k_j)/2, (n_0-k_j)/2)}\frac{\left|X_j^{I\prime}X_j^I\right|^{1/2}}{\left|X_j'X_j\right|^{1/2}} \left\|y - X_j\hat{\beta}\right\|^{-n}. \tag{3.6}$$

where $\hat{\beta}$ is the MLE of β.

When l types of the standard experiments, $\tilde{x}^{I_1}, \cdots \tilde{x}^{I_l}$, are considered, the subadditive probability is defined to be

$$\underline{p}(M_j|x) = \min\{p(M_j|x, I_k) : k = 1, \cdots, l\}, \tag{3.7}$$

where $p(M_j|x, I_k)$ is the posterior probability based on the standard experiment $\tilde{x}^{I_k}$. The quantity

$$\sum_{j=1}^{m} \underline{p}(M_j|x), \tag{3.8}$$

could be an index of the similarity among the standard experiments. Apart from the philosophical arguments concerning the subadditive probability, the quantity in (3.7) may be useful practically as a summary of the results. For philosophical discussions, see Walley (1991) and its reference.

We specialize the formulas to the ANOVA1 model. We consider 2 models whose explanatory variable matrix is $X_1 = \iota_{n_1+n_2}$, and

$$X_2 = \begin{bmatrix} \iota_{n_1} & 0 \\ 0 & \iota_{n_2} \end{bmatrix}.$$

Fig. 1 shows the performance of the posterior probabilities of the model1 versus n_1 with $n_2 = n_1^2$. The true values are $\beta_2 = (0, 0.4)$ and $\tau = 1$; thus M_1 is false. The solid line is for (3.3) , the dotted line is for (3.4) . The other line is for the quantity (3.8) . In this example, it always holds that $\underline{p}(M_1|x) = p(M_1|x, I_1)$ and $\underline{p}(M_2|x) = p(M_2|x, I_2)$.

3.2 Partially Noninformative Prior

Let $\tilde{x}$ be an observation whose distribution is parameterized by (θ, τ). The models are $M_1 : \theta = \theta_0$ and $M_2 : \theta \not= \theta_0$. In this case, it may happen that the value θ_0 is preferable than other values even under the model M_2, since this model selection problem itself is posited based on some

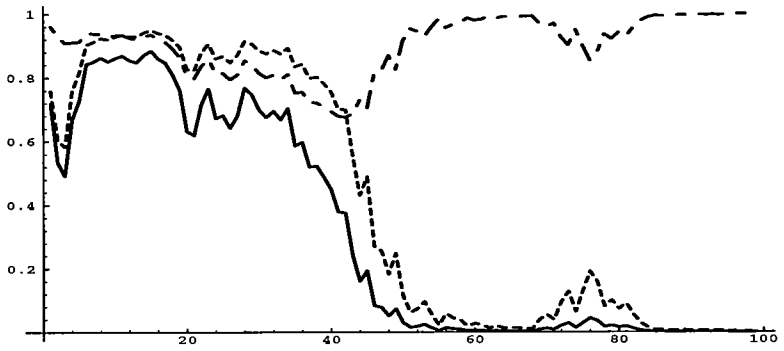

Figure 1: The Posterior Probabilities of M_1 based on the Different Standard Experiments and the index of their Similarity versus n_1 when M_1 is False: I_1: (Solid Line), I_2:(Dotted Line).

prior information which suggest the value θ_0. This partial prior knowledge is incorporated into the model M_2 by defining the prior based on the standard experiment $\tilde{x}^I$ by

$$\pi_2(\theta, \tau) = E\left[p_2(\theta, \tau|\tilde{x}^I)|\theta_0, \tau\right], \tag{3.9}$$

while

$$\pi_1(\tau) = E\left[p_1(\tau|\tilde{x}^I)|\theta_0, \tau\right]. \tag{3.10}$$

Example 1 Let $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ be an i.i.d. sample from $N(\theta, 1)$. Let $M_1 : \theta = \theta_0$ and $M_2 : \theta \in \Re$. In this case $\dim(\tau) = 0$. The standard experiment is supposed to be $\tilde{x}^I|M_2, \theta \sim N(\theta, 1)$. The initial prior is supposed to be $p_2(\theta) \propto 1$. Then the prior based on the standard

experiment is

$$\pi_2(\theta) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{1}{4}(\theta - \theta_0)^2\right). \tag{3.11}$$

