POSTERIOR EXPECTED MARGINAL LIKELIHOOD FOR TESTING HYPOTHESES

by

Katsuaki Iwaki
Purdue University and Asia University, Japan

Technical Report #96-13

Department of Statistics
Purdue University
West Lafayette, IN USA

April 1996

Posterior Expected Marginal Likelihood for Testing Hypotheses

Katsuaki Iwaki *

Abstract

In this paper we introduce the *posterior expected marginal likelihood* for testing hypotheses when prior information is vague under some of the hypotheses. This method is based on a *data dependent prior* which is derived from a noninformative prior using the idea of an *imaginary minimal training sample*. In a sense, this data dependent prior is that which is most compatible with the data among the reasonable priors. This method is compared with existing (modified) Bayesian methods for testing hypotheses. The examples suggest that the method is widely applicable compared with other methods.

KEY WORDS: Baysian inference; Data dependent prior; Imaginary minimal training sample; Intermediate prior; Intrinsic Bayes factor; Intrinsic prior; Multiple comparison of hypotheses; Noninformative prior; Nonnested hypotheses; Post-experimental probability distribution; Post-experimental odds

*Katsuaki Iwaki is Assistant Professor, Faculty of Economics, Asia University, Musashino-shi, Tokyo, 180, Japan. The author thanks James O. Berger, Luis R.Pericchi for helpful comments.

1 INTRODUCTION

In this paper we discuss a unified approach towards testing statistical hypotheses. It will be shown that this approach has wide applicability; it is applicable not only to usual testing problems but also to testing multiple hypotheses and testing non-nested hypotheses. Model selection is also covered.

The approach is aimed at obtaining the post-experimental probability of hypotheses as Bayesians do. When informative proper priors are not available, there is controversy as to how to develop automatic Bayesian procedures for testing hypotheses. One tradition is the training sample idea, where the data is divided into two parts, the first part being used to obtain proper priors and the second part being used to compute probabilities of the hypotheses. Example of this approach can be found in , e.g., Lempers (1971), Frühwirth-Schnatter, S. (1995), O'Hagan (1995) and Berger and Pericchi (1996). This paper also follows this line.

In Section 2, we define the data dependent prior and the related notion of the posterior expected marginal likelihood (PEML), giving its motivations and an example. In Section 3, we review existing Bayesian and modified Bayesian methods briefly but critically, using the linear regression model as an example. In Section 4, we give further examples in order to show the wide applicability of the suggested method and to point out difficulties with other methods.

In this paper, all random variables are distinguished from their realizations by tilde.

2 POSTERIOR EXPECTED MARGINAL LIKELIHOOD

In statistical practice, it is usual that a statistician has a natural idea of the sample size for his or her problem. Thus we assume this in this section. The ambiguity about

this notion will be discussed in the concluding remarks.

Suppose that the observable random variable with sample size n, $\tilde{x}_{(n)} = \langle \tilde{x}_1, \cdots, \tilde{x}_n \rangle$, has a parameterized probability density belonging to $\{p(x|H, \theta_H) \mid H \in \mathcal{H}, \theta_H \in \Theta_H\}$, w.r.t. an appropriate measure $m^{(n)}(\cdot)$ on $rng(\tilde{x}_{(n)})$ (the range of $\tilde{x}_{(n)}$). This set is called the *model*. The set $\mathcal{H}$ is a finite set and its elements are called the hypotheses. The hypothesis-specific set Θ_H is a set of adjustable parameters and it is an open subset of $\Re^{d_H}$ or a singleton for a simple hypothesis. In the former case, $\dim \Theta_H = d_H$ and in the latter case, $\dim \Theta_H = 0$.

We now introduce a density on Θ_H denoted by $p_H(\theta_H)$. This is a density w.r.t. $\nu_H(\cdot)$, which is Lebesgue measure if $\dim \Theta_H > 0$ or the one point probability measure if $\dim \Theta_H = 0$. In the standard situation where we have a *probability* distribution on $\mathcal{H}$, $\langle p(H) \mid H \in \mathcal{H} \rangle$, and a conditional *probability* density on Θ_H, $p(\theta_H \mid H)$, for H satisfying $\dim \Theta_H > 0$, this density $p_H(\theta_H)$ is defined by

$$p_H(\theta_H) = \begin{cases} p(H), & \dim \Theta_H = 0 \\ p(H)p(\theta_H|H), & \dim \Theta_H > 0 \end{cases}. \tag{2.1}$$

However, we also need to consider improper prior distribution, and the conditional distribution in the right-hand side of (2.1) can not necessarily be defined in the improper prior distribution case. For further elaboration of this point, see the Appendix and Dawid(1995). Thus we instead start with $p_H(\theta_H)$ and define

$$p(H) := \int_{\Theta_H} p_H(\theta_H)\nu_H(d\theta_H), \tag{2.2}$$

and, if $0 < p(H) < \infty$, we define

$$p(\theta_H|H) := p_H(\theta_H)/p(H). \tag{2.3}$$

for $\theta_H \in \Theta_H$. This notation is also applied to the posterior distributions. Incidentally, if $\Theta_H = \{\theta_0\}$, then $\int_{\Theta_H} f(\theta_H)\nu_H(d\theta_H) = f(\theta_0)$. This expression is introduced only for notational convenience.

Example. Assume that

$$\tilde{x}_{(n)} = \langle \tilde{x}_1, ..., \tilde{x}_n \rangle | \theta \ i.i.d. \sim N(\theta, 1), \qquad \theta \in \Re,$$

and the problem is to test $H_1 : \theta = 0$ vs $H_2 : \theta \neq 0$. Then we have $\mathcal{H} = \{H_1, H_2\}$, $\Theta_{H_1} = \{0\}$ and $\Theta_{H_2} = \Re$ (or $\Theta_{H_2} = \Re - \{0\}$). In this case, a noninfomative prior that might be considered is $p_{H_1}(0) = c_1$ and $p_{H_2}(\theta_{H_2}) = c_2$, for some appropriate choice of the ratio of c_1 and c_2. Note that, whatever this ratio may be, $p(H_2) = \infty$ and $p(\theta_{H_2}|H_2)$ can not be defined.

We assume that $p_H(\cdot)$ is determined only up to multiplicative constant. Let n_0 be the minimal sample size defined to be the sample size for which it holds that

$$\int_{\Theta_H} p_H(\theta_H | x_{(n_0)}) \nu_H(d\theta_H) < \infty, \qquad (m^{(n_0)}(\cdot) - a.e. x_{(n_0)}), \tag{2.4}$$

for any $H \in \mathcal{H}$ and it holds that

$$\int_{\Theta_H} p_H(\theta_H | x_{(n_0-1)}) \nu_H(d\theta_H) = \infty, \qquad (m^{(n_0-1)}(\cdot) - a.e. x_{(n_0-1)}), \tag{2.5}$$

for some $H \in \mathcal{H}$. We assume the existence of the minimal sample size. Let $\tilde{x}^I$ be a random variable which is independent of $\tilde{x}_{(n)}$ and has the same distribution as $\tilde{x}_{(n_0)}$ given $H \in \mathcal{H}$ and $\theta_H \in \Theta_H$. We call $\tilde{x}^I$ the imaginary minimal training sample. If we could observe $\tilde{x}^I = x^I$, we would have

$$p(\theta_H | H, x^I) \propto p(x^I | H, \theta_H) p_H(\theta_H), \tag{2.6}$$

for H satisfying $p(H) = \infty$. Although this distribution would not be noninformative, it would be "minimally informative", since n_0 is the minimal sample size for the posterior to be proper. Thus

$$\{p(\theta_H | H, x^I) \mid x^I \in rng(\tilde{x}^I)\}, \tag{2.7}$$

is a set of reasonable candidates for the prior for θ_H. Although we can not observe $\tilde{x}^I = x^I$, we can estimate the value of $p(\theta_H | H, \tilde{x}^I)$ by its posterior expectation,

$$\bar{p}_{x_{(n)}}(\theta_H \mid H) \ := \ E[p(\theta_H | H, \tilde{x}^I) \mid H, \tilde{x}_{(n)} = x_{(n)}], \tag{2.8}$$

$$= \int_{rng(\tilde{x}^I)} p(\theta_H \mid H, x^I) p(x^I | H, x_{(n)}) m^{(n_0)}(dx^I),$$

where

$$p(x^I | H, x_{(n)}) = \int_{\Theta_H} p(x^I | H, \theta_H) p(\theta_H | H, x_{(n)}) d\theta_H.$$

With this data dependent prior, we define a modified marginal likelihood to be

$$\bar{p}(x_{(n)} | H) := \int_{\Theta_H} p(x_{(n)} | H, \theta_H) \bar{p}_{x_{(n)}}(\theta_H | H) d\theta_H, \tag{2.9}$$

which is called the posterior expected marginal likelihood (PEML). Then we define the post-experimental probability of the hypothesis to be

$$\bar{p}(H | x_{(n)}) \propto \begin{cases} \bar{p}(x_{(n)} | H) & \text{when } p(H) = \infty \\ \int_{\Theta_H} p(x | H, \theta_H) p(\theta_H | H) \nu_H(d\theta_H) & \text{when } p(H) < \infty. \end{cases} \tag{2.10}$$

See the Appendix concerning this definition.

