RESEARCH ARTICLE OPEN ACCESS

A new Bayesian information criterion for the linear regression model: case of maximum a posteriori estimator

Théophile RABENANTENAINA*, Parfait BEMARISIKA**, André TOTOHASINA**

**(Thematic Doctoral School “Science, Culture, Society and Development”, of the University of Toamasina-Madagascar)*

*** (Department of Mathematics and computer science, Ecole Normale Supérieure pour l’Enseignement Technique of the University of Antsiranana-Madagascar)*

ABSTRACT

This article is concerned with the Bayesian selection of linear regression models. In order to achieve this objective, we will successively present the estimators obtained by maximum a posteriori in the two cases of a priori laws of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma^2}$: cases informative and non-informative one. In the non-informative case, we put forth the use of Jeffreys’ a priori law, which is based on Fisher’s information. In the informative case, we examine the a priori joint distribution of these two parameters, which follow a normal-gamma distribution. We then present some properties of these estimators for both cases. Based on these properties, we propose a new model selection criterion based on the calculation of the Kullback-Leibler divergence between the operational model (approximate) and the true model (unknown).

Keywords–A posteriori law; a priori law; Bayesian information criterion; Kullaback-Leibler divergence; maximum a posteriori estimator

Date of Submission: 02-10-2024 Date of acceptance: 16-10-2024

I. INTRODUCTION

The topic of Bayesian model selection in linear regression has been the subject of extensive study. Over the years, numerous approaches have been proposed to address this key issue in Bayesian statistical modeling. Among the most notable works are the following: The BIC (Bayesian Information Criterion) was proposed by Schwarz (1978) [11], the Bayes factor by Kass and Raftery (1995) [5], the GBIC (Generalized Bayesian Information Criterion) Konishi et al. (2004) [7], and the MAPNI (Maximum a Posteriori Non Informatif) by G. Celeux et al. (2006) [2]. Nevertheless, some challenges remain unsolved. The calculation of the Bayes factor is often challenging when the model contains multiple explanatory variables. Furthermore, when multiple candidate models are available, the time required to validate the two hypotheses increases exponentially. In the case of the GBIC, we encountered a certain degree of difficulty in calculating this criterion, which is related to the challenge of determining its numerical value. This is due to the difficulty in determining the covariance matrix in the expression of this criterion. Finally, with regard to MAPNI, the criterion remains for the non-informative model, that is to say, the consideration of an a priori non-informative law. In this article, we put forth a Bayesian information criterion for a regression model. The criterion is obtained by calculating the average Kullback-Leibler divergence is a measure of dissimilarity between two probability distributions, P and Q. A small value of this quantity indicates that the chosen model is close to the true model. In this case, the minimum criterion will be a selected model that is hopefully close to the optimal choice. The remainder of this papers organized as follows: Section 2 provides a detailed account of Bayesian regression, followed by a proposed criterion for model selection. Discussions are presented before a conclusion and outlook in Section 3.

II. CONTENTS

2.1. Materials and methods

Bayesian analysis, as developed by T. Bayes (1763) [1] and P. S. Laplace (1795) [9], begins with an examination of a given situation and the identification of an uncertainty pertaining to an unknown parameter θ. This uncertainty is then quantified through the application of probabilistic distributions, utilizing fundamental principles of probability calculus. The uncertainty about θ is modeled in the form of a distribution, known as an a priori distribution, which provides information about θ taken as a random variable. This is in contrast to frequentist analysis, which w it as a constant. This a priori distribution is updated by extracting information from the observations of the variable X, to obtain another master distribution known as the a posteriori distribution.

We'll apply the Bayesian approach to regression models. It is different from the classical approach. It takes into account data and information from previous studies. The basic Bayes formula combines this information with new observed data.

2.1.1. Linear regression model

Let$y = (y_1, y_2, \dots, y_n)'$ be an dependent variable to be explained, and let $X = [\mathbb{1}|X_1|X_2| \dots |X_{p-1}]$ an explanatory variable where X_j is the vector of size n corresponding to the j −th variable and $\mathbb{1} = (1, \dots ,1)'$, the linear regression model relating these two variables is represented as follows: $y = \beta_0\mathbb{1} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1}X_{p-1} + \varepsilon$ ou $y = X\beta + \varepsilon$ [10]

where:

1. $X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}$ is the matrix of explanatory variables;
2. $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is the vector of the unknown parameters;
3. $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is the error vector.

