Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 74-86
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# Bayes Factor with Lindley Paradox and Tow Standard Methods in Model

**Xiaoting Nie**

Changsha Normal University, Changsha, China
Email: 1176292212@qq.com

**How to cite this paper:** Nie, X.T. (2020) Bayes Factor with Lindley Paradox and Tow Standard Methods in Model. *Open Journal of Statistics*, **10**, 74-86.
https://doi.org/10.4236/ojs.2020.101006

**Received:** December 1, 2019
**Accepted:** February 10, 2020
**Published:** February 13, 2020



## Abstract

For any statistical analysis, Model selection is necessary and required. In many cases of selection, Bayes factor is one of the important basic elements. For the unilateral hypothesis testing problem, we extend the harmony of frequency and Bayesian evidence to the generalized p-value of unilateral hypothesis testing problem, and study the harmony of generalized P-value and posterior probability of original hypothesis. For the problem of single point hypothesis testing, the posterior probability of the Bayes evidence under the traditional Bayes testing method, that is, the Bayes factor or the single point original hypothesis is established, is analyzed, a phenomenon known as the Lindley paradox, which is at odds with the classical frequency evidence of p-value. At this point, many statisticians have been worked for this from both frequentist and Bayesian perspective. In this paper, I am going to focus on Bayesian approach to model selection, starting from Bayes factors and going within Lindley Paradox, which also briefly talks about partial and fractional Bayes factor. Trying to use a simple way to consider this paradox is the thing what I want to do in the paper. On the other hand, a detailed derivation of BIC and AIC is given in Section 4. The guiding principle of selecting the optimal model is to investigate from two aspects: one is to maximize the likelihood function, the other is to minimize the number of unknown parameters in the model. The larger the likelihood function value, the better the model fitting, but we can not simply measure the model fitting accuracy, which leads to more and more unknown parameters in the model, and the model that becomes more and more complex would have caused an overmatch. Therefore, a good model should be the combination of the fitting accuracy and the number of unknown parameters to optimize the configuration.

## Keywords

Bayes Factors, Lindley Paradox, Fractional Bayes Factor, BIC, AIC

## 1. Introduction

Many statisticians are naturally involved in the question of model selection [1], in case to define the "best model" to fit real data, different approaches have been proposed since last century, many well-known methods such as F-test [2], AIC, BIC [3], Bayesian model averaging [4]. We are focusing on Bayesian approach, as we analyze data from some possible models $M_1,\cdots,M_k$. We denote $\theta\in\Theta$ as parameter and $\pi(\theta)$ as prior probability, then for likelihoods $f(Data\,|\,\theta)$ and prior $g(\theta)$. The posterior for model $M_k$ with parameter $\theta_k$ is proportional to $f_k(Data\,|\,\theta_k)g_k(\theta_k)\pi_k$, we get posterior probability as

$$P(M_k\,|\,Data)\propto\pi_k\int_{\Omega_k}f_k(Data\,|\,\theta_k)g_k(\theta_k)\mathrm{d}\theta_k$$

$$=\frac{\pi_k\int_{\Omega_k}f_k(Data\,|\,\theta_k)g_k(\theta_k)\mathrm{d}\theta_k}{\sum_{j=1}^{K}\pi_j\int_{\Omega_j}f_j(Data\,|\,\theta_j)g_j(\theta_j)\mathrm{d}\theta_j}$$

In a Bayesian analysis, the priors $\pi_k$ on each model and $g_k(\theta_k)$ on the parameters of model *k* are proper and subjective. And the Bayesian solutions to do questions are to compute the posterior probability $P(M_k\,|\,Data)$ for each model. For model selection, we would choose the model from Bayesian conclusion as maximizes $P(M_k\,|\,Data)$.

However, Bayes factor has its only limitation, that is Bayes factors itself can only show the difference of how hypothesis model is against a null model [5]. Also, Bayes factor has a close connection with priors, if we change the width of the prior, it will also change the Bayes factor. At this point, we may need to consider about Lindley Paradox.

In Section 2, we give a simple and general explanation of Bayes factor. Following, in Section 3, we will talk about Lindley's Paradox. And Section 4 can be one of the main parts of the theoretical approach for AIC and BIC, for which we give the derivation. A simple example is given as well to use AIC and BIC.

## 2. Bayes Factor

Before talking about all things, first we would construct one of the most important variables within Bayesion Methods-Bayes Factor [6].