Example (continued). The usual improper prior is given by $p_{H_2}(\theta_{H_2}) \propto 1$. The minimal sample size is then $n_0 = 1$. The distribution of the imaginary minimal training sample $\tilde{x}^I$ is given by

$$p(x^I | H_2, \theta_{H_2}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x^I - \theta_{H_2})^2).$$

The conditional distribution on Θ_{H_2}, given $\tilde{x}^I = x^I$, is

$$p(\theta_{H_2} | H_2, x^I) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\theta_{H_2} - x^I)^2).$$

The predictive distribution of $\tilde{x}^I$ given the actual data is

$$p(x^I | H_2, x_{(n)}) = \frac{\sqrt{n}}{\sqrt{2\pi(n+1)}} \exp(-\frac{n}{2(n+1)}(x^I - \bar{x}_n)^2).$$

where $\bar{x}_n = (x_1 + \cdots + x_n)/n$. Thus the data dependent prior is

$$\bar{p}_{x_{(n)}}(\theta_{H_2} | H_2) = \frac{\sqrt{n}}{\sqrt{2\pi(2n+1)}} \exp(-\frac{n}{2(2n+1)}(\theta_{H_2} - \bar{x}_{(n)})^2). \tag{2.11}$$

For this data dependent prior, the post-experimental odds ratio of H_1 to H_2 is given by

$$\frac{\bar{p}(H_1|x_{(n)})}{\bar{p}(H_2|x_{(n)})} = \sqrt{2(n+1)}\exp(-\frac{n\bar{x}_n^2}{2}). \tag{2.12}$$

The data dependent prior has the following asymptotic property. Let the true distribution of $\tilde{x}_{(\infty)} = \langle \tilde{x}_1, \cdots, \tilde{x}_n, \cdots \rangle$ be $P_0(\cdot)$. We assume that there exists θ_H^* such that

$$\lim_{n\to\infty} \frac{\int_{U(\theta_H^*,\varepsilon)^c} p(\tilde{x}_{(n)}|H,\theta_H)p_H(\theta_H)d\theta_H}{\int_{U(\theta_H^*,\varepsilon)} p(\tilde{x}_{(n)}|H,\theta_H)p_H(\theta_H)d\theta_H} = 0 \qquad (P_0(\cdot) - a.e.), \tag{2.13}$$

for any $\varepsilon > 0$, where $U(\theta_H^*, \varepsilon) = \{\theta_H \in \Theta_H | \, \|\theta_H - \theta_H^*\| < \varepsilon\}$. We can refer to Berk (1966, 1970), Sono (1986) and Dmochowski (1995) concerning the assumption (2.13), which implies that the posterior probability mass on Θ_H accumulates in the neighborhood of the value $\theta_H^* \in \Theta_H$.

Now we define

$$p^0(\theta_H|H,\theta_H^0) := \int_{rng(\tilde{x}^I)} p(\theta_H|H,x^I)p(x^I|H,\theta_H^0)m^{(n_0)}(dx^I). \tag{2.14}$$

Then we have the following proposition under the assumption (2.13).

Proposition 1. If $p^0(\theta_H|H,\theta_H^0)$ is continuous w.r.t. θ_H^0,

$$\bar{p}_{\tilde{x}_{(n)}}(\theta_H|H) \to p^0(\theta_H|H,\theta_H^*) \qquad (P_0(\cdot) - a.e.). \tag{2.15}$$

Example (continued) When the true distribution of an i.i.d.sequence $\langle \tilde{x}_1, \cdots, \tilde{x}_n, \cdots \rangle$ is the normal distribution with mean θ_0 and variance 1, then $\theta_{H_1}^* = \theta_0$ and

$$\bar{p}_{\tilde{x}_{(n)}}(\theta_{H_2}|H_2) \to \frac{1}{\sqrt{4\pi}}\exp\left(-\frac{(\theta_{H_1}-\theta_0)^2}{4}\right) \qquad (a.e.). \tag{2.16}$$

The following lemma is important in the field of Bayesian testing hypothesis.

Lemma Let $p_H^S(\theta_H)$ be a proper prior distribution. Assume that $p^S(\theta_H|H)/p_H(\theta_H)$ is continuous and bounded above. Then under the assumption (2.13), it holds that

$$\frac{\int_{\Theta_H} p(\tilde{x}_{(n)}|\theta_H)p^S(\theta_H|H)d\theta_H}{\int_{\Theta_H} p(\tilde{x}_{(n)}|\theta_H)p_H(\theta_H)d\theta_H} = O(1) \qquad (P_0(\cdot) - a.e.) \tag{2.17}$$

Note that these three assumptions are corresponding to the conditions for the princple of stable estimation. See Edwards, Lindman and Savage (1963). By the proposition 1 and the lemma, the following proposition holds.

Proposition 2. Assume (2.13). If $p(\theta_H|H, \theta_H^*)/p_H(\theta_H)$ is continuous and bounded above, then it holds that

$$\frac{\bar{p}(\tilde{x}_{(n)}|H)}{\int_{\Theta_H} p(\tilde{x}_{(n)}|\theta_H)p_H(\theta_H)d\theta_H} = O(1) \qquad (P_0(\cdot) - a.e.) \tag{2.18}$$

The data dependent prior here is sensible, although its result is not sequentially coherent, i.e., it does not hold that

$$\bar{p}(H|x_1, x_2) \propto \bar{p}(H|x_1)p(x_2|H, x_1)$$

Given $\tilde{x}_{(n)} = x_{(n)}$, there could exsit a subjective Bayesian who happens to have the same prior distribution as the data dependent prior. Since his or her inference is meaningful, the inference based on the data dependent prior is also *meaningful*. Recall, also, that this data dependent prior is minimally informative, so that the ensuing analysis essentially allows the data to "speak for itselt." Finally, its limiting distribution is the expectation on a set of reasonable priors (2.6) w.r.t. the nearest distribution to the true distribution.

Aitkin (1991) proposed using the posterior distribution from the entire data and a noninformative prior as the data dependent prior for this problem. Note that it becomes increasingly informative as the sample size becomes larger, while the data dependent prior proposed here stays minimally informative. Thus our method agrees with proper Bayes solutions asymptotically, as is shown above.

3 APPLICATION TO LINEAR MODELS AND COMPARISON WITH OTHER METHODS

3.1 The PEML Approach

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

$$\tilde{y}|H, \beta_H, \sigma_H \sim N_n\left(X_H\beta_H, \sigma_H^2 I_n\right), \tag{3.1}$$

where X_H is an $n \times k_H$ matrix of rank k_H. We shall now consider the PEML for this problem. It is easy to see that $n_0 = \max\{k_H \mid H \in \mathcal{H}\} + 1$. We choose

$$p_H(\beta_H, \sigma_H) \propto 1/\sigma_H, \tag{3.2}$$

as the noninformative prior under the hypothesis H. The imaginary minimal training sample is chosen to be

$$\tilde{y}^I|H, \beta_H, \sigma_H \sim N_{n_0}\left(X^I\beta_H, \sigma_H^2 I_{n_0}\right), \tag{3.3}$$

where X^I is an $n_0 \times k_H$ matrix satisfying the condition that

$$nX_H^{I\prime}X_H^I = n_0 X_H'X_H. \tag{3.4}$$

Note that (3.4) can be rewritten as

$$n\mathcal{I}_H[\tilde{y}|\beta, \sigma] = n_0\mathcal{I}_H[\tilde{y}^I|\beta, \sigma], \tag{3.5}$$

where $\mathcal{I}_H[\tilde{y}|\beta,\sigma]$ is Fisher information matrix for the model (3.1), so (3.4) is a natural condition ensuing that the imaginary minimal training sample has information which is appropriately proportional to the information in the actual data. With some computation the PEML is given by

$$\begin{aligned}\bar{p}(y|H,X_H) &= \frac{\sqrt{n_0}^{k_H}\Gamma\left((n+n_0-k_H)/2\right)^2}{\sqrt{n+n_0}^{k_H}\sqrt{\pi}^{n}\Gamma\left((n_0-k_H)/2\right)^2} \\ &\quad \frac{\Gamma(n_0-k_H)\Gamma\left(n-k_H/2\right)}{\Gamma\left((n-k_H)/2\right)\Gamma\left(n+n_0-k_H\right)}\left\|y-X_H\hat{\beta}_H\right\|^{-n}. \qquad (3.6)\end{aligned}$$

3.2 Zellner and Siow's Method

For testing hypotheses concerning whether the mean in the normal law is zero or not, Jeffreys (1961) employed the Cauchy distribution as his prior. Zellner and Siow (1980) generalized his method to the linear regression model.