Example 2 Let $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ be an i.i.d. sample from $N(\theta, \tau^{-1})$. Let $M_1 : \theta = \theta_0$ and $M_2 : \theta \in \Re$. The initial prior is supposed to be $p_2(\theta, \tau) \propto \tau^{-1}$. If the standard experiment $\tilde{x}^I$ is a 2-dimensional random vector such that

$$\tilde{x}^I | M_2, \theta, \tau \sim N_2(\theta \iota_2, I_2), \qquad (\theta, \tau) \in \Re \times (0, \infty), \tag{3.12}$$

then the prior based on the standard experiment is

$$\pi_2(\theta, \tau) = \frac{1}{\pi^{3/2} 2^{3/2} \sqrt{\tau}} \exp\left(-\frac{\tau}{2}(\theta - \theta_0)^2\right). \tag{3.13}$$

However, if we choose the standard experiment $\tilde{x}^I = (\tilde{x}_1^I, \tilde{x}_2^I)$ such that

$$\tilde{x}^I | M_2, \theta, \tau, \tilde{x}_2^I = x_2^I \sim N(\theta, (x_2^I)^{-1}), \tag{3.14}$$

and

$$\tilde{x}_2^I | M_2, \theta, \tau \sim Gamma\left(\frac{1}{2}, 2\tau\right), \tag{3.15}$$

for $(\theta, \tau) \in \Re \times (0, \infty)$, then

$$\pi_2(\theta, \tau) = \frac{\Gamma(3/2)}{4\pi^{3/2}\sqrt{\tau}} \left(1 + \tau \frac{(\theta - \theta_0)^2}{4}\right)^{-3/2}. \tag{3.16}$$

If the prior independence between θ and τ is preferred, the latter analysis is favorable because it holds that

$$p(\theta, \tau | x^I) = p(\theta | x^I) p(\tau | x^I), \tag{3.17}$$

for the latter standard experiment.

3.3 Lindley's Paradox

Let $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$ be an i.i.d. sample from $N(\theta, 1)$. The models are $M_1 : \theta = 0$ and $M_2 : \theta \in \Re$. If the conditional distribution under M_2 is assumed to be $p(\theta|M_2) = n(\theta|\mu, \tau^{-1})$ with $p(M_1) = p(M_2) = 1/2$, then the posterior odds ratio is

$$\frac{\sqrt{n+\tau}}{\sqrt{\tau}} \exp\left(-\frac{1}{2}\left(n\bar{x}^2 - \frac{n\tau}{n+\tau}(\bar{x}-\mu)^2\right)\right), \tag{3.18}$$

where $\bar{x}$ is the sample mean. If $\tau \to 0$, this ratio goes to ∞ for all data. Thus even in the proper case, if the variance of the prior distribution τ^{-1} is very large compared with the variance of the sample mean $\bar{x}$, i.e. $\tau^{-1} \gg n^{-1}$, it tends to overwhelm the information from data. This phenomenon is called Lindley paradox by Shafer (1982), though Lindley (1957) pays attention to the behavior when $n \to \infty$. Although this is a controversial problem, it may be interesting to consider it from the information theoretic view.

We consider the case where the proper conditional distribution $p(\theta_j|M_j)$ is given based on the evidence from other sources. When all the models are simple, the maximum entropy rule or other information theoretic arguments (e.g.,Bernardo (1979)) may yield $p(M_j) \propto 1$. However, if some models are not simple, this is not the case. Indeed, in our discussion, we can define the entropy w.r.t. $\pi_j(\theta_j)$ to be

$$-\sum_{j=1}^{m} p(M_j) \log p(M_j) + \sum_{j=1}^{m} p(M_j) Ent(M_j), \tag{3.19}$$

where

$$Ent(M_j) = -\int_{\Theta_j} p(\theta_j|M_j) \log \frac{p(\theta_j|M_j)}{\pi_j(\theta_j)} d\theta_j, \tag{3.20}$$

for the composite models and $Ent(M_j) = 0$ for the simple models. Note that the underlying measure and the conditional probability distribution are both one point probability measure for the simple models. Given $p(\theta_j|M_j)$, the most uninformative prior distribution on the whole structure is obtained by maximizing (3.19) w.r.t. $p(M_1), \cdots, p(M_m)$ under the constraint that $\sum_{j=1}^{m} p(M_j) = 1$. The solution is

$$p(M_j) \propto \exp(Ent(M_j)). \tag{3.21}$$

Example Let us return to the problem above. The noninformative prior is given by $\pi_2(\theta) = 1/2\sqrt{\pi}$. The prior probability of the models are given by $p(M_1) \propto 1$ and

$$p(M_2) \propto \exp\left(\frac{1-\log 2}{2}\right)\frac{1}{\sqrt{\tau}}. \tag{3.22}$$