Suppose we have data $D$ with prior $\theta$ and $M_1,M_2$ as two different models. By Condition Rule, we have:

$$P(M_1\,|\,D)=\frac{P(D\,|\,M_1)P(M_1)}{P(D)}$$

Recall for Odds we have $P(M_1)=1-P(M_2)$. And for $P(D\,|\,M_1)$ is the marginal likelihood, which $P(D\,|\,M_1)=\int P(D\,|\,\theta,M_1)P(\theta\,|\,M_1)\mathrm{d}\theta$. $\theta$ denotes prior. Then, by Bayes' Rule,

$$P(M_1\,|\,D)=\frac{P(D\,|\,M_1)P(M_1)}{P(D\,|\,M_1)P(M_1)+P(D\,|\,M_2)P(M_2)}$$

$$\frac{P(M_1\,|\,D)}{P(M_2\,|\,D)}=\frac{P(D\,|\,M_1)}{P(D\,|\,M_2)}\frac{P(M_1)}{P(M_2)}$$

where $\frac{P(D|M_1)}{P(D|M_2)}$ is defined as Bayes' factor, and realized it is also the ratio of marginal likelihood. Furthermore, we denote Bayes' factor as:

$$B_{1,2}(y)=\frac{P(y|M_1)}{P(y|M_2)}$$

Bayesian method fits in many models for testing because it can provide a decisiveness of the evidence agree the null model in contrast *p*-values [7] which are usually just regarded as evidence mearsurement against the alternative [8]. Also, the Bayes factor (Jefferys, 1961) [9] is used in Bayesian hypothesis. Assuming that $p_i(D|\theta_i), i=1,2$ are the likelihoods for $D$ under two competing models $H_1$ and $H_2$, and the parameters are $\theta_i$. Meanwhile, let $\pi_i(\theta_i)$ be their prior distributions [10]. The Bayes factor for $H_2$ against $H_1$:

$$B_{2,1}=\frac{\int P_2(D|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}{\int P_1(D|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}$$

Above these, evidence from the data agrees $H_2$, against $H_1$. So Bayes factor can avoid many limitations in p-value testing. The development of Bayes factor in statistical models test can applicate in many areas of research [11].

## 3. Priors and Lindley Paradox

### 3.1. Introduction to Lindley Paradox

The Lindley's Paradox shows how a value (or the number of standard deviations) is used in a Frequent Assumption [12] test results in a completely different inference from Bayesian hypothesis [13].

When we faced with improper priors (like priors can't integrate to one) in the null hypothesis and model selection, we will find some problems. Such priors can be acceptable, but for other purposes it is also acceptable. So we consider testing the hypotheses:

$$H_0:\theta\in\Theta_0 \text{ vs } H_1:\theta\in\Theta_1$$

Defining $\theta$ for marginal density, so we can use the following model:

$$p(\theta)=p(\theta|H_0)p(H_0)+p(\theta|H_1)p(H_1)$$

Making $p(\theta|H_0)$ and $p(\theta|H_1)$ are proper density functions, the posterior is given by:

$$\begin{aligned}&p(H_0|D)\\&=\frac{p(H_0)p(D|H_0)}{p(H_0)P(D|H_0)+p(H_1)p(D|H_1)}\\&=\frac{p(H_0)\int_{\Theta_0}p(D|\theta)p(\theta|H_0)\mathrm{d}\theta}{p(H_0)\int_{\Theta_0}p(D|\theta)p(\theta|H_0)\mathrm{d}\theta+p(H_1)\int_{\Theta_1}p(D|\theta)p(\theta|H_1)\mathrm{d}\theta}\end{aligned}$$

Then we can suppose that we use improper priors, making $p(\theta|H_0)\propto z_0$ and $p(\theta|H_1)\propto z_1$. So:

$$p(H_0 \mid D) = \frac{p(H_0) z_0 \int_{\Theta_0} p(D \mid \theta) \mathrm{d}\theta}{p(H_0) z_0 \int_{\Theta_0} p(D \mid \theta) \mathrm{d}\theta + p(H_1) z_1 \int_{\Theta_1} p(D \mid \theta) \mathrm{d}\theta}$$
$$= \frac{p(H_0) z_0 s_0}{p(H_0) z_0 s_0 + p(H_1) z_1 s_1}$$

Establishing model $i$ that $s_i = \int_{\Theta_i} p(x \mid \theta) \mathrm{d}\theta$ is the marginal likelihood or the integrated. So we assume that $p(H_0) = p(H_1) = \frac{1}{2}$

Then an equation can be obtained:

$$p(H_0 \mid D) = \frac{z_0 s_0}{z_0 s_0 + z_1 s_1} = \frac{s_0}{s_0 + \left(\frac{z_1}{z_0}\right) s_1}$$

So we can use different $z$ that we want to change the posterior arbitrarily. Meanwhile, when using proper and not clear priors might cause similar problems. Because the probability of data in a complex model with a diffuse prior will be very small. So one thing we must know, when we do research in Bayes factor a clearer and simper model is better. It was called the Lindley paradox.