Let H_0 be the largest hypothesis, i.e. $\{X_H\beta \mid \beta \in \Re^{k_H}\} \subset \{X_{H_0}\beta \mid \beta \in \Re^{k_{H_0}}\}$ for any $H \in \mathcal{H}$. If there is no such hypothesis, we can create it with $p(H_0) = 0$. Any hypothesis H is now compared with H_0 as follows. First, they reparametrize β_{H_0} as $\beta_{H_01} = (X_H'X_H)^{-1}X_H'X_{H_0}\beta_{H_0}$ and $\beta_{H_02} = (X_H^{\perp\prime}X_H^{\perp})^{-1}X_H^{\perp\prime}X_{H_0}\beta_{H_0}$ where $X_H^{\perp}$ satisfies $\{X\beta|\beta \in \Re^{k_{H_0}}\} = \{\left(X_H, X_H^{\perp}\right)\beta|\beta \in \Re^{K_{H_0}}\}$, and $X_H'X_H^{\perp} = 0$. Then, the Zellner and Siow prior is given by

$$\begin{aligned}&p_{H_0}(\beta_{H_01},\beta_{H_02},\sigma_{H_0}) \\ &\quad = c\frac{\Gamma\left((k_{H_0}-k_H+1)/2\right)\sqrt{\pi}^{k_H-k_{H_0}-1}}{\left(1+\beta_{H_02}'X_H^{\perp\prime}X_H^{\perp}\beta_{H_02}/n\sigma_{H_0}^2\right)^{(k_{H_0}-k_H+1)/2}}\left|\frac{X_H^{\perp\prime}X_H^{\perp}}{n\sigma_{H_0}^2}\right|^{1/2}\frac{1}{\sigma_{H_0}}, \qquad (3.7)\end{aligned}$$

and $p_H(\beta_H,\sigma_H) = c/\sigma_{H_0}$.

When $\#\mathcal{H} > 2$ ($\#\mathcal{H}$ is the number of elements in $\mathcal{H}$), this prior distribution varies with each "alternative" hypothesis, H, which is compared with H_0. In this sense, the mothod is not a pure Bayesian method. As far as these two hypotheses are concerned, their posterior odds ratio, PO_{H/H_0}, is given by

$$\frac{\sqrt{\pi}}{2^{(k_{H_0}-k_H)/2}}\left\|y-X_H\hat{\beta}_H\right\|^{-n+k_H}/$$

$$\int_0^\infty \frac{\alpha^{(k_{H_0}-k_H-1)/2}e^{-\alpha}}{(2\alpha+n)^{(k_{H_0}-k_H)/2}} \left(\left\|y - X_{H_0}\hat{\beta}_{H_0}\right\|^2 + \frac{2\alpha}{2\alpha+n} y' X_H^{\perp} \left(X_H^{\perp\prime} X_H^{\perp}\right)^{-1} X_H^{\perp\prime} y\right)^{-(n-k_H)/2} d\alpha, \quad (3.8)$$

where $\hat{\beta}_H = (X_H' X_H)^{-1} X_H' y$. When $\#\mathcal{H} > 2$, the post-experimental probability of the hypotheses can be seen to be

$$p(H|y) \propto \begin{cases} PO_{H/H_0} & \text{when } H \neq H_0 \\ 1, & \text{when } H = H_0. \end{cases} \quad (3.9)$$

Note that the exact solution (3.8) is due to the present author; Zellner and Siow(1980) gave an approximation to (3.8) based on Jeffreys's idea. Jeffreys(1961) also gave a formula for the exact solution in a special case but it is computationally more burdensome since it contains the confluent hypergeometric function as an integrand.

Zellner and Siow's method is not a pure Bayesian method because it uses the largest hypothesis as a "pivot". Interestingly, it is possible to generalize Jeffreys's method in a pure Bayesian way by using the smallest hypothesis as a pivot, although we do not pursue this here.

3.3 Spiegelhalter and Smith's Method

Suppose that

$$p_H(\theta_H) = c_H \left(\mathcal{I}_H[\tilde{x}^I|\theta_H]\right)^{1/2}. \quad (3.10)$$

Once $c_{H_1/H_2} := c_{H_1}/c_{H_2}$ is determined, the posterior odds ratio in favour of H_1 against H_2 is given by

$$PO(x_{(n)}) = c_{H_1/H_2} \frac{\int_{\Theta_{H_1}} p(x_{(n)}|H_1,\theta_{H_1}) p_{H_1}(\theta_{H_1}) \nu_{H_1}(d\theta_{H_1})}{\int_{\Theta_{H_2}} p(x_{(n)}|H_2,\theta_{H_2}) p_{H_2}(\theta_{H_2}) \nu_{H_2}(d\theta_{H_2})}, \quad (3.11)$$

which is somewhat misleadingly called the Bayes Factor by Spiegelhalter and Smith (1982) (see the Appendix). They introduced the following idea to determine c_{H_1/H_2} when H_1 is nested within H_2:

$$1 = \sup\left\{PO(x^I)|x^I \in rng(\tilde{x}^I)\right\}. \quad (3.12)$$

Since the relation, $c_{H_1/H_2}c_{H_2/H_3} = c_{H_1/H_3}$, does not hold, one must define $c_{H_1/H_2} := c_{H_1/H_0}/c_{H_2/H_0}$ using one of the hypotheses, H_0, as a pivot. This idea can be seen in Akman and Raftery (1986), using the smallest hypothesis as a pivot. Since one could as well use the largest hypothesis as a pivot, there is a degree of arbitrariness here.

For the linear model, the choice of the explanatory variable for the minimal training sample is controversial, as can been seen in O'Hagan(1995). Here we choose the same explanatory variable as for the PEML, given by (3.4). Then, using the largest hypothesis,H_0, as a pivot, we have the posterior probability of the hypotheses given by

$$p(H|y, X_H) \propto \left(\frac{k_{H_0}+1}{n}\right)^{k_H/2} \left\|y - X_H\hat{\beta}_H\right\|^{-n}, \tag{3.13}$$

where $\hat{\beta}_H = (X_H'X_H)^{-1}X_H'y$. Note that the assumption in (3.10) is essential for this method. Indeed, if we instead use the prior $p_H(\beta_H, \sigma_H) \propto \sigma_H^{-1}$ for the linear model, the right-hand side of (3.12) is infinity.

This method is defined for nested hypotheses. However, if $\inf\left\{PO(x^I)| x^I \in rng(\tilde{x}^I)\right\} = -\infty$ and $\sup\left\{PO\left(x^I\right)|x^I \in rng(\tilde{x}^I)\right\} < \infty$, this method can be generalized to the case of non-nested hypotheses. In this case, we say that H_2 is more complex that H_1. For example, it can be shown that the i.i.d. log-normal hypothesis with two-dimensional unknown parameter is more complex than the i.i.d. exponential hypothesis with unknown scale parameter, and that the i.i.d. negative binomial hypothesis with unknown ratio is more complex than the i.i.d. Poisson hypothesis with unknown mean. Although this generalization is interesting, the method does not handle non-nested problems in general. For example, if we compare the i.i.d. log-normal hypothesis with two-dimensional unknown parameter and the i.i.d. Weibull hypothesis with two-dimensional unknown parameter, neither the infimum nor the supremum is finite, and this method is not applicable.

3.4 Suzuki's method

Suzuki(1992) introduced the notion of relative covergence. His definition is equivalent to

the following definition in our situation.

Definition Let $p_H(\theta_H)$ be a density on Θ_H w.r.t. the measure $\nu_H(\cdot)$ and let $\langle p_H(\theta_H|\lambda) \mid \lambda \in N\rangle$ be a sequence of densities on Θ_H w.r.t. $\nu_H(\cdot)$. We say that $\langle\langle p_H(\theta_H|\lambda) \mid H \in \mathcal{H}\rangle \mid \lambda \in N\rangle$ converges relatively to $\langle p_H(\theta_H) \mid H \in \mathcal{H}\rangle$ iff there exists a sequence $\langle c_\lambda \in \Re|\lambda \in N\rangle$ such that for any $H \in \mathcal{H}$

$$p_H(\theta_H) = \lim_{\lambda\to\infty} c_\lambda p_H(\theta_H|\lambda). \tag{3.14}$$

The $p_H(\theta_H|\lambda)$ are called the intermediate priors of $p_H(\theta_H)$. Let $p_H(\theta_H|x_{(n)})$ and $p_H(\theta_H|x_{(n)},\lambda)$ be posterior densities of $p_H(\theta_H)$ and $p_H(\theta_H|\lambda)$, respectively. Under regularity conditions which allow the limit to pass inside the integral, it can be shown that

$$p_H(\theta_H|x_{(n)}) = \lim_{\lambda\to\infty} p_H(\theta_H|x_{(n)},\lambda). \tag{3.15}$$

Example (continued from Section 2). Let $p_{H_1}(0|\lambda) = p(H_1) = 1/2$. Let

$$p_{H_2}(\theta_{H_2}|\lambda) = \frac{1}{2}\frac{D\,(-\lambda < \theta_{H_2} < \lambda)}{2\lambda},$$

where $D(proposition) = 1$ if the proposition is true and $D(proposition) = 0$ otherwise. Then it holds that this sequence converges relatively to the one point probability measure with $p(H_1) = 1$ and $p(H_2) = 0$, although $p(H_1|\lambda) = p(H_2|\lambda) = 1/2$ for any $\lambda \in N$. Thus, for any data, $p(H_1|x_{(n)}) = 1$ and $p(H_2|x_{(n)}) = 0$.