Since $X\beta = E(y|\beta)$ and $D(y|\sigma^2) = D(\varepsilon|\beta, \sigma^2) = \sigma^2 P^{-1}$ is the covariance the matrix of y, then, $y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 P^{-1})$.

Thus, the likelihood function of y knowing β and σ^2 is defined as follows:

$p(y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'P(y - X\beta)\right\}$.

Therefore,

$$\begin{aligned}\ln\big(p(y|\beta, \sigma^2)\big) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'P(y - X\beta) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y'Py - 2\beta'X'Py + \beta'X'PX\beta) \qquad (1)\end{aligned}$$

Deriving (1) with respect to the variable β, we obtain $\hat{\beta} = (X'PX)^{-1}X'Py$, the maximum likelihood estimator of β. Similarly, for the parameter σ^2, we obtain: $\hat{\sigma}^2 = \frac{1}{n}\big(y - X\hat{\beta}\big)'P\big(y - X\hat{\beta}\big)$, the maximum likelihood estimator of σ^2.

2.1.2. Case of a priori non-informative law

One of the commonly used a priori laws is the so-called standard law (Jeffreys' a priori law [3]). Although it is improper, the resulting a posteriori law is proper probability density. For this standard a priori law, we have no information about the parameters other than σ^2, so β and σ^2 are assumed to be uniformly distributed and independent; this gives the joint a priori law of this parameters as follows: $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$.

By putting $\tau = 1/\sigma^2$, we therefore have$\pi(\beta, \tau) \propto \frac{1}{\tau}$. Using the likelihood function with the standard a priori joint distribution, we obtain a posteriori joint distribution:

$$\begin{aligned}\pi(\beta, \tau|y) &\propto (2\pi)^{-n/2}\tau^{n/2} exp\left\{-\frac{\tau}{2}(y - X\beta)'P(y - X\beta)\right\} \\ &\propto \tau^{n/2-1} exp\left\{-\frac{\tau}{2}(y - X\beta)'P(y - X\beta)\right\} \qquad (2)\end{aligned}$$

The exponent of the term (2) can be written as follows:

$$\begin{aligned}(y-X\beta)'P(y-X\beta) &= y'Py-2\beta'X'Py+\beta'X'PX\beta\\ &= y'Py-2(\mu^*)'X'Py+(\mu^*)'X'PX\mu^*+(\beta-\mu^*)'X'PX(\beta-\mu^*)\\ &= (y-X\mu^*)'P(y-X\mu^*)+(\beta-\mu^*)'X'PX(\beta-\mu^*) \qquad (3)\end{aligned}$$

With $\mu^*=(X'PX)^{-1}X'Py$ [6].

Remark 2.1. L*et Y be an $m\times 1$ random vector and X a random variable. Assume that $Y,X\sim NG(\mu,V,a,b)$, so the joint density function $f(y,x|\mu,V,a,b)$ is written as [6]:*

$$f(y,x|\mu,Va,b)=(2\pi)^{m/2}(det\,V)^{-1/2}a^b\big(\Gamma(b)\big)^{-1}$$
$$x^{-m/2+b-1}exp\left\{-\frac{x}{2}[2a+(y-\mu)^TV^{-1}(y-\mu)]\right\} \qquad (4)$$

with $a>0$, $b>0$, $0<x<+\infty$ and $-\infty<y_i<+\infty$.

If the random variables Y and X are distributed according to the normal-gamma (NG) distribution, $Y,X\sim NG(\mu,V,a,b)$, then the random vector Y has a marginal distribution that can be expressed as a t −multivariate distribution, also known as a multivariate Student:

$$Y\sim t(\mu,aV/b\,,2b) \qquad (5)$$

and the random variable X has marginal distribution that is the gamma distribution:

$$X\sim G(a,b) \qquad (6)$$

Substitution of the expression in (2) into (3) and comparison with the result obtained with $n/2-1=p/2+(n-p)/2-1$ in (4) reveals that the a posteriori density function (2) is such that:

$$\beta,\tau|y\sim NG(\mu^*,(X'PX)^{-1},n\hat{\sigma}^2/2,(n-p)/2). \qquad (7)$$

According to (5), the a posteriori marginal distribution for the vector β of unknown parameters is determined by the t −multivariate distribution [6]:

$$\beta|y\sim t\left(\mu^*,\frac{n}{n-p}\hat{\sigma}^2(X'PX)^{-1},n-p\right).$$

Therefore, the maximum a posteriori estimator $\hat{\beta}_{MAP}$ of β is written by $arg\max_{\beta}\pi(\beta|y)$. Let $\hat{\beta}_{MAP}=(X'PX)^{-1}X'Py$. And the covariance matrix $D(\beta|y)$ is written $D(\beta|y)=\frac{n}{n-p-2}\hat{\sigma}^2(X'PX)^{-1}$. Next, the marginal distribution for the weight parameter τ obtained from the a posteriori (7) is the gamma distribution, as can be seen from (6) $\tau|y\sim G(n\hat{\sigma}^2/2,(n-p)/2)$.

The inverse gamma distribution is therefore the a posteriori distribution of variance σ^2: $\sigma^2|y\sim IG(n\hat{\sigma}^2/2,(n-p)/2)$.

The maximum a posteriori estimator of σ^2 is thus defined by:

$$\hat{\sigma}^2_{MAP}=arg\max_{\sigma^2}\pi(\sigma^2|y)=\frac{n}{n-p+2}\hat{\sigma}^2=\frac{\big(y-X\hat{\beta}_{MAP}\big)'P\big(y-X\hat{\beta}_{MAP}\big)}{n-p+2}.$$

The variance of σ^2 is $V(\sigma^2|y)=\frac{2n^2(\hat{\sigma}^2)^2}{(n-p-2)^2(n-p-4)}$.

Properties 2.1. *(Laws of estimators). The Bayesian estimators $\hat{\beta}_{MAP}$ and $\hat{\sigma}^2_{MAP}$ have the following properties:*

- $(n-p+2)\hat{\sigma}^2_{MAP}/\sigma^2$ *follows a Chi-square with $n-p$ degrees of freedom $\left(\chi^2_{n-p}\right)$,*
- $\frac{(n-p)(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{p(n-p+2)\,\hat{\sigma}^2_{MAP}}$ *follows a Fisher distribution with $(p,n-p)$ degrees of freedom $(F(p,n-p))$.*

2.1.3. Case of a priori informative law

Let the variance factor σ^2 now be a random and unknown variable. To obtain a conjugate prior for the unknown parameters β and σ^2, we introduce in place of σ^2 the parameter of unknown weight τ with $\tau=1/\sigma^2$. As $\left(\det(\tau^{-1}P^{-1})\right)^{-1/2}=(\det(P))^{1/2}\tau^{n/2}$, the likelihood function is written as follows :

$$p(y|\beta,\tau)=(2\pi)^{-n/2}(\det(P))^{1/2}\tau^{n/2}exp\left[-\frac{\tau}{2}(y-X\beta)'P(y-X\beta)\right]$$

As a priori for β and τ, the density function (4) of the normal-gamma distribution:

$$\beta,\tau\sim NG(\mu,V,a,b) \qquad (8)$$

is chosen [8]. The joint a posteriori distribution is then obtained by combining the likelihood function with the joint a priori distribution:

$$\pi(\beta,\tau|y) \propto \tau^{p/2+b-1} exp\left\{-\frac{\tau}{2}[2a+(\beta-\mu)'V^{-1}(\beta-\mu)]\right\}\tau^{n/2} exp\left[-\frac{\tau}{2}(y-X\beta)'P(y-X\beta)\right]$$
$$= \tau^{n/2+b+p/2-1} exp\left\{-\frac{\tau}{2}[2a+(\beta-\mu)'V^{-1}(\beta-\mu)+(y-X\beta)'P(y-X\beta)]\right\} \qquad (9)$$

The bracketed expression for the exponent can be written as:

$$2a+y'Py+\mu'V^{-1}\mu-2\beta'(X'Py+V^{-1}\mu)+\beta'(X'PX+V^{-1})\beta$$
$$=2a+y'Py+\mu'V^{-1}\mu-(\mu^*)'(X'PX+V^{-1})\mu^*+(\beta-\mu^*)'(X'PX+V^{-1})(\beta-\mu^*)$$
$$=2a+y'Py+\mu'V^{-1}\mu-2(\mu^*)'(X'PX+V^{-1}\mu)+(\mu^*)'(X'PX+V^{-1})(\mu^*)'+(\beta-\mu^*)'(X'PX+V^{-1})(\beta-\mu^*)$$
$$=2a+(\mu-\mu^*)'V^{-1}(\mu-\mu^*)+(y-X\mu^*)'P(y-X\mu^*)+(\beta-\mu^*)'(X'PX+V^{-1})(\beta-\mu^*) \qquad (10)$$