### 3.2. A Simple Model in Lindley Paradox

Many authors [14] have discussed this so-called paradox [15] in different ways [16]. So I want to find a simple way to consider this problem. The usual point null hypothesis testing problem is to test:

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

In normal model $N(x \mid \theta, 1)$. The prior probability is $p(H_0)\left(p(H_0) = \rho_0\right)$.

Let $\pi(\theta) = N\left(\theta \mid 0, \sigma^2\right)(\sigma > 0)$ be the prior distribution for the unknowm parameter $\theta$ in the model.

The Bayes factor is given by:

$$B = \frac{N(x \mid 0,1)}{\int N(x \mid \theta, 1) \pi(\theta) \mathrm{d}\theta}$$

In order to consider the paradox, we can formalise it and compare the two following normal models:

$$M_0 = \left\{ N(x \mid 0,1) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^2\right) \right\}$$

$$M_1 = \left\{ N(x \mid \theta, 1) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\} \right\}$$

Consider a physical system where quantity X may be measured and assume. And we need to use the $\sigma$ to define both the priors. The prior of the null hypothesis is $\rho_0$ supposing the $\rho_0$ can depend on $\sigma$.

Computing the Bayes factor representing the odds of the null hypothesis $H_0$ is:

$$B_0 = \frac{N(x|0,1)}{\int N(x|\theta,1)N(\theta|0,\sigma^2)\mathrm{d}\theta} = \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{\mathrm{e}^{\frac{-\frac{1}{2}x^2}{\sigma^2+1}}}\sqrt{\sigma^2+1}$$

In this case, prior probabilities $p(H_0)$ and $p(H_1)=1-P(H_0)$ for two hypotheses can be expressed. Given the result x, in Bayes theory that:

$$p(H_m|x)p(x) = p(x|H_m)p(H_m)$$

for $m=0,1$, $p(H_m)$ is prior probabilities and $p(x|H_m)$ is the conditional distribution, $p(x) = p(x|H_0)P(H_0) + P(x|H_1)p(H_1)$ can outcome the overall distribution. Posterior probability $p(H_m|x)$ is in the hypothesis $H_m$. In Bayes theory we can evaluate the posterior probabilities, $p(H_0|x)$ is given by:

$$\begin{aligned} p(H_0|x) &= \frac{p(x|H_0)p(H_0)}{p(x)} \\ &= \frac{p(x|H_0)p(H_0)}{p(x|H_0)p(H_0)+p(x|H_1)p(H_1)} \\ &= \left[1+\frac{p(x|H_1)p(H_1)}{p(x|H_0)p(H_0)}\right]^{-1} \\ &= \left[1+\frac{1-p(H_0)}{p(H_0)}\frac{p(x|H_1)}{p(x|H_0)}\right]^{-1} \end{aligned}$$

Then, we can use the mean value in prior distribution with $\pi_m(\theta) \equiv p(\theta|H_m)$ and make the rest of the prior probability as a normal distribution with variance $\tau$, so:

$$\pi_1(\theta) = \left(2\pi\tau^2\right)^{-\frac{1}{2}}\exp\left\{\frac{-\theta^2}{2\tau^2}\right\}$$

Evaluating the conditional probabilities:

$$p(x|H_m) = \int \pi_m(\theta)p(x|\theta)\mathrm{d}\theta$$

We can evaluate $p(x|H_0)$ and $p(x|H_1)$, overall:

$$p(H_0|x) = \left[1+\frac{1-\rho_0}{\rho_0}\frac{\mathrm{e}^{-\frac{1}{2}x^2/(\sigma^2+1)}}{\mathrm{e}^{-\frac{1}{2}x^2}}\frac{1}{\sqrt{\sigma^2+1}}\right]^{-1} = \left[1+\frac{1-\rho_0}{\rho_0}\frac{1}{B_0}\right]^{-1}$$

So we have an equation like before, we can talk about the prior $\rho(\sigma)$. Our approach is to measure the value of alternative assumptions about zero. In Asymptotically Bayesian attribute, if the model is incorrectly specified, the posterior will accumulates in the model. In the case of the Kullback-Leibler divergence, the closest to the real model [17]. As a result, divergence $D_K L(N|x(\theta,1)\|N(x|0,1))$ represents the loss. Because we know the prior before. The excepted loss can be given:

$$\int_\Theta D_K L(N(x\theta,1)\|N(x|0,1))\pi(\theta)\mathrm{d}\theta = \int\frac{1}{2}\theta^2\pi(\theta)\mathrm{d}\theta = \frac{1}{2}\sigma^2$$

The model prior represent the loss relatied with a probability statement, it also determined self-information loss function. So we have the prior on the alternative model is:

$$1-\rho_0(\sigma) \propto e^{\sigma^2}$$

The prior of the null hypothesis is $\rho(\sigma) \propto 1$, then we can get:

$$\rho_0(\sigma) = \frac{1}{1+\exp\left\{\frac{1}{2}\sigma^2\right\}}$$

Then, this applies to the category of large $\sigma$ and $p(H_0 \mid x, \sigma)$ goes to zero, so $p(H_0) \to 0$. Therefore, this method is consistent, we do not advocate the choice of big $\sigma$.