This last phenominon was noted by Bartlett (1957). Since the phenomenon should be avoided, Suzuki(1983) proposed that uncertainty concerning each hypothesis should be balanced. Specifically, when there are no simple hypotheses, he proposed the condition

for the intermediate prior probability distributions that the entropies of $p(\theta_H|H,\lambda)$ are the same. Here we put $p(H|\lambda) = 1/\#\mathcal{H}$, because of the reason in the Appendix. For the linear model, the intermediate prior probability distributions,

$$p(\beta_H, \sigma_H|H,\lambda) = \frac{1}{\lambda\sigma_H} D(-\lambda^{\frac{1}{(k_H+1)}} / 2 < \log \sigma_H < \lambda^{\frac{1}{(k_H+1)}} / 2)$$
$$\prod_{i=1}^{k_H} D(-\lambda^{\frac{1}{(k_H+1)}} /2 < \beta_{Hi} < \lambda^{\frac{1}{(k_H+1)}} /2), \tag{3.16}$$

satisfy the above condition where $\beta_H = \langle \beta_{H1}, \cdots, \beta_{Hk_H} \rangle$. The posterior probability is given by

$$p(H|y, X_H) \propto \frac{\pi^{k_H/2}}{|X_H' X_H|^{1/2}} \Gamma\left(\frac{n-k_H}{2}\right) \left\| y - X_H \hat{\beta}_H \right\|^{-(n-k_H)}, \tag{3.17}$$

where $\hat{\beta}_H = (X_H' X_H)^{-1} X_H y$. However, if we replace $(X_H' X_H/n)^{1/2}\beta_H$ by β_H, this reparametrization yields

$$p(H|y, X_H) \propto \frac{\pi^{k_H/2}}{n^{k_H/2}} \Gamma\left(\frac{n-k_H}{2}\right) \left\| y - X_H \hat{\beta}_H \right\|^{-(n-k_H)}. \tag{3.18}$$

Thus Suzuki's method is not invariant under reparametrization. Furthermore, the solution depends on the choice of unit of the dependent variable. The idea of obtaining a "balanced" prior through intermediate priors is, however, of considerable importance. See the reference in Suzuki (1992) for more about this idea.

3.5 Klein and Brown's Method

Klein and Brown(1984) considered a different type of intermediate prior. Given a sequence of probability densities, $\langle p(\theta_H|H,\lambda)|\lambda \in N\rangle$, for H satisfying $\dim \Theta_H > 0$, they chose $\langle\langle p(H|\lambda)|H \in \mathcal{H}\rangle|\lambda \in N\rangle$ to minimize

$$\sum_{H\in\mathcal{H}} p(H|\lambda) \log p(H|\lambda) + \sum_{H:\dim\Theta_H>0} p(H|\lambda) I(H|\lambda), \tag{3.19}$$

where $I(H|\lambda)$, the index of the prior information of the conditional distribution under H, is either

$$I_A(H|\lambda) := E[\log p(\tilde{\theta}_H|H,\lambda)|H,\lambda] - E[\log p(\tilde{x}^I|H,\tilde{\theta}_H)|H,\lambda], \tag{3.20}$$

or

$$I_B(H|\lambda) := E[-\log \frac{p(\tilde{\theta}_H|H, \tilde{x}^I, \lambda)}{p(\tilde{\theta}_H|H, \lambda)}|H, \lambda], \tag{3.21}$$

where $\tilde{x}^I$ is an imaginary training sample and $\tilde{\theta}_H$ is a random variable whose density is $p(\theta_H|H, \lambda)$. The quantity (3.19) can be considered to be the total prior information. The resulting intermediate posterior probailities of the hypotheses are given by

$$p(H|x_{(n)}, \lambda) \propto p(H|\lambda) \int_{\Theta_H} p(x_{(n)}|H, \theta_H) p(\theta_H|H, \lambda) d\theta_H, \tag{3.22}$$

where $p(H|\lambda) \propto \exp(-I(H|\lambda))$. The final posterior probability is obtained by taking the limit of the intermediate posterior probabilities.

Note that (3.22) depends on the choice of intermediate priors. For instance, in the example in Section 2, the posterior probability when the intermediate prior is chosen to be normal and that when the intermediate prior is chosen to be uniform are different, even if both of the intermediate priors converge to Lebesgue measure relatively. Since $I_A(H|\lambda)$ is not invariant under reparametrization, $I_B(H|\lambda)$ is theoretically more appealing.

For the linear model, Klein and Brown's intermediate prior is given by

$$p(\beta_H|H, \sigma_H, \lambda) = \frac{1}{(2\pi)^{k_H/2} \sigma_H^{k_H} |V_\lambda|^{1/2}} \exp\left(-\frac{(\beta_H - \beta_{0H})' V_\lambda^{-1} (\beta_H - \beta_{0H})}{2\sigma_{k_H}^2}\right), \tag{3.23}$$

and

$$p(\log \sigma_H|H, \lambda) = \frac{D(\log \delta_{1\lambda} < \log \sigma_H < \log \delta_{2\lambda})}{\log \delta_{2\lambda} - \log \delta_{1\lambda}}, \tag{3.24}$$

where V_λ is a $k_H \times k_H$ positive definite symmetric matrix and the minimum eigenvalue of V_λ goes to infinity as $\lambda \to \infty$. It is also assumed that $\delta_{1\lambda} \to 0$ and $\delta_{2\lambda} \to \infty$ as $\lambda \to \infty$. Their imaginary minimal training sample is a k_H-dimensional random vector when H is true and its conditional distribution is

$$\tilde{y}^I|H, \beta_H, \sigma_H \sim N_{k_H}\left(X_H^I \beta_H, \sigma_H^2 I_{k_H}\right), \tag{3.25}$$

where X_H^I is a $k_H \times k_H$ matrix such that

$$X_H^{I\prime} X_H^I = X_H' X_H / n. \tag{3.26}$$

Their intermediate prior under a hypothesis converges relatively to the prior (3.10). The limiting posterior probability of the hypothesis is given by

$$p(H|y, X_H) \propto \frac{1}{n^{k_H/2}} \|y - X_H\beta_H\|^{-n}. \tag{3.27}$$

Klein and Brown discussed the relationship between parametrization and invariance w.r.t. the choice of explanatory variables which leave the subspace spanned by their columns unchanged. Use of their suggestion is compatible with use of (3.26), which, in addition, leaves the posterior probability invariant.

Bernardo(1980) and Pericchi (1984) tried a similar idea, but they defined the information index differently to be

$$I_B^{BP}(H|\lambda, n) := E[-\log \frac{p(\tilde{\theta}_H|H, \tilde{x}_{(n)}, \lambda)}{p(\tilde{\theta}_H|H, \lambda)} \mid H, \lambda], \tag{3.28}$$

where $\tilde{x}_{(n)}$ is the real observable random variable. Since $I_B^{BP}(H|\lambda, n)$ is a function not only of the prior information but also of the sample size, it is not a pure index of the prior information. Thus it is more controversial.

3.6 Berger and Pericchi's Method

Berger and Pericchi (1993,1995) proposed the intrinsic Bayes factor. There are two versions of the intrinsic Bayes factor, the arithmetic and the geometric, but here we consider only the former because the authors think it more important.

First they assume that $p_H(\theta_H)$ is proportional to the usual noninformative prior. Let $x(l) := \langle x_{i_1}, \cdots, x_{i_k} \rangle$ for $l = \{i_1 < \cdots < i_k\} \subset \{1, \cdots, n\}$. Let L be a set of $l \in \{1, \cdots, n\}$ such that for any $H \in \mathcal{H}$, $p(H|x(l)) < \infty$ and for any proper subset $l' \subset l$, there exists a hypothesis $H \in \mathcal{H}$ such that $p(H|x(l')) = \infty$.