With $\mu^* = (X'PX+V^{-1})^{-1}(X'Py+V^{-1}\mu)$. Substituting (10) into (9) and comparing with (4), we obtain:

$$\beta,\tau|y \sim NG(\mu^*,V^*,a^*,b^*) \qquad (11)$$

where:

- $V^* = (X'PX+V^{-1})^{-1}$;
- $a^* = [2a+(\mu-\mu^*)'V^{-1}(\mu-\mu^*)+(y-X\mu^*)'P(y-X\mu^*)]/2$;
- $b^* = n/2+b$.

In accordance with (5) and (11), the a posteriori marginal distribution of β is identified as the t −multivariate distribution, specifically:

$\beta|y \sim t(\mu^*, a^*V^*/b^*, 2b^*)$

Consequently, the maximum a posteriori $\hat{\beta}_{MAP}$ estimator of β is given $arg\max_{\beta}\pi(\beta|y)=\mu^*$. Therefore, $\hat{\beta}_{MAP}$ can be expressed as:

$\hat{\beta}_{MAP} = (X'PX+V^{-1})^{-1}(X'Py+V^{-1}\mu)$.

Moreover, according to (6), the a posteriori marginal distribution for the weight parameter τ is the gamma distribution. Consequently, the variance, $\sigma^2 = \frac{1}{\tau}$ has an inverse gamma distribution [6], [4]:

$\sigma^2|y \sim IG(a^*,b^*)$

Therefore, the maximum a posteriori estimator $\hat{\sigma}^2_{MAP}$ of σ^2 is given by $arg\max_{\sigma^2}\pi(\sigma^2|y)=\frac{a^*}{b^*+1}$. So we have:

$$\hat{\sigma}^2_{MAP} = \frac{2a+\left(\mu-\hat{\beta}_{MAP}\right)^T V^{-1}\left(\mu-\hat{\beta}_{MAP}\right)+\left(y-X\hat{\beta}_{MAP}\right)'P\left(y-X\hat{\beta}_{MAP}\right)}{n+2b+2}.$$

And the variance of σ^2 is $V(\sigma^2|y)=\left(\frac{b^*+1}{b^*-1}\right)^2\frac{\left(\hat{\sigma}^2_{MAP}\right)^2}{b^*-2}$. The covariance matrix of β is $D(\beta|y)=E\big((\beta-E(\beta|y))'(\beta-E(\beta|y))\big)=\left(\frac{b^*+1}{b^*-1}\right)\hat{\sigma}^2_{MAP}(X'PX+V^{-1})^{-1}$.

Properties 2.2. *Bayesian estimators for informative a priori law have the following properties:*

1. $2(b^*+1)\frac{\hat{\sigma}^2_{MAP}}{\sigma^2}$ *follows a Chi-square distribution with* $2b^*$ *degrees of freedom;*
2. $\frac{b^*(\beta-\hat{\beta}_{MAP})'(X'PX+V^{-1})(\beta-\hat{\beta}_{MAP})}{p(b^*+1)\hat{\sigma}^2_{MAP}}$ *follows a Fisher distribution with* $(p,2b^*)$ *degrees of freedom.*

2.2. Results

A useful measure of the deviation between the operational model and the approximate model is defined from the Kullback-Leibler divergence [8]:

$$\Delta(\beta,\sigma^2) = E_y(-2\ln(p(y|\beta,\sigma^2))) \qquad (12)$$

Where $p(y|\beta,\sigma^2)$ is the likelihood function under the approximate model. A reasonable criterion for judging the quality of the family of approximations given the data is $E_{(\hat{\beta},\hat{\sigma}^2)}\{\Delta(\beta,\sigma^2)\}$. Given a collection of competing families of approximations, the one that minimizes $E_{(\hat{\beta},\hat{\sigma}^2)}\{\Delta(\beta,\sigma^2)\}$ is in some sense closest to truth and should be preferred. Of course, $E_{(\hat{\beta},\hat{\sigma}^2)}\{\Delta(\beta,\sigma^2)\}$ is unknown, but it can be estimated if certain additional assumptions are made. In this subsection, we will compute this quantity in the two previous cases of maximum a posteriori estimators (non-informative and informative a priori law), i.e. $\left(\hat{\beta},\hat{\sigma}^2\right)=\left(\hat{\beta}_{MAP},\hat{\sigma}^2_{MAP}\right)$ in order to propose a Bayesian regression model selection criterion. And this criterion will be called Bayesian Kullback-Leibler Information Criterion, denoted as BKLIC.