## 4. BIC and AIC

### 4.1. BIC

#### 4.1.1. Notation (Table 1)

**Table 1.** Notation 1.

| | |
|---|---|
| $y$ | observed data $y_1, \cdots, y_n$ |
| $M_i$ | candidate model |
| $\theta_i$ | vector of parameters in the model |
| $g(\theta_i)$ | the prior density of the parameters $\theta_i$ |
| $P(y \mid M_i)$ | marginal likelihood |
| $f(y \mid \theta_i)$ | the density of the data given |
| $L(\theta_i \mid y)$ | the likelihood of $y$ given the model $M_i$ |

#### 4.1.2. Derivation of BIC

In this section we are going to talk about the basic idea [18] of how BIC (Bayesian information criterion) constructed and given the derivation of BIC [4].

As what we have showed in section one, $B_{1,2}(y) = \frac{P(y \mid M_1)}{P(y \mid M_2)}$ as Bayes factor for two models, then we consider more models $M_i$ which $i \in \{1, \cdots, n\}$

$$P(y \mid M_i) = \int f(y \mid \theta_i) = \int \exp\left(\log\left(f(y \mid \theta_i) g_i(\theta_i)\right)\right) \mathrm{d}\theta_i$$

$f(y \mid \theta_i) g_i(\theta_i)$ where $\theta_i$ is the vector of parameters in the model $M_i$, $L$ is the likelihood function and $g_i(\theta_i)$ is the p.d.f. of the distribution of parameters $\theta_i$

Denoting $\tilde{\theta}_i$ as the posterior mode, then we use Taylor expansion, let $Q(\theta_i) = \log\left(f(y \mid \theta_i) g_i(\theta_i)\right)$,

$$Q(\theta_i) = \log\left(f(y \mid \theta_i) g_i(\theta_i)\right)$$

$$\approx \log\left(f(y \mid \tilde{\theta}_i)\right) g_i(\tilde{\theta}_i) + (\theta_i - \tilde{\theta}_i) \nabla_{\theta_i} Q\big|_{\tilde{\theta}_i} + \frac{1}{2}(\theta_i - \tilde{\theta}_i)^{\mathrm{T}} H_{\theta_i}(\theta_i - \tilde{\theta}_i)$$

where $H_{\theta_i}$ is a $|\theta_i||\theta_i|$ matrix such that $H_{mn}=\left.\frac{\partial^2 Q}{\partial\theta_m\partial\theta_n}\right|_{\tilde{\theta}_i}$, where $|\theta_i|=d_i=dimension(\theta_i)$. since $Q$ attains its maximum, the Hessian matrix $H_{\theta_i}$ is negative definite. Let us denote $\bar{H}_{\theta_i}=-H_{\theta_i}$, and then approximate $P(y\mid M_i)$:

$$P(y\mid M_i)\approx\int\exp\left\{Q|_{\tilde{\theta}_i}+(\theta_i-\tilde{\theta}_i)\nabla_{\theta_i}Q|_{\tilde{\theta}_i}+\frac{1}{2}(\theta_i-\tilde{\theta}_i)^{\mathrm{T}}H_{\theta_i}(\theta_i-\tilde{\theta}_i)\right\}\mathrm{d}\theta_i$$

Then, by higher dimension normal distribution,

$$\because\int\frac{1}{(2\pi)^{\frac{d_i}{2}}\left|\tilde{H}_{\theta_i}^{-1}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\theta_i-\hat{\theta}_i)^{\mathrm{T}}\vec{H}_{\theta_i}(\theta_i-\hat{\theta}_i)\right)\mathrm{d}\theta_i=1$$

$$\Rightarrow\int\exp\left(-\frac{1}{2}(\theta_i-\hat{\theta}_i)^{\mathrm{T}}\vec{H}_{\theta_i}(\theta_i-\hat{\theta}_i)\right)\mathrm{d}\theta_i=(2\pi)^{\frac{d_i}{2}}\left|\tilde{H}_{\theta_i}^{-1}\right|^{\frac{1}{2}}$$

$$\Rightarrow P(y\mid M_i)=f(y\mid\tilde{\theta}_i)g_i(\tilde{\theta}_i)(2\pi)^{\frac{d_i}{2}}\left|\tilde{H}_{\theta_i}^{-1}\right|^{\frac{1}{2}}$$

$$\Rightarrow\log P(y\mid M_i)=\log f(y\mid\tilde{\theta}_i)+\log g_i(\tilde{\theta}_i)+\frac{d_i}{2}\log(2\pi)+\frac{1}{2}\log\left|\tilde{H}_{\theta_i}^{-1}\right|$$

Furthermore, let us think about Weak Law of Large Numbers. For $y$ is given data, $f(y\mid\theta_i)$ is the likelihood $L(\theta_i\mid y)$ and $L$ attains its maximum at the maximum likelihood estimate $\theta_i-\hat{\theta}_i$.