If H_1 is nested in H_2, their intrinsic Bayes factor is defined to be

$$B_{21}^{AI} := \frac{1}{\#L} \sum_{l \in L} \frac{\int_{\Theta_{H_2}} p(x(l^c)|H_2, \theta_{H_2}) p(\theta_{H_2}|H_2, x(l)) d\theta_{H_2}}{\int_{\Theta_{H_1}} p(x(l^c)|H_1, \theta_{H_1}) p(\theta_{H_1}|H_1, x(l)) d\theta_{H_1}}, \tag{3.29}$$

$$= PO(x_{(n)})_{21} \frac{1}{\#L} \sum_{l \in L} PO(x(l))_{12}, \tag{3.30}$$

where $PO(x_{(n)})_{ij}$ is the posterior odds ratio in favour of the hypothesis H_i against H_j based on the noninformative prior above. The latter factor in (3.30) is called the correction factor. Note that the undefined constants are canceled out by multiplying the two factors.

Pure Bayes factors possess the properties that $B_{ij} = 1/B_{ji}$ and $B_{ij} = B_{ik}/B_{jk}$. The arithmetic intrinsic Bayes factor does not possess these properties, so that generalizations to non-nested cases and multiple comparison cases need rather involved modifications. In the nested case,B_{12}^{AI} must be defined to be $1/B_{21}^{AI}$ because, if one tries to define B_{12}^{AI} by (3.29), the correction factor does not converge when n is large. Since, under the assumptions in the Lemma in Section 2, the ratio of $PO(x_{(n)})_{21}$ and any subjective Bayes factor is $O(1)$, the divergence should be avoided. However, in non-nested situations, it can be unclear as to which hypothesis should be H_1 in (3.29). Furthermore, there are cases where correction factors diverge for both definitions. For example, consider the model $\tilde{x}_{(n)}|\theta_1, \theta_2$ *i.i.d.* $\sim N(\theta_1, \theta_2)$ and the hypotheses, $H_1 : \theta_1 = 1$ v.s. $H_2 : \theta_1 = 2$. The most natural solution to this difficulty is to create, if necessary and if possible, an encompassing hypothesis, within which the other hypotheses are nested, and to define the intrinsic Bayes factors by using this hypothesis as a pivot.

An interesting feature of the intrinsic Bayes factor is that, in many interesting cases, there exists a pair of proper priors on Θ_{H_1} and Θ_{H_2} which yield a genuine Bayes factor which is equivalent to the intrinsic Bayes factor in the sense that the ratio between the two converges to 1 as the sample size grows. These priors are called the intrinsic priors. Intrinsic priors need not be unique. In many interesting nested models, the conditional probability of the parameter of interest given the nuisance parameter under the alternative hypothesis is proper. The existence of intrinsic priors not only reinforces the validity of the method, but the approach also provides a systematic way of deriving a default priors for testing hypotheses. More details about the existence and integrability of the intrinsic prior can be found in Dmochowski (1995).

In the linear regression model, there are intrinsic priors for each hypothesis compared

with the encompassing hypothesis. However, this does not hold in general for multiple comparison problems. For example, consider the model $\tilde{x}_{(n)}|\theta_1, \theta_2$ *i.i.d.* $\sim N(\theta_1, \theta_2)$ and the hypotheses, $H_1 : \theta_1 = 0, \theta_2 = 1$, $H_2 : \theta_1 = 0, \theta_2 \in (0, \infty)$ and $H_0 : \theta_1 \in \Re, \theta_2 \in (0, \infty)$. Although H_0 is the encompassing hypothesis in this model, the intrinsic prior on H_0 is incompatible when it is compared with H_1 and H_2. But, in this case, it is possible to construct a set of coherent intrinsic priors by choosing the smallest hypothesis as the pivot. This strategy is also useful when the method is applied to the change-point problem discussed in Booth and Smith (1982), Broemling and Tsurumi (1987), Iwaki (1988) and Perez (1994).

For the linear regression model with the prior $p_H(\beta_H, \sigma_H) \propto \sigma_H^{-1}$ the intrinsic Bayes factor in favour of H_0 against H is

$$\frac{p(y|X_{H_0})}{p(y|X_H)} \frac{1}{\#L} \sum_{l \in L} \frac{p(y(l)|X_{H_0}(l))}{p(y(l)|X_H(l))}, \tag{3.31}$$

where

$$p(y|X) = |X'X|^{-1/2} \Gamma\left(\frac{n-k}{2}\right) \left\|y - X\hat{\beta}\right\|^{-(n-k)},$$

$$k = rankX,$$

$$\hat{\beta} = (X'X)^{-1}X'y,$$

$$y(\{i_1 < \cdots < i_{n_0}\}) = \langle y_{i_1}, \cdots, y_{i_{n_0}} \rangle',$$

$$X(\{i_1 < \cdots < i_{n_0}\}) = \langle x_{i_1}, \cdots, x_{i_{n_0}} \rangle' \qquad \text{for} \qquad X = \langle x_1, \cdots, x_n \rangle',$$

$$L = \{l \subset \{1, \cdots, n\} \mid \#l = n_0 \wedge \forall \lambda \in [rankX_H(l) = rankX_H]\},$$

and

$$n_0 = rankX_{H_0} + 1.$$

It should be remarked that the result of this method is invariant w.r.t. the choice of the explanatory variables, in so far as the subspaces spanned by their columns are the same, while additional assumptions concerning the explanatory variables are needed to obtain the results for the other methods. For example, the assumption (3.4) is important for the method of the PEML.

3.7 O'Hagan's Method

O'Hagan (1995) proposed the fractional Bayes factor for testing hypotheses. The fractional Bayes factor is defined to be

$$B_{12}(b) := \frac{q_b(x_{(n)}|H_1)}{q_b(x_{(n)}|H_2)}, \tag{3.32}$$

where

$$q_b(x_{(n)}|H) := \frac{\int_{\Theta_H} p(x|\theta_H)p_H(\theta_H)\nu_H(d\theta_H)}{\int_{\Theta_H} p(x|\theta_H)^b p_H(\theta_H)\nu_H(d\theta_H)}, \tag{3.33}$$

for $0 < b < 1$. Note that this definition requires $p_H(\theta_H)$ to be defined only up to a multiplicative constant. If we apply this method to the regression model with the noninformative prior (3.2), we have

$$B_{12}(\frac{n_0}{n}) = \frac{\Gamma((n-k_{H_1})/2)\Gamma((n_0-k_{H_2})/2)}{\Gamma((n-k_{H_2})/2)\Gamma((n_0-k_{H_1})/2)} \left(\frac{\left\|y - X_{H_2}\hat{\beta}_{H_2}\right\|}{\left\|y - X_{H_1}\hat{\beta}_{H_1}\right\|}\right)^{n-n_0}, \tag{3.34}$$

where $\hat{\beta}_H = (X_H'X_H)^{-1}X_H'y$ $(H = H_1, H_2)$. O'Hagan is not specific as to how to choose n_0 or b. However, we choose $n_0 = 1 + max\{k_H|H \in \mathcal{H}\}$, following the spirit of his numerical examples and Berger and Pericchi (1995).

3.8 Numerical Examples and Comparisons

We first apply the above methods to Hald's regression data from Draper and Smith (1966). There are four potential regressors, which we denote by 1,2,3,4 and a constant term (included in all hypotheses) which we denote by c. The sample size is $n = 13$, and the minimal training sample size is $n_0 = 6$.

In Table 1, we give the post-experimental probabilities of the hypotheses that the indicated subset of the regressors is correct, for the PEML, for Zellner and Siow's method (denoted by ZS1), for the pure Bayesian extension of Jeffreys's method (denoted by ZS2), for the method of Spiegelhalter and Smith (denoted by SS), for the method of Suzuki with (3.18), for the method of Klein and Brown (denoted by KB), for the AIBF with (3.2),

and for the method of O'Hagan (denoted by FBF). We also give the expected values of the dimension (EVD) of the hypotheses for each method.

We next specialize the formulas to the ANOVA1 model. The independent observations are

$$\tilde{y}_{ij}|\beta_1, \cdots, \beta_I, \sigma^2 \sim N(\beta_i, \sigma^2) \qquad i = 1, \cdots, I; j = 1, \cdots, J,$$

and we wish to compare the hypotheses,

$$H_1 : \beta_1 = \cdots = \beta_I \in \Re \text{ vs } H_2 : \langle \beta_1, \cdots \beta_I \rangle \in \Re^I.$$

We analyze two data sets which appeared in Box and Tiao (1972, pp.246-247), with $I = 6$ and $J = 5$, for the various methods. Note that the difference between ZS1 and ZS2 disappears because we have only two hypotheses. So we write just ZS.

It seems that the method of Suzuki prefers is seriously weighted towards the more complex hypotheses. However, it should be remembered that the result of this method depends on the choice of the scale in the dependent variable. If this choice is not sensible, the interpretation of the results may not be sensible. This choice requires a possibly difficult subjective judgement.

Spiegelhalter and Smith's method also seems to prefer the more complex hypotheses. This is because their posterior odds ratio in favour of the simpler hypothesis is always less than 1 for the minimal sample size. This ratio might be deemed to be the infimum of the reasonable automatic post-experimental odds ratios.

It should be noted that the difference between ZS-1 and ZS-2 in Table 1 is not negligible. This suggests that the choice between the largest hypothesis and the smallest hypothesis as a pivot is an important decision in general.