2.2.1. For the case a priori non-informative law

The maximum a posteriori estimators $\hat{\beta}_{MAP}$ and $\hat{\sigma}^2_{MAP}$ of β and σ^2 are:

- $\hat{\beta}_{MAP} = (X'PX)^{-1}X'Py$;

- $\hat{\sigma}_{MAP}^2 = \frac{(y-X\hat{\beta}_{MAP})'P(y-X\hat{\beta}_{MAP})}{n-p+2}$.

We'll determine the quantity $-2E_y E_{\hat{\theta}_{MAP}(x)}\left[\ln\left(p(y|\hat{\theta}_{MAP}(x)\right)\right]$ where $\hat{\theta}_{MAP} = \left(\hat{\beta}_{MAP}, \hat{\sigma}_{MAP}^2\right)$.

We know that $\ln(p(y|\beta,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\ln(\det(P)) - \frac{1}{2\sigma^2}(y-X\beta)'P(y-X\beta)$.

So, $\ln\left(p(y|\hat{\beta}_{MAP},\hat{\sigma}_{MAP}^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) + \frac{1}{2}\ln(\det(P)) - \frac{1}{2\hat{\sigma}_{MAP}^2}\left(y-X\hat{\beta}_{MAP}\right)'P\left(y-X\hat{\beta}_{MAP}\right)$.

Therefore, $\ln\left(p(y|\hat{\beta}_{MAP},\hat{\sigma}_{MAP}^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) - \frac{n-p+2}{2}$.

Hence,

$$T = E_y E_{\hat{\theta}_{MAP}(x)}\left(\ln\left(p(y|\hat{\theta}_{MAP}(x)\right)\right)$$

$$= E_{\hat{\theta}_{MAP}(x)} E_y\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) + \frac{1}{2}\ln(\det(P)) - \frac{1}{2\hat{\sigma}_{MAP}^2}\left(y-X\hat{\beta}_{MAP}\right)'P\left(y-X\hat{\beta}_{MAP}\right)\right]$$

$$= E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) + \frac{1}{2}\ln(\det(P)) - \frac{1}{2\hat{\sigma}_{MAP}^2}E_y\left(\left(y-X\hat{\beta}_{MAP}\right)'P\left(y-X\hat{\beta}_{MAP}\right)\right)\right]$$

We know that:

$$E_y\left[\left(y-X\hat{\beta}_{MAP}\right)'P\left(y-X\hat{\beta}_{MAP}\right)\right]$$

$$= E_y\left[\left((y-X\beta)+\left(X\beta-X\hat{\beta}_{MAP}\right)\right)'P\left((y-X\beta)+\left(X\beta-X\hat{\beta}_{MAP}\right)\right)\right]$$

$$= E_y[(y-X\beta)'P(y-X\beta)] + E_y\left[2\left(X\beta-X\hat{\beta}_{MAP}\right)'P(y-X\beta)\right]$$
$$+E_y\left[\left(X\beta-X\hat{\beta}_{MAP}\right)'P\left(X\beta-X\hat{\beta}_{MAP}\right)\right]$$

$$= E_y[(y-X\beta)'P(y-X\beta)] + \left[2\left(X\beta-X\hat{\beta}_{MAP}\right)'P\left(E_y(y)-X\beta\right)\right]$$
$$+\left[\left(X\beta-X\hat{\beta}_{MAP}\right)'P\left(X\beta-X\hat{\beta}_{MAP}\right)\right]$$