We set $g_i(\theta_i)=\begin{cases}1, & \theta\in\left[\tilde{\theta}_i-\frac{1}{2},\tilde{\theta}_i+\frac{1}{2}\right]\\ 0, & \text{else}\end{cases}$, then each element in the matrix, $\tilde{H}_{\theta_i}$, can be expressed as:

$$\tilde{H}_{mn}=-\left.\frac{\partial^2\log L(\theta_i\mid y)}{\partial\theta_m\partial\theta_n}\right|_{\theta_i=\hat{\theta}_i}$$

Then, for $\tilde{H}_{\theta_i}$ as a Fisher information matrix that,

$$\begin{aligned}\tilde{H}_{mn}&=-\left.\frac{\partial^2\log\left(\prod_{j=1}^{n}L(\theta_i\mid y_j)\right)}{\partial\theta_m\partial\theta_n}\right|_{\theta_i=\hat{\theta}_i}\\&=-\left.\frac{\partial^2\sum_{j=1}^{n}\log L(\theta_i\mid y_j)}{\partial\theta_m\partial\theta_n}\right|_{\theta_i=\hat{\theta}_i}\\&=-\left.\frac{\partial^2\left(\frac{1}{n}\sum_{j=1}^{n}n\log L(\theta_i\mid y_j)\right)}{\partial\theta_m\partial\theta_n}\right|_{\theta_i=\hat{\theta}_i}\end{aligned}$$

In this case, for the data $y_1,\cdots,y_n$ is IID, and $n$ is large, we would apply Weak Law of Large number here, as random variable $X_j=n\log L(\theta_i\mid y_j)$ we have $\frac{1}{n}\sum_{j=1}^{n}n\log L(\theta_i\mid y_j)\xrightarrow{P}E\left(n\log L(\theta_i\mid y_j)\right)$, Moreover, for Fisher in-

formation matrix:

$$\begin{aligned}\tilde{H}_{mn} &= -\frac{\partial^2 E\left[n\log L\left(\theta_i \mid y_j\right)\right]}{\partial\theta_m\partial\theta_n}\Bigg|_{\theta_i=\hat{\theta}_i} \\ &= -n\frac{\partial^2 E\left[\log L\left(\theta_i \mid y_j\right)\right]}{\partial\theta_m\partial\theta_n}\Bigg|_{\theta_i=\hat{\theta}_i} \\ &= -n\frac{\partial^2 E\left[\log L\left(\theta_i \mid y_1\right)\right]}{\partial\theta_m\partial\theta_n}\Bigg|_{\theta_i=\hat{\theta}_i} \\ &= nI_{mn} \\ &\Rightarrow \left|\tilde{H}_{\theta_i}\right| = n^{|\theta_i|} \mid I_{\theta_i}\end{aligned}$$

For which $I_{\theta_i}$ is the Fisher information matrix for a single data point $y_1$, and after substituting we final get for BIC:

$$2\log P\left(\mathbf{y} \mid M_i\right) = 2\log L\left(\hat{\theta}_i \mid \mathbf{y}\right) + 2\log g_i\left(\tilde{\theta}_i\right) + \left|\theta_i\right|\log\left(2\pi\right) - \left|\theta_i\right|\log n - \log\left|I_{\theta_i}\right|$$

## 4.2. AIC

### 4.2.1. Notation (Table 2)

**Table 2.** Notation 2.

| | |
|---|---|
| $M_j = \left\{P\left(y \mid \theta_j\right) : \theta_j \in \Theta_j\right\}$, | Different models(each is a set of density) |
| $K\left(x, y\right)$ | The Kullback-Leibler distance between *x*, *y* |
| $\ell_j\left(\theta_j\right)$ | the log-likelihood function for model *j* |
| $\hat{P}_j\left(y\right) = P\left(y \mid \hat{\theta}_j\right)$ | An estimate of *P* based on model *j* |
| $d_j$ | The dimension of $\Theta_j$ |
| $Y_j$ | The Data drawn from density *P* |
| $\hat{\theta}_j$ | The MLE of model *j* |
| $s\left(y \mid \theta_j\right) = \frac{\partial \log P\left(y \mid \theta_j\right)}{\partial \theta_j}$ | The Jaccobi Matrix of log $P\left(y \mid \theta_j\right)$ |