Hypothesis	ZS-2	AIBF	ZS-1	FBF	PEML	KB	SS	Suzuki
1,2,3,4,c	0.01	0.05	0.07	0.06	0.05	0.07	0.18	0.36
1,2,3,c	0.09	0.17	0.18	0.20	0.23	0.23	0.25	0.20
1,2,4,c	0.10	0.19	0.19	0.20	0.24	0.23	0.25	0.20
1,3,4,c	0.08	0.16	0.15	0.16	0.16	0.16	0.17	0.15
2,3,4,c	0.02	0.04	0.03	0.04	0.01	0.01	0.02	0.03
1,2,c	0.54	0.28	0.28	0.24	0.25	0.25	0.11	0.05
1,3,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1,4,c	0.17	0.11	0.10	0.10	0.05	0.05	0.02	0.01
2,3,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2,4,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3,4,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4,c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EVD	3.30	3.66	3.69	3.72	3.76	3.77	4.04	4.30

Table 1: The Post-experimental Probabilities for Hald's Data.

data set	ZS	AIBF	FBF	PEML	KB	SS	Suzuki
D1	0.15	0.16	0.18	0.05	0.17	0.00	0.00
D2	0.99	0.96	0.99	1.00	1.00	0.88	0.01

Table 2: The Post-experimental Probabilities of H_1 for the Data from Box and Tiao (1972).

4 OTHER EXAMPLES

4.1 Testing for a Particular Value of a Normal Mean

Let $\tilde{x}_{(n)}|\theta \ i.i.d \sim N(\theta, 1)$ and let $H_1 : \theta = 0$ and $H_2 : \theta \in \Re$. In Table 3, the post-experimental odds ratios are listed. Zellner (1984) said "the posterior odds ratio's value reflects (a) the prior odds ratio's value, (b) relative precision of prior and posterior distributions for parameters, (c) relative goodness of fit of the two models (hypotheses in our terminology), and (d) extent to which prior and sample information regarding parameters' values are in agreement." In Table 3, we decompose the odds ratios into the product of three factors corresponding to (a)×(b), (c) and (d). The relative goodness of fit (c) is common to all these methods. The factor corresponding to (d) is normalized to be 1 when $\bar{x} = 0$. Jeffreys (1961, p.274) uses the Cauchy prior for this model, and the approximation to his posterior odds is also listed in the Table. The posterior odds ratio of Spiegelhalter and Smith (1982) and Klein and Brown (1984) are the same for this model, if we use the normal distribution as the intermediate prior for the latter method. This common value is listed for SS and KB. The intrinsic prior for the adjustable parameter of the hypothesis H_2 is

$$p^I(\theta|H_2) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{\theta^2}{4}\right). \tag{4.1}$$

The Bayes factor corresponding to this proper prior is listed as IP. O'Hagan's fractional Bayes factor with $b = 1/n$ is listed as FBF. Suzuki's method is not applicable to this model.

When we view this Table, we notice that the factor (c) is dominant and the results of all the methods are close to each other if n is large. But when the factor (c) is not dominant, the factor (d) may cause non-negligible differences.

The PEML does not cause disagreement between the prior belief regarding the parameter and the data information because the data dependent prior is used. The method of SS and KB does not cause disagreement because they use noninformative

	(a)×(b)	(c)	(d)
PEML	$\sqrt{2(n+1)/n}$		1
Jeffreys	$\sqrt{\pi/2}$		$1+\bar{x}^2$
SS and KB	1	$\sqrt{n}\exp(-n\bar{x}^2/2)$	1
IP	$\sqrt{(2n+1)/n}$		$\exp(n\bar{x}^2/2(2n+1))$
O'Hagan	1		$\exp(\bar{x}^2/2)$

Table 3: The Post-experimental Odds Ratios in favour of H_1 decomposed into three factors.

priors. On the other hand, the intrinsic prior and Jeffreys's prior under the alternative hypothesis prefer the value suggested by the null hypothesis, and this causes disagreement. The fractional Bayes factor behaves the same way. The methods of the former type are characterized by totally noninformative prior information, while the methods of the latter type presuppose at least partial prior information, explicitly or implicitly. Such prior information might be natural in many applications. But if these latter methods are used automatically and the information is not from real prior belief, the larger discrepancy between this fictitious prior information and the data makes the results more dubious.

When the discrepancy is small, the PEML and the intrinsic prior are almost the same. When it is large, the PEML prefers the more complex model to a greater extent than does the intrinsic prior. Spiegelhalter and Smith's method prefers the more complex hypothesis to an even larger extent.

4.2 Testing for Equality of Two Exponential Parameters

Let the density of $\tilde{x} = \langle \tilde{x}_{ij} \mid i = 1, 2; j = 1, \cdots, n_i \rangle$ be

$$p(x_{11}, \cdots, x_{2n_2} | \theta_1, \theta_2) = \theta_1^{n_1} e^{-t_1\theta_1} \theta_2^{n_2} e^{-t_2\theta_2}, \tag{4.2}$$

where $t_i = \sum_{j=1}^{n_i} x_{ij}$. Let $H_1 : \theta_1 = \theta_2 = \theta \in \Re$ and $H_2 : \langle \theta_1, \theta_2 \rangle \in \Re^2$. The priors

are $p_{H_1}(\theta) \propto \theta^{-1}$ and $p_{H_2}(\theta_1, \theta_2) \propto \theta_1^{-1}\theta_2^{-1}$. The imaginary minimal training sample is $\tilde{x}^I = \langle \tilde{x}_1^I, \tilde{x}_2^I \rangle$, and its density is

$$p(x_1^I, x_2^I | H_1, \theta) = \theta^2 e^{-(x_1^I + x_2^I)\theta}, \tag{4.3}$$

and

$$p(x_1^I, x_2^I \mid H_2, \theta_1, \theta_2) = \theta_1 e^{-x_1^I \theta_1} \theta_2 e^{-x_2^I \theta_2}. \tag{4.4}$$

The data dependent prior under each hypothesis is, respectively, given by

$$\bar{p}(\theta | H_1) = n(n+1)(t_1 + t_2)^n \theta \int_0^\infty \frac{z^3 e^{-z\theta}}{(t_1 + t_2 + z)^{n+2}} dz, \tag{4.5}$$

and

$$\bar{p}(\theta_1, \theta_2 | H_2) = \prod_{i=1}^{2} n_i t_i^{n_i} \int_0^\infty \frac{z_i e^{-z_i \theta_i}}{(t_i + z_i)^{n_i + 1}} dz_i, \tag{4.6}$$

where $n = n_1 + n_2$. The PEML is given by

$$\bar{p}(x | H_1) = \frac{3\Gamma(n+2)}{2(2n+1)(2n+3)(t_1 + t_2)^n}, \tag{4.7}$$

and

$$\bar{p}(x | H_2) = \frac{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}{4(2n_1 + 1)(2n_2 + 1) t_1^{n_1} t_2^{n_2}}. \tag{4.8}$$

In this example, we point out a difficulty with O'Hagan's method. For this model, the fractional Bayes factor is given by

$$B_{12}^b = PO_{12}^N \frac{\Gamma(bn_1)\Gamma(bn_2)}{\Gamma(b(n_1 + n_2))} \left(\frac{(t_1 + t_2)^{n_1 + n_2}}{t_1^{n_1} t_2^{n_2}} \right)^b, \tag{4.9}$$

where

$$\begin{aligned} PO_{12}^N &= \frac{\int_0^\infty p(x | H_1, \theta) p_{H_1}(\theta) d\theta}{\int_0^\infty \int_0^\infty p(x | H_2, \theta) p_{H_2}(\theta_1, \theta_2) d\theta_1 d\theta_2} \\ &= \frac{\Gamma(n_1 + n_2) t_1^{n_1} t_2^{n_2}}{\Gamma(n_1)\Gamma(n_2)(t_1 + t_2)^{n_1 + n_2}}. \end{aligned}$$

Let us consider the situation where we have far more data about θ_2 than θ_1; specifically suppose $n_2 = n_1^2$. If we choose the fraction $b = 2/(n_1 + n_2)$, then it is easily seen that

$\widetilde{PO}_{12}^{N} / \tilde{B}_{12}^{b} = o(1)$ $(a.e.)$. Let B_{12}^{S} be the Bayes factor of a subjective Bayesian whose prior density w.r.t. the noninformative prior is continuous and bounded above. Since the assumptions in the lemma in Section 2 hold, it follows that $\tilde{B}_{12}^{S} / \widetilde{PO}_{12}^{N} = O(1)$ $(a.e.)$, so that

$$\tilde{B}_{12}^{S} / \tilde{B}_{12}^{b} \to 0 \qquad (a.e.). \tag{4.10}$$

This problem may be avoided by introducing two fractions, b_1 and b_2, for the hypothesis H_2 but the approach then becomes considerably more complicated.