The term in the middle of the previous expression cancels out because $E_y(y) = X\beta$. Moreover, the first term can be written as follows: $E_y[(y-X\beta)'P(y-X\beta)] = E_y[\varepsilon' P\varepsilon] = E_y[tr(\varepsilon' P\varepsilon)] = tr\left[PE_y(\varepsilon'\varepsilon)\right] = tr(\sigma^2 PP^{-1}) = n\sigma^2$. So we have the result: $E_y\left[\left(y-X\hat{\beta}_{MAP}\right)^T\left(y-X\hat{\beta}_{MAP}\right)\right] = n\sigma^2 + \left[\left(X\beta - X\hat{\beta}_{MAP}\right)'P\left(X\beta-X\hat{\beta}_{MAP}\right)\right]$. Using this partial result, we thus have:

$$T = E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) + \frac{1}{2}\ln(\det(P))\right] - E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2 + \left[\left(X\beta-X\hat{\beta}_{MAP}\right)'P\left(X\beta-X\hat{\beta}_{MAP}\right)\right]}{2\hat{\sigma}_{MAP}^2}\right]$$

$$= E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_{MAP}^2) + \frac{1}{2}\ln(\det(P))\right] - \frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2}{\hat{\sigma}_{MAP}^2}\right]$$
$$-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{\left(\beta-\hat{\beta}_{MAP}\right)'X'PX\left(\beta-\hat{\beta}_{MAP}\right)}{\hat{\sigma}_{MAP}^2}\right] \qquad (13)$$

We know that $\frac{(n-p+2)\hat{\sigma}_{MAP}^2}{\sigma^2} \sim \chi_{n-p}^2$ from property (2.1). Since, $\frac{\hat{\sigma}_{MAP}^2}{\sigma^2} = \frac{1}{n-p+2}\frac{(n-p+2)\hat{\sigma}_B^2}{\sigma^2}$, we have:

$$\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2}{\hat{\sigma}_{MAP}^2}\right] = \frac{n}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{1}{\frac{\hat{\sigma}_{MAP}^2}{\sigma^2}}\right]$$

$$= \frac{n}{2}E_{MAP(x)}\left[\frac{1}{\frac{1}{n-p+2}\frac{(n-p+2)\hat{\sigma}_{MAP}^2}{\sigma^2}}\right]$$

$$= \frac{n(n-p+2)}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{1}{\frac{(n-p+2)\hat{\sigma}_{MAP}^2}{\sigma^2}}\right]$$

$$= \frac{n(n-p+2)}{2(n-p-2)}$$

Furthermore, $\frac{(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{\hat{\sigma}^2_{MAP}}=\frac{p(n-p+2)}{n-p}\frac{(n-p)(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{p(n-p+2)\hat{\sigma}^2_{MAP}}$.

From property (2.2) we have $\frac{(n-p)(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{p(n-p+2)\hat{\sigma}^2_{MAP}}\sim F(p,n-p)$. This will give the following result:

$$\begin{aligned}\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{\hat{\sigma}^2_{MAP}}\right]&=\frac{p(n-p+2)}{2(n-p)}E_{\hat{\theta}_B(x)}\left[\frac{(n-p)(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{p(n-p+2)\hat{\sigma}^2_{MAP}}\right]\\&=\frac{p(n-p+2)}{2(n-p)}\frac{(n-p)}{(n-p-2)}\\&=\frac{p(n-p+2)}{2(n-p-2)}\end{aligned}$$

Using the previous results, we have:

$$\begin{aligned}T&=E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})+\frac{1}{2}\ln(\det(P))\right]-\frac{(n+p)(n-p+2)}{2(n-p-2)}\\&=E\left[-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})+\frac{1}{2}\ln(\det(P))\right]-\frac{(n+p)(n-p+2)}{2(n-p-2)}\end{aligned}$$

Let, $-2T=n\ln(2\pi)+n\ln(\hat{\sigma}^2_{MAP})-\ln(\det(P))+\frac{(n+p)(n-p+2)}{n-p-2}$.

Therefore, a selection criterion for a multiple linear regression model in the Bayesian estimation framework can be obtained when the a priori distribution of the unknown parameters is non-informative:

$$BKLIC=n\ln(2\pi)+n\ln(\hat{\sigma}^2_{MAP})-\ln(\det(P))+\frac{(n+p)(n-p+2)}{2(n-p-2)}.$$

2.2.2. For the case of a priori informative

The maximum a posteriori estimators $\hat{\beta}_{MAP}$ and $\hat{\sigma}^2_{MAP}$ of β and σ^2 are:

- $\hat{\beta}_{MAP}=(X'PX+V^{-1})^{-1}(X'Py+V^{-1}\mu)$;
- $\hat{\sigma}^2_{MAP}=\frac{2a+(\mu-\hat{\beta}_{MAP})'V^{-1}(\mu-\hat{\beta}_{MAP})+(y-X\hat{\beta}_{