### 4.2.2. Derivation of AIC

We can measure the quality of $\hat{p}_j\left(y\right)$ (as an estimate of $p$) by the Kullback-Leibler distance [19]:

$$\begin{aligned}K\left(p, \hat{p}_j\right) &= \int p\left(y\right)\log\left(\frac{p\left(y\right)}{\hat{p}_j\left(y\right)}\right)\mathrm{d}y \\ &= \int p\left(y\right)\log p\left(y\right)\mathrm{d}y - \int p\left(y\right)\log \hat{p}_j\left(y\right)\mathrm{d}y\end{aligned}$$

So, we want to minimize $K\left(p, \hat{p}_j\right)$ over *j*, which is the same as maximizing

$$K_j = \int p\left(y\right)\log p\left(y \mid \hat{\theta}_j\right)\mathrm{d}y$$

For calculating $K_j$, we can use Monte Carlo method to do an estimate

$$\bar{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log p\left(Y_i \mid \hat{\theta}_j\right) = \frac{\ell_j\left(\hat{\theta}_j\right)}{n}$$

However, this estimate is very biased because the data are being used twice: first to get the MLE and second to estimate the integral by Monte Carlo method, and the bias is approximayely $\frac{d_j}{n}$. That means we should prove [20]

$$\bar{K}_j - \frac{d}{n} \approx K_j$$

Choose $\theta_{j_0}$, s.t. $p\left(y \mid \theta_{j_0}\right) = \max_{\theta_j \in \Theta_j} p\left(y \mid \theta_j\right)$, and let

$$s\left(y, \theta_j\right) = \frac{\partial \log p\left(y \mid \theta_j\right)}{\partial \theta_j}, \quad H\left(y, \theta_j\right) = \frac{\partial^2 \log p\left(y \mid \theta_j\right)}{\partial \theta_j^2}$$

So, $s\left(y, \theta_j\right)$ is the Jacobi martix of $\log p\left(y \mid \theta_j\right)$, and $H\left(y, \theta_j\right)$ is the Hessian martix of $\log p\left(y \mid \theta_j\right)$.

$$\begin{aligned} \Rightarrow K_j &\approx \int p(y)\Big(\log p\left(y \mid \theta_{j_0}\right) + \left(\hat{\theta} - \theta_{j_0}\right)^{\mathrm{T}} s\left(y, \theta_{j_0}\right) \\ &\quad + \frac{1}{2}\left(\hat{\theta} - \theta_{j_0}\right)^{\mathrm{T}} H\left(y, \theta_{j_0}\right)\left(\hat{\theta} - \theta_{j_0}\right)\Big)\mathrm{d}y \\ &= K_0 - \frac{1}{2n} Z_n^{\mathrm{T}} J Z_n \end{aligned}$$

where

$$K_0 = \int p(y) \log p\left(y \mid \theta_{j_0}\right)\mathrm{d}y, \quad Z_n = \sqrt{n}\left(\hat{\theta}_j - \theta_{j_0}\right), \quad J = -E\left[H\left(y, \theta_{j_0}\right)\right].$$

$$\begin{aligned} \Rightarrow \bar{K}_j &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(Y_i, \theta_{j_0}\right) + \left(\hat{\theta} - \theta_{j_0}\right)^{\mathrm{T}} s\left(Y_i, \theta_{j_0}\right) + \frac{1}{2}\left(\hat{\theta} - \theta_{j_0}\right)^{\mathrm{T}} H\left(Y_i, \theta_{j_0}\right)\left(\hat{\theta} - \theta_{j_0}\right)\right) \\ &= K_0 + A_n + \left(\hat{\theta} - \theta_0\right)^{\mathrm{T}} S_n - \frac{1}{2n} Z_n^{\mathrm{T}} J_n Z_n \\ &= K_0 + A_n + \frac{Z_n^{\mathrm{T}} S_n}{\sqrt{n}} - \frac{1}{2n} Z_n^{\mathrm{T}} J Z_n \end{aligned}$$

where

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(Y_i, \theta_0\right) - K_0\right), \quad S_n = \frac{1}{n}\sum_{i=1}^{n} s\left(Y_i, \theta_0\right)$$

and

$$J_n = -\frac{1}{n}\sum_{i=1}^{n} H\left(Y_i, \theta_0\right) \xrightarrow{P} J$$

$$\bar{K}_j - K_j \approx A_n + \frac{\sqrt{n} Z_n^{\mathrm{T}} S_n}{n} \approx A_n + \frac{Z_n^{\mathrm{T}} J Z_n}{n}$$

From the knowledge of asymptotic distribution, we have three claims [21]:

**Claim 4.1** $Z_n = \sqrt{n}\left(\hat{\theta} - \theta_{j_0}\right) \to N\left(0, J^{-1} V J^{-1}\right)$, *where* $V = Var\left(s\left(Y, \theta_{j_0}\right)\right) = J^{-1}$.