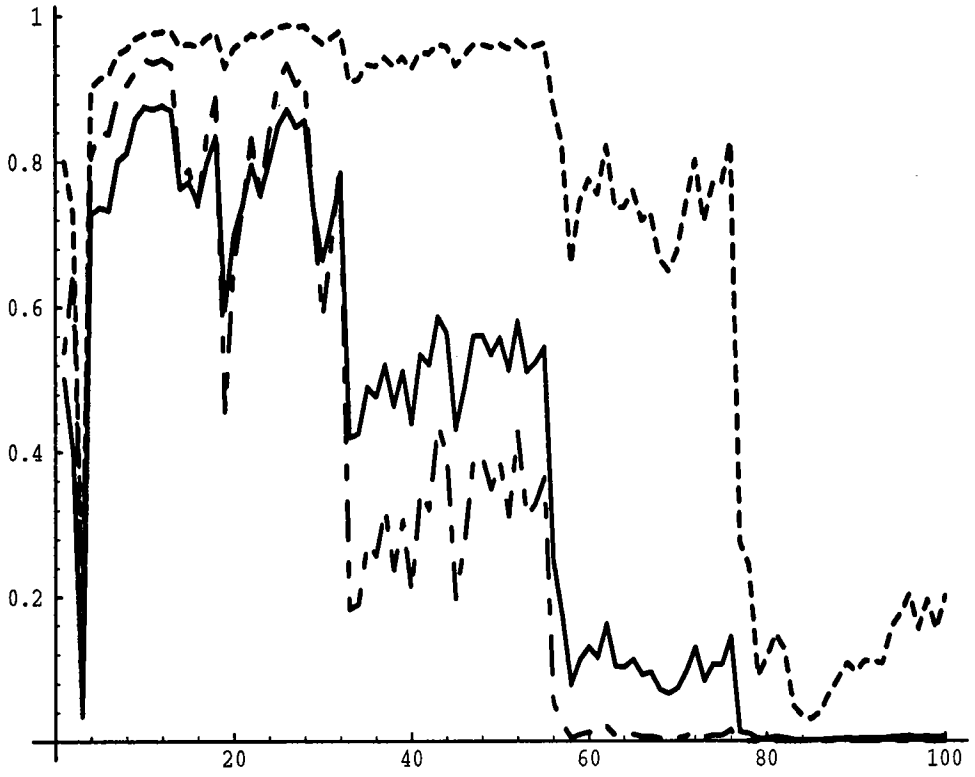

Figure 1: The Post-experimental Probabilities of H_1 versus n_1 when H_1 is False: PEML (Solid Line), FBF (Dotted Line) and Subjective Bayesian's .

Fig.1 shows the performance of the post-experimental probabilites of these methods versus n_1 with $n_2 = n_1^2$. The true values are $\theta_1 = 1$ and $\theta_2 = 0.75$; thus H_1 is false.

The prior of the subjective Bayesian is $p(\theta|H_1) = (2.5)^2\theta\exp(-2.5\theta)$, and $p(\theta_1,\theta_2|H_2) = \exp(-\theta_1 - 1.5\theta_2)$.

4.3 Testing for a Particular Value of a Poisson Parameter

Let $\tilde{x}_{(n)}$ be an i.i.d. sequence of random variables from the Poisson distribution with mean θ. Let $H_1 : \theta = \theta_0$ and $H_2 : \theta \in (0,\infty)$. The noninformative prior density under the hypothesis H_2 is given by $p_{H_2}(\theta) \propto \theta^{-1/2}$. The minimal sample size is 1. Then the data dependent prior concerning θ is given by

$$\bar{p}(\theta|H_2) = \frac{e^{-\theta}}{\Gamma(t+1/2)\sqrt{\theta}}\left(\frac{n}{n+1}\right)^{t+1/2}\sum_{z=0}^{\infty}\frac{\theta^z\Gamma(t+z+1/2)}{(n+1)^z z!\Gamma(z+1/2)}, \tag{4.11}$$

and the PEML of H_2 is given by

$$\bar{p}(x_{(n)}|H_2) = \frac{n^{t+1/2}}{x_1!\cdots x_n!\Gamma(t+1/2)(n+1)^{2t+1}}\sum_{z=0}^{\infty}\frac{\Gamma(t+z+1/2)^2}{\Gamma(z+1/2)z!(n+1)^{2z}}, \tag{4.12}$$

where $t = \sum_{i=1}^{n} x_i$

There is a problem with the method of Klein and Brown when it is applied to this model. Indeed, if we use the gamma distribution with scale parameter λ^{-1} as the intermediate prior for θ, which converges to $p_{H_2}(\theta)$, under the hypotheses H_2, and we adopt I_B as the information index, it follows that

$$\lim_{\lambda\to 0}\frac{p(H_1|x_{(n)})}{p(H_2|x_{(n)},\lambda)} = \infty. \tag{4.13}$$

This fact means that the hypothesis H_1 is always selected without regard to the data if λ is small enough.

4.4 A Non-nested Multicomparison Example

Let $\tilde{x}_{(n)}$ be an i.i.d. sequence of random variables. Suppose that they are from the exponential distribution with mean θ^{-1} under H_1, their logarithms are from $N(\mu,\sigma^2)$ under H_2 and they are from Weibull(α,β) under H_3. The noninfomative priors are $p_{H_1}(\theta) \propto \theta^{-1}$,

$p_{H_2}(\mu,\sigma) \propto \sigma^{-1}$ and $p_{H_3}(\alpha,\beta) \propto \alpha^{-1}\beta^{-1}$. The minimal sample size is 2. The PEMLs are given by

$$\bar{p}(x_{(n)}|H_1) = \frac{n\Gamma(n)}{2(2n+1)t^n}, \tag{4.14}$$

$$\bar{p}(x_{(n)}|H_2) = \frac{\sqrt{2}\Gamma((n+1)/2)^2\Gamma(n+1/2)}{\prod_{i+1}^{n} x_i\sqrt{n+2}\sqrt{\pi}^{n+2}\Gamma((n-1)/2)\Gamma(n+1)\left(\sum_{i=1}^{n}(\log x_i - \hat{\mu})^2\right)^{n/2}}, \tag{4.15}$$

and

$$\bar{p}(x_{(n)}|H_3) = \frac{2x_1^I x_2^I}{q(x_{(n)})}\int_0^\infty\int_0^\infty q\left(x_{(n)}, x_1^I, x_2^I\right)^2 \left|\log x_1^I - \log x_2^I\right| dx_1^I dx_2^I, \tag{4.16}$$

where $\hat{\mu} = \sum_{i=1}^{n} \log\ x_i\,/n$ and

$$q(x_{(n)}) = \Gamma(n)\int_0^\infty \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}\left(\sum_{i=1}^{n} x_i^\alpha\right)^{-n}\alpha^{n-2}d\alpha.$$

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 post-experimental probabilities by the PEML: $\bar{p}(H_1|x_{(n)}) = 0.414, \bar{p}(H_2|x_{(n)}) = 0.3699$ and $\bar{p}(H_3|x_{(n)}) = 0.217$.

4.5 A Non-stationary Example

Let $\tilde{x}_{(n)}$ be an independent sequence and let $\tilde{x}_i|\theta \sim N(\theta, i)$. Let $H_1 : \theta = 0$ and $H_2 : \theta \in \Re$. If we choose

$$\tilde{x}^I|\theta \sim N(\theta, f_0^{-1}) \tag{4.17}$$

as the imaginary training sample, the posterior odds ratio in favour of H_1 against H_2 is

$$\frac{\sqrt{2(f_0+F_n)}}{\sqrt{f_0}}\exp(-\frac{F_n}{2}\hat{\theta}^2), \tag{4.18}$$

where $F_n = \sum_{i=1}^{n} i^{-1}$, and $\hat{\theta} = F_n^{-1}\sum_{i=1}^{n} x_i/i$.

The IBF is

$$B_{12}^{AI} = PO_{12}^{N}\left(\frac{1}{n}\sum_{I=1}^{n}\frac{1}{\sqrt{i}}\exp\left(-\frac{x_i^2}{2i}\right)\right)^{-1}, \tag{4.19}$$

where

$$PO_{12}^{N} = \sqrt{F_n} \exp\left(-\frac{F_n \hat{\theta}^2}{2}\right). \tag{4.20}$$

However, since

$$\frac{1}{n} E\left[\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \exp\left(-\frac{\tilde{x}_i^2}{2i}\right) \middle| \theta\right] \to 0, \ (n \to \infty), \tag{4.21}$$

and

$$\sum_{i=1}^{\infty} V\left[\frac{1}{\sqrt{i}} \exp\left(-\frac{\tilde{x}_i^2}{2i}\right) \middle| \theta\right] / i^2 < \infty,$$

by the strong law of large numbers,

$$\tilde{B}_{12}^{AI} \Big/ \widetilde{PO}_{12}^{N} \to \infty, \qquad (a.e.). \tag{4.22}$$

Since the assumption (2.13) holds, the ratio between $\widetilde{PO}_{12}^{N}$ and a subjective Bayes factors is $O(1)$ almost everywhere under the assumptions concerning subjective priors in the lemma in Section 2. Therefore, the ratio between the intrinsic Bayes factor and a subjective Bayes factor also diverges almost everywhere. This example shows that, in non-stationary models, the choice of simple arithmetric averaging in the IBF may not be appropriate. Also, it might be difficult to choose statistically meaningful weights among the many possibilities. On the other hand, non-independency is not a serious obstacle to this method and, indeed, Varshavsky (1995) applied this method successfully to a stationary time series model.

