Marginal Probability. Its use in Bayesian Statistics as the Evidence of Models and Bayes Factors

Luis Raúl Pericchi, Department of Mathematics and Biostatistics and Bioinformatics Ce
University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico.*

Keywords:*Bayes Factors, Evidence of Models, Intrinsic Bayes Factors, Intrinsic Priors, Posterior Model Probabiities*

1 Definition

Suppose that we have vectors of random variables $[\mathbf{v,w}] = [v_1, v_2, \ldots, v_I, w_1, \ldots, w_J]$ in $\Re^{(I+J)}$. Denote as the **joint** density function: $f_{\mathbf{v,w}}$, which obeys:$f_{\mathbf{v,w}}(v, w) \geq 0$ and
$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{v,w}}(v, w) dv_1 \ldots dv_I dw_1 \ldots dw_I = 1$. Then the probability of the set $[A_v, B_w]$ is given by

$$P(A_v, B_w) = \int \ldots \int_{A_v, B_w} f_{\mathbf{v,w}}(v, w) \mathbf{dvdw}.$$

The the **marginal** density $f_{\mathbf{v}}$ is obtained as

$$f_{\mathbf{v}}(v) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{v,w}}(v, w) dw_1 \ldots dw_I.$$

The the **marginal probability** of the set A_v is then obtained as,

$$P(A_v) = \int \ldots \int_{A_v} f_{\mathbf{v}}(v) dv.$$

We have assumed that the random variables are continuous. When they are discrete, integrals are substituted by sums.
We proceed to present an important application of marginal densities to construct the *Evidence of the Model* and marginal probabilities for measuring the *Bayesian Probability of a Model*.

*email address: luarpr@uprrp.edu,This work sponsored in part by NIH Grant: P20-RR016470

2 Measuring the Evidence in Favor of a Model

In Statistics, a parametric model, is denoted as $f(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k)$, where $\mathbf{x} = (x_1, \ldots, x_n)$ is the vector of n observations and $\theta = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_k})$ is the vector of k parameters. For instance we may have n observations normally distributed and the vector of parameters is (θ_1, θ_2) the location and scale respectively, denoted by $f_{Normal}(\mathbf{x}|\theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi}\boldsymbol{\theta_2}} \exp(-\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\theta_2^2}}(\mathbf{x_i} - \boldsymbol{\theta_1})^{\mathbf{2}})$.
Assume now that there is reason to suspect that the location is zero. As a second example, it may be suspected that the sampling model which usually has been assumed Normally distributed, is instead a Cauchy, $f_{Cauchy}(\mathbf{x}|\theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{\mathbf{1}}{\pi\boldsymbol{\theta_2}} \frac{\mathbf{1}}{(\mathbf{1}+(\frac{\mathbf{x_i}-\boldsymbol{\theta_1}}{\boldsymbol{\theta_2}})^{\mathbf{2}})}$. The first problem is a *hypothesis test* denoted by

$$H_0 : \theta_1 = 0 \text{ VS } H_1 : \theta_1 \neq 0,$$

and the second problem is a *model selection* problem:

$$M_0 : f_{Normal} \text{ VS } M_1 : f_{Cauchy}.$$

How to measure the evidence in favor of H_0 or M_0? Instead of maximizing likelihoods as it is done in traditional significance testing, in Bayesian statistics the central concept is *the evidence* or *marginal probability density*

$$m_j(\mathbf{x}) = \int f_j(\mathbf{x}|\boldsymbol{\theta_j})\pi(\boldsymbol{\theta_j})\mathbf{d}\boldsymbol{\theta_j},$$

where j denotes either model or hypothesis j and $\pi(\boldsymbol{\theta_j})$ denotes the prior for the parameters under model or hypothesis j.
Marginal probabilities embodies the likelihood of a model or hypothesis in great generality and can be claimed it is the natural probabilistic quantity to compare models.

3 Marginal Probability of a Model

Once the marginal densities of the model j, for $j = 1, \ldots, J$ models have been calculated and assuming the prior model probabilities $P(M_j), j = 1, \ldots, J$ with $\sum_{j=1}^{J} P(M_j) = 1$ then, using Bayes Theorem, *the marginal probability of a model* $P(M_j|\mathbf{x})$ can be calculated as,

$$P(M_j|\mathbf{x}) = \frac{\mathbf{m_j}(\mathbf{x}) \cdot \mathbf{P}(\mathbf{M_j})}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{m_i}(\mathbf{x}) \cdot \mathbf{P}(\mathbf{M_i})}.$$