**Claim 4.2** $\sqrt{n}S_n = \frac{\sqrt{n}}{n}\sum_{i=1}^{n} s\left(Y_i, \theta_{j_0}\right) \to N\left(0, V\right)$

**Claim 4.3** *Let* $\epsilon$ *be a random vector with mean* $\mu$ *and covariance* $\Sigma$, *and* $Q = \epsilon^{\mathrm{T}} A\epsilon$, *then,*

$$E\left(Q\right) = trace\left(A\Sigma\right) + \mu^{\mathrm{T}} A\mu$$

So, with these calims above,

$$\Rightarrow E\left(\bar{K} - K\right) \approx E\left(A_n\right) + E\left(\frac{Z_n^{\mathrm{T}} J Z_n}{n}\right) = 0 + \frac{E\left(Z_n^{\mathrm{T}} J Z_n\right)}{n}$$

$$= \frac{trace\left(JJ^{-1}VJ^{-1}\right)}{n} + \mathbf{0}^{\mathrm{T}} J\mathbf{0}$$

$$= \frac{trace\left(VJ^{-1}\right)}{n} = \frac{trace\left(I\right)}{n} = \frac{d_j}{n}$$

$$\Rightarrow \hat{K}_j = \frac{\ell_j\left(\hat{\theta}_j\right)}{n} - \frac{d_j}{n} = \bar{K}_j - \frac{d_j}{n}$$

So, we define

$$AIC\left(j\right) = 2n\hat{K}_j = 2\ell_j\left(\hat{\theta}_j\right) - 2d_j$$

### 4.3. Example of Simple Model

Let us consider again with the example in section 3, if we take data $Y_1, \cdots, Y_n \sim N\left(\theta, 1\right)$, and compare it with two models, such that, $M_0 : N\left(0,1\right)$ and $M_1 : N\left(\theta, 1\right)$. Then take the same hypothesis as in section 3.2, we test:

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

By standard normal distribution we have,

$$Z = \frac{\bar{Y} - 0}{\sqrt{Var\left(\bar{Y}\right)}} = \sqrt{n}\bar{Y}$$

In case to avoid Type I error in our test, for $\alpha = 0.05$, by $Z$ table, we would reject $H_0$ if $\left|Z\right| > \frac{z_\alpha}{2} \approx 1.96$ (we take $\left|Z\right| > 2$). Which implies if $\left|\bar{Y}\right| > \frac{2}{\sqrt{n}}$, we reflect $H_0$.

**Case 1:** ***BIC***

For what we have showed in section 4.1, we proved that $BIC = 2\log L\left(\hat{\theta}_i \mid \mathbf{y}\right) + 2\log g_i\left(\tilde{\theta}_i\right) + \left|\theta_i\right|\log\left(2\pi\right) - \left|\theta_i\right|\log n - \log\left|I_{\theta_i}\right|$. However, in case to make comparison with two models, we could get away some unnecessary part, we take $BIC = \log L\left(\hat{\theta}_i \mid y\right) - \frac{\left|\theta_i\right|}{2}\log n$. Thus,

For $H_0$,

$$BIC = \log L\left(0\right) - \frac{0}{2}\log n = -\frac{n\bar{Y}^2}{2} - \frac{nS^2}{2}$$

and $H_1$,

$$BIC = \log L\left(\hat{\theta}\right) - \frac{1}{2}\log n = -\frac{nS^2}{2} - \frac{1}{2}\log n$$

where $S^2 = \sum_i \left(Y_i - \bar{Y}\right)^2$. If we want to choose $M_1$ as a better model, then we would make $-\frac{n\bar{Y}^2}{2} - \frac{nS^2}{2} < -\frac{nS^2}{2} - \frac{1}{2}\log n$, in other words, $\left|\bar{Y}\right| > \sqrt{\frac{\log n}{n}}$. And BIC is an estimate of a function of the posterior probability of a model under Bayesian setup.

**Case 2: *AIC***

And from section 4.2, for $AIC = 2\ell_j\left(\hat{\theta}_j\right) - 2d_j$, for which as what we have defined above $S^2 = \sum_i \left(Y_i - \bar{Y}\right)^2$ that $\ell(\theta) = -\frac{n\left(\bar{Y} - \theta\right)^2}{2} - \frac{nS^2}{2}$. Further deduce $AIC = \ell_S - |S|$ Thus,

For $H_0$,

$$AIC = \ell(0) - 0 = -\frac{n\bar{Y}^2}{2} - \frac{nS^2}{2}$$

and $H_1$,

$$AIC = \ell(\theta) - 1 = -\frac{nS^2}{2} - 1$$

If we want to choose $M_1$ as a better model at this point, we would take $-\frac{n\bar{Y}^2}{2} - \frac{nS^2}{2} < \frac{nS^2}{2} - 1$, implies $\left|\bar{Y}\right| > \frac{\sqrt{2}}{\sqrt{n}}$. Which AIC is estimate a constant plus the relative distance between unknow likelihood function.