Then the posterior odds ratio in favor of M_1 is

$$\exp\left(\frac{\log 2 - 1}{2}\right)\sqrt{n+\tau}\exp\left(-\frac{1}{2}\left(n\bar{x}^2 - \frac{n\tau}{n+\tau}(\bar{x}-\mu)^2\right)\right). \tag{3.23}$$

Thus Lindley's paradox disappeared in the sense of Shafer (1982), while it does not disappear in the sense of Lindley (1957).

REFERENCES

Aitkin, M. (1991), " Posterior Bayes Factors," (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 53, 111-142.

Akaike, H. (1991), Comment on Aitkin's "Posterior Bayes Factors," *Journal of the Royal Staistical Society*, Ser. B. 53,

Akman, V. E. and Raftery, A. E. (1986), "Bayes Factors for Non-Homogenious Poisson Processes with Vague Prior Information," *Journal of the Royal Satistical Soceiety,* Ser. B. 48, 322-329.

Bartlett, M. S. (1957), Comment on D. V. Lindley's "Statistical Paradox", *Biometrika,* 44, 533-534.

Berger, J. O. and Pericchi, L. R. (1995), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo et al., London: Oxford University Press, pp. 23-42.

Berger, J. O. and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Predicition," *Journal of the American Statistical Association*, 91, 109-122.

Berk, R. H. (1966), "Limiting Behavior of Posterior Distributions when the Model is Incorrect," *The Annals of Mathematical Statistics,* 37, 51-58.

Berk, R. H. (1970), "Consistency a Posteriori," *The Annals of Mathematical Statistics,* 41, 894-906.

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference," (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 113-147 .

Bernardo, J. M. (1980),"A Bayesina Analysis of Classical Hypotheis Testing," in *Bayesian Statistics 1*, eds. J. M. Bernardo et al., London: Oxford University Press, pp. 605-618.

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

Dawid, A. P. (1995), Comment on O'Hagan's "Fractional Bayes Factors for Model Comparisons," *Journal of Royal Statistical Society,* Ser. B, 57, 124-124.

Dmochowski, J. (1995), "Properties of Intrinsic Bayes Factors," Ph. D. Dissertation, Purdue Universtiy, Department. of Statistics.

Edwards, W., Lindman, H. and Savage, L. J. (1963). "Bayesian Statistical Inference for Psychological Research," *Psychological Review* 70, 193-242. [Reprinted in *Robustness of Bayesian Analysis*, 1984, (eds. J. Kadane), Amsterdam: North-Holland.]

Früwirth-Schnatter, S. (1995). "Bayesian Model Discrimination and Bayes Factors for Linear Gaussian State Space Models, " *Journal of the Royal Statistical Society,* Ser. B, 57, 237-246.

Iwaki, K. (1988). "A Bayesian Inference on a Statistical Model with a Structural Change," *The Journal of Economics Studies (The University of Tokyo),* 31, 1-10, (in Japanese).

Iwaki, K. (1992). "Bayesian Testing in the Growth Curve Model," *Journal of Economics* (Asia University), 17-2, 1-27, (in Japanese).

Iwaki, K. (1996) "Posterior Expected Marginal Likelihood for Testing Hypotheses," Technical

Report 96-13, Purdue University, Department of Statistics.

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

Kass, R. E. and Wasserman, L. (1995) "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterioni," *Journal of the American Statistical Association*, 90, 928-934.

Klein, R. W. and Brown, S. J. (1984), "Model Selection when There is "Minimal" Prior Information", *Econometrica*, 52, 1291-1312

Lempers, F. B. (1971), *Posterior Probabilities of Alternative Linear Models,* Rotterdam: University of Rotterdam Press.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrka,* 44, 187-192

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparisons," (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 57, 99-138.

Pericchi, L. R. (1984), "An Alternative to the Standard Bayesian Procedure for Discrimination between Normal Linear Models," *Biometrika,* 44, 187-192.