The fractional Bayes factor is

$$B_{b_n}(x) = \exp\left(-\frac{1 - b_n}{2} F_n \hat{\theta}^2\right) \frac{1}{b_n}. \tag{4.23}$$

Thus the automatic choice of $b_n = 1/n$ is inappropriate and the choice should satisfy $b_n F_n = O(1)$

5 CONCLUDING REMARKS

Jeffreys's method and its extension by Zellner and Siow (1980) can not be generalized to non-nested models such as the model in Subsection 4.4. For multiple comparison

models, there is arbitrariness concerning the choice of the "pivot." For non-normal cases, Jeffreys (1961) gives an interesting suggestion. See also Kass and Wasserman (1995).

Spiegelhalter and Smith's method was originally defined only for nested models. Generalization to non-nested models is possible in some cases but not always. This method also has a degree of arbitrariness about the "pivot" for multiple comparison models. Finally, it seems to excessivly favour the more complex models.

Suzuki's method can be applied only when there are no hypotheses which have proper priors concerning the adjustable parameters. For this method, subjective elements such as the choice of unit of measurement and parametrization are crucial. We have an example where Klein and Brown's method is inappropriate in Subsection 4.3. Berger and Pericchi's method needs rather involved modifications when it is applied to non-nested or multiple comparison models, so it is more a "strategy" (Berger and Pericchi (1995)) than an algorithm. We gave an example where O'Hagan's method results in a serious discrepancy from reasonable subjective Bayesian methods in Subsection 4.2.

The most important notion is the balance of prior uncertainty. Since this can not be formulated in terms of an improper prior directly, Suzuki (1983) introduced the notion of an intermediate prior. Jeffreys keeps the balance by, roughly speaking, relating the amount of information in the prior to the amount of information contained in a unit observation, as is shown in (3.7). See Kass and Wasserman (1995) for more details. Other methods use the notion of the (imaginary) minimal training sample in their own ways.

However the general definition of the (imaginary) minimal training sample is difficult. The definition of Berger and Pericchi (1993) is general to some extent, but their idea is connected with using simple arithmetic or geometric means, which is inappropriate for non-stationary models such as the model in Subsection 4.5.

For non-stationary models or models with complex experiments, the PEML method needs a non-automatic definition of the imaginary training sample. However, this sub-

jective judgement is mainly concerned with the degree of uncertainty, and other aspects of subjective judgement are determined automatically. Also this judgement is directly related to the statistical model. Furthermore, for any finite sample, we can investigate the data dependent prior to judge wheter or not it is reasonable.

REFERENCES

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

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

Booth, N. B. and Smith, A. F. M. (1982), "A Bayesian Approach to Retrospective Identification of Change-Points," *Journal of Econometrics,* 19, 7-22

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

Broemling, L. D. and Tsurumi, H. (1987). *Econometrics and Structural Change,* New York; Marcel Dekker, Inc.

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.

Draper, N. R. and Smith, H. (1966). *Applied Regression Analysis,* NewYork: John Wiley & Sons.

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).

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 Statstical Paradox," *Biometrka,* 44, 187-192

Nihon Sugakkai (1977), *Encyclopedic Dictionary of Mathematics* , Cambridge Mass : MIT Press.

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

Perez, J. M. (1994) "Modelos de Intercambiabilidad Parcial," Master Thesis, Universidad Simon Bolivar, Carcas, Venezuela.

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

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

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 Liner 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.

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 REMARKS ON FORMAL ASPECTS OF THE DISCUSSION

For testing hypothesis problems, the statistical model of the observable random variable $\tilde{x}$ is given by the set of densities w.r.t. an appropriate underlying measure,

$$\{p(x|\bar{\theta})|\bar{\theta} \in \bar{\Theta}\}. \tag{1.1}$$

Here $\bar{\Theta}$ is the direct sum of $\langle \Theta_H | H \in \mathcal{H} \rangle$ defined as

$$\bar{\Theta} := \coprod_{H \in \mathcal{H}} \Theta_H \tag{1.2}$$

$$= \bigcup_{H \in \mathcal{H}} \bar{\Theta}_H, \tag{1.3}$$

where $\bar{\Theta}_H = \{H\} \times \Theta$. Note that $H_1 \neq H_2 \Rightarrow \bar{\Theta}_{H_1} \cap \bar{\Theta}_{H_2} = \emptyset$. A more abstract definition of the direct sum of sets can be found in Nihon Sugakkai (1977). In this paper, Θ_H is an open subset of $\Re^{d_H}$ or a singleton. For example, the parameter space of the model in Subsection 4.4 is

$$\bar{\Theta} = \{H_1\} \times \Re \cup \{H_2\} \times \Re^2 \cup \{H_3\} \times \Re^2, \tag{1.4}$$

and the model is

$$\tilde{x}_{(n)} i.i.d. \sim \begin{cases} Exp(\theta^{-1}) & \text{when } \bar{\theta} = \langle H, \theta \rangle \in \{H_1\} \times \Re \\ \log N(\mu, \sigma^2) & \text{when } \bar{\theta} = \langle H, \langle \mu, \sigma \rangle \rangle \in \{H_2\} \times \Re^2 \\ Weibull(\alpha, \beta) & \text{when } \bar{\theta} = \langle H, \langle \alpha, \beta \rangle \rangle \in \{H_3\} \times \Re^2 \end{cases}. \tag{1.5}$$

A Bayesian has a probability distribution on $\bar{\Theta}$, $P(\cdot)$. ¿From this, we define the probabilities of hypotheses, the conditional probability distribution given the hypothesis H and the restriction of the probability distribution on the hypotheses H:

$$p(H) := P(\bar{\Theta}_H), \tag{1.6}$$

$$P(A|H) := \frac{P(A \cap \bar{\Theta}_H)}{p(H)}, \tag{1.7}$$

$$P_H(\cdot) : \{B \cap \bar{\Theta}_H | B \in dom P(\cdot)\} \ni A \mapsto P(A) \in [0, 1]. \tag{1.8}$$

It is assumed that $p(H) > 0$. In our case, $P(\cdot|H)$ and $P_H(\cdot)$ are absolutely continuous w.r.t. Lebesgue measure, and their densities are written as $p(\cdot|H)$ and $p_H(\cdot)$, respectively, when Θ_H is an open subset of $\Re^{d_H}$. This notion is generalized to the case where $P(\bar{\Theta}) = \infty$ when no prior information is available. Even in this case, the analysis is done by Bayes formula:

$$P(d\bar{\theta}|x) \propto p(x|\bar{\theta}) P(d\bar{\theta}). \tag{1.9}$$

However, since $\exists H \in \mathcal{H}[p(H) = \infty]$, the conditional probability, $P(\cdot|H)$, can not be defined for this H. Thus the Bayes factor

$$\frac{\int_{\Theta_{H_1}} p(x|H_1, \theta_{H_1}) p(\theta_{H_1}|H_1) d\theta_{H_1}}{\int_{\Theta_{H_2}} p(x|H_2, \theta_{H_2}) p(\theta_{H_2}|H_2) d\theta_{H_2}}, \tag{1.10}$$

can not be defined. However, we can define the posterior odds ratio by

$$\frac{\int_{\Theta_{H_1}} p(x|H_1, \theta_{H_1}) p_{H_1}(\theta_{H_1}) d\theta_{H_1}}{\int_{\Theta_{H_2}} p(x|H_2, \theta_{H_2}) p_{H_2}(\theta_{H_2}) d\theta_{H_2}}. \tag{1.11}$$

This argument can be found in Dawid (1995).

It is natural to define $p_H(\theta_H) = c_H q_H(\theta_H)$ where $q_H(\theta_H)$ is the usual noninformative prior of the model $\{p(x|\theta_H)|\theta_H \in \Theta_H\}$. The problem here is how to choose the ratio

between $c_H/c_{H'}$ for the noninformative case. Several possible answers were discussed in the text. Once this ratio is determined to be, say 1, for the noninformative case and the statistician feels that H is k times more likely than H', this information may be expressed by setting this ratio k. However, since this strategy seems to require more judicious study on foundations of Bayesian statistics, we restricted consideration to the noninformative case. For the methods using the training sample idea, we can formally define the prior probability of hypotheses after observing the training sample. However, we also restricted consideration only to the noninformative case, and we set this probability being $1/\#\mathcal{H}$.