## 5. Conclusion

The question of how to choose a best model and what is a best model, it is hard to define. More precise, the controversy has existed for a long time, and no doubt it will continue longer. In this paper, we have discussed Bayes factor in hypothesis. It is obviously that Bayes factor is increasingly used in many fields of statistic research. For Bayes factor standard methods, AIC and BIC, we would consider to use for model selection. However, we also should notice that for all methods they all have their own limitation, such as the sensitivity of priors in Lindley's paradox. Even both frequentist and Bayesian statisticians have came up with different new ideas, it is still hard to be implemented or understand by all other. Moreover, from statistic point, the method also needs to be general enough to apply. Such as for Lindley's paradox, the partial Bayes factor in case to avoid the sensitive of priors, it takes the minimal training sample from data set to get prior and then apply with rest of the data. Partial Bayes factor at some point did deduce the influence of sensitivity of prior, but how to find the minimal training sample could also be a hard problem. Same as fractional Beyes factor, even it proves the method of choosing data for partial Bayes facto, it still has many limitations we need consider.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References


[1] Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC. *Statistics and Computing*, **27**, 1413-1432.

[2] Hartman, B.M. and Groendyke, C. (2013) Model Selection and Averaging in Financial Risk Management. *North American Actuarial Journal*, **17**, 216-228. https://doi.org/10.1080/10920277.2013.824374

[3] Bayarri, M.J., Berger, J.O., Forte, A. and Garca-Donato, G. (2012) Criteria for Bayesian Model Choice with Application to Variable Selection. *The Annals of Statistics*, **40**, 1550-1577. https://doi.org/10.1214/12-AOS1013

[4] Wasserman, L. (2000) Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92-107. https://doi.org/10.1006/jmps.1999.1278

[5] Ardila, F., *et al.* (2008) Root Polytopes and Growth Series of Root Lattices. *SIAM Journal on Discrete Mathematics*, **25**, 1-17.

[6] Kass, R.E. and Raftery, A.E. (1999) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.

[7] Bayarri, M.J. and Berger, J.O. (2004) The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, **19**, 58-80.

[8] Spezzaferri, F., Verdinelli, I. and Zeppieri, M. (2007) Bayes Factors for Goodness of Fit Testing. *Journal of Statistical Planning and Inference*, **137**, 43-56. https://doi.org/10.1016/j.jspi.2005.09.002

[9] Jeffery, H. (1961) Theory of Probability.

[10] Bhat, H.S. and Kumar, N. (2010) On the Derivation of the Bayesian Information Criterion.

[11] Kadane, J.B. and Lazar, N.A. (2004) Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, **99**, 279-290.

[12] Merhav, N. and Feder, M. (1998). Universal Prediction. *IEEE Transactions on Information Theory*, **44**, 2124-2147. https://doi.org/10.1109/18.720534

[13] Lindley, D.V. (1957) A Statistical Paradox. *Biometrika*, **44**, 187-192. https://doi.org/10.1093/biomet/44.1-2.187

[14] Bartlett, M.S. (1957) A Comment on D. V. Lindley's Statistical Paradox. *Biometrika*, **44**, 533-534. https://doi.org/10.2307/2332888

[15] Robert, C.P. (1993) A Note on Jeffreys-Lindley Paradox. *Statistica Sinica*, **3**, 601-608.

[16] Villa, C. and Walker, S. (2015) On the Mathematics of the Jeffreys-Lindley Paradox. *Communications in Statistics—Theory and Methods*, **46**, 12290-12298.

[17] Cousins, R.D. (2014) The Jeffreys-Lindley Paradox Discovery Criteria in High Energy Physics. *Synthese*, **194**, 395-432. https://doi.org/10.1007/s11229-015-0687-3

[18] Shively, T. and Walker, S.G. (2013) On the Equivalence between Bayesian and Classical Hypothesis Testing.

[19] Vardeman, S.B. (1987) Testing a Point Null Hypothesis: The Irreconcilability of p-Values and Evidence Comment. *Journal of the American Statistical Association*,

**82**, 130-131. http://www.jstor.org/stable/2289136
https://doi.org/10.2307/2289136

[20] Penny, W.D. (2012) Comparing Dynamic Causal Models Using AIC, BIC and Free Energy. *Neurolmage*, **59**, 319-330. https://doi.org/10.1016/j.neuroimage.2011.07.039

[21] Zellner, A. and Siow, A. (1980) Posterior Odds Ratios for Selected Regression Hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, **31**, 585-603. https://doi.org/10.1007/BF02888369