We have then the following formula for any two models or hypotheses:

$$\frac{P(M_j|\mathbf{x})}{P(M_i|\mathbf{x})} = \frac{P(M_j)}{P(M_i)} \times \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

or in words: Posterior Odds equals Prior Odds times Bayes Factor, where the Bayes Factor of M_j over M_i is

$$B_{j,i} = \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

Jeffreys (1961).
In contrast to *p-values*, which have interpretations heavily dependent on the sample size n, and its definition is not the same as the scientific question, the posterior probabilities and Bayes Factors address the scientific question: "how probable is model or hypothesis j as compared with model or hypothesis i?", and the interpretation is the same for any sample size, Berger and Pericchi (1996a, 2001). Bayes Factors and Marginal Posterior Model Probabilities have several advantages, like for example large sample consistency, that is as the sample size grows the Posterior Model Probability of the sampling model tends to one. Furthermore, if the goal is to predict future observations y_f it is **not** necessary to select one model as *the* predicting model since we may predict by the so called Bayesian Model Averaging, which if quadratic loss is assumed, the optimal predictor takes the form,

$$E[Y_f|\mathbf{x}] = \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{E}[\mathbf{Y_f}|\mathbf{x}, \mathbf{M_j}] \times \mathbf{P}(\mathbf{M_j}|\mathbf{x}),$$

where $E[Y_f|\mathbf{x}, \mathbf{M_j}]$ is the expected value of a future observation under the model or hypothesis M_j.

4 Intrinsic Priors for Model Selection and Hypothesis Testing

Having said some of the advantages of the marginal probabilities of models, the question arises: how to assign the conditional priors $\pi(\theta_j)$? In the two examples above which priors are sensible to use? The problem is **not** a simple one since it is not possible to use the usual Uniform priors since then the Bayes Factors are undetermined. To solve this problem with some generality, Berger and Pericchi (1996a,b) introduced the concepts of Intrinsic Bayes Factors and Intrinsic Priors. Start by splitting the sample in two sub-samples $\mathbf{x} = [\mathbf{x}(\mathbf{l}), \mathbf{x}(-\mathbf{l})]$ where the training sample $\mathbf{x}(\mathbf{l})$ is as small as possible such that for $j = 1, \ldots, J : 0 < m_j(\mathbf{x}(\mathbf{l})) < \infty$. Thus starting with an improper prior $\pi^N(\theta_j)$, which does not integrate to one (for example the Uniform), by using the minimal training sample $\mathbf{x}(\mathbf{l})$, all the conditional prior densities $\pi(\theta_j|\mathbf{x}(\mathbf{l}))$ **become** proper. So we may form the Bayes Factor using the training sample $\mathbf{x}(\mathbf{l})$ as

$$B_{ji}(\mathbf{x}(\mathbf{l})) = \frac{\mathbf{m_j}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}{\mathbf{m_i}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}.$$

This however depends on the particular training sample $\mathbf{x}(\mathbf{l})$. So some sort of average of Bayes Factor is necessary. In Berger and Pericchi (1996) it is shown that the average should be the arithmetic average. It is also found a theoretical prior that is an approximation to the procedure just described as the sample size grows. This is called an *Intrinsic Prior*. In the examples above: **Example 1**: in the normal case, assuming first that the variance is known $\theta_2^2 = \theta_{2,0}^2$ then it turns out that the Intrinsic Prior is Normal centered at the null hypothesis $\theta_1 = 0$ and with variance $2 \cdot \theta_{2,0}^2$. More generally when the variance is unknown

$$\pi^I(\theta_1|\theta_2) = \frac{1 - \exp(-\theta_1^2/\theta_2^2)}{2\sqrt{\pi} \cdot (\theta_1^2/\theta_2)}, \text{ and } \pi^I(\theta_2) = \frac{1}{\theta_2}.$$

It turns out that $\pi^I(\theta_1|\theta_2)$ is a proper density, Berger and Pericchi (1996ab), Pericchi(2005).
Example 2: in the Normal vs Cauchy example, it turns out that the improper prior $\pi^I(\theta_1, \theta_2) = 1/\theta_2$ is the appropriate prior for comparing the models, Pericchi (2005). For other examples of Intrinsic Priors see for instance, Berger and Pericchi (1996a, 1996b, 2001), Moreno, Bertolino and Racugno (1998), Pericchi (2005) and Casella and Moreno (2009), among others.
This article is based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC

5 References

Berger J.O. and Pericchi L.R. (1996a). The Intrinsic Bayes Factor for Model Selection and Prediction. *Jour. Amer. Stat. Assoc.*,**91**, p. 109-122.
Berger J.O. and Pericchi L.R. (1996b). The Intrinsic Bayes Factors for Linear Models. In *Bayesian Statistics 5*, Bernardo J.M. et. al, editors, p. 23-42, Oxford University Press.
Berger J.O. and Pericchi L.R. (2001) Objective Bayesian Methods for Model Selection: Introduction and Comparison. *IMS LectureNotes-Monograph Series*, **38**, p. 135-207.
Casella, G. and Moreno, E. (2009) Assessing robustness of intrinsic tests of independence in two-way contingency tables. Journal of the American Statistical Association, 104, 1261-1271.
Jeffreys, H. (1961) *Theory of Probability*. 3rd Ed. Oxford University Press.
Moreno E., Bertolino F. and Racugno W. (1998) An Intrinsic Limiting Procedure for Model Selection and Hypothesis Testing. *Jour. of the Amer Statist Assoc*, **93**, 444, pp. 1451-1460.
Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC
Pericchi, L.R. (2005)Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors. *Handbook of Statistics*, Vol. 25: Bayesian Thinking: Modeling and Computation. Dey D.K. and Rao C.R. Editors. *Elsevier*, North-Holland.