Proschan, F. (1963), "Theoretical Explanation of Observed Decreasing Failure Rate," *Technometics* 5, 375-383.

Shafer,G. (1982), "Lindley's Paradox," (with discussion), *Journal of the American Statistical Association*, 77, 325-351.

Sono, S. (1986), "On Simple Bayesian Learning, Constraint of Parameter Space, and Berger's View." *Economic Studies (Hokkaido University),* 35-4, 130-139, (in Japanese).

Spiegelhater, D. J. and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information," *Journal of the Royal Statistical Society,* Ser. B, 44, 377-387.

Suzuki, Y. (1983), "On Bayesian Approach to Model Selection," in *ISI contributed papers,* Madrid Vol.1, 288-291.

Suzuki, Y. (1992), "Class-convergence of Measure and its Application to Bayesian Statistics," *TIMIS Journal* (Tama Institute of Management & Information Sciences) 25.

Varshavsky, J. A. (1995), "Intrinsic Bayes Factors for Model Selection with Autoregresive Data," Techinical Report 95-23, Purdue University, Department of Statistics.

Wally, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, London:Chapman and Hall.

Zellner, A. (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and some Specific Results," in *Basic Issues in Econometrics,* Chicago: University of Chicago Press, 275-305.

Zellner, A. and Siow, A. (1980), "Posterior Odds for Selected Regression Hypotheses," in *Bayesian Statistics 1*, eds. J. M. Berenardo et al., Valencia : Valencia University Press, pp. 585-603. Reply, pp. 638-643.

A APPENDIX:PROOFS OF THE THEOREM

Let $u(\theta) = p(\theta|\theta^*)/\pi(\theta)$. Let

$$p(\tilde{x}_{(n)}) = \int_\Theta p(\tilde{x}_{(n)}|\theta)p(\theta|\theta^*)d\theta,$$

and

$$q(\tilde{x}_{(n)}) = \int_\Theta p(\tilde{x}_{(n)}|\theta)\pi(\theta)d\theta.$$

Let

$$\tilde{\alpha}(n,\varepsilon) = \frac{\int_{U(\theta^*,\varepsilon)^c} p(\tilde{x}_{(n)}|\theta)\pi(\theta)d\theta}{\int_{U(\theta^*,\varepsilon)} p(\tilde{x}_{(n)}|\theta)\pi(\theta)d\theta}.$$

Let

$$\varphi_1(\varepsilon) = \inf\{u(\theta) : \theta \in U(\theta^*,\varepsilon)\},$$

and

$$\varphi_2(\varepsilon) = \sup\{u(\theta) : \theta \in U(\theta^*,\varepsilon)\}.$$

Since $u(\theta)$ is continuous, we have

$$\lim_{\varepsilon \to 0} \varphi_i(\varepsilon) = u(\theta^*), \qquad (i = 1, 2).$$

Since $u(\theta)$ is bounded, we have $\bar{u} = \sup\{u(\theta) | \theta \in \Theta\}$. Then it holds that

$$\frac{\varphi_1(\varepsilon)}{1 + \tilde{\alpha}(n, \varepsilon)} q(\tilde{x}_{(n)}) \leq p(\tilde{x}_{(n)}).$$

It also holds that

$$p(\tilde{x}_{(n)}) \leq \varphi_2(\varepsilon)(1 + \bar{u}\tilde{\alpha}(n, \varepsilon)\varphi_1(\varepsilon)^{-1}) q(\tilde{x}_{(n)}).$$

Thus

$$\frac{\varphi_1(\varepsilon)}{1 + \tilde{\alpha}(n, \varepsilon)} \leq \frac{p(\tilde{x}_{(n)})}{q(\tilde{x}_{(n)})} \leq \varphi_2(\varepsilon)(1 + \bar{u}\tilde{\alpha}(n, \varepsilon)\varphi_1(\varepsilon)^{-1}).$$

Thus

$$\varphi_1(\varepsilon) \leq \lim_{n \to \infty} \frac{p(\tilde{x}_{(n)})}{q(\tilde{x}_{(n)})} \leq \varphi_2(\varepsilon), \qquad (P(\cdot | \theta^*) - a.e.).$$

Since this holds for any $\varepsilon > 0$, it holds that

$$\lim_{n \to \infty} \frac{p(\tilde{x}_{(n)})}{q(\tilde{x}_{(n)})} = u(\theta^*) = 1, \qquad (P(\cdot | \theta^*) - a.e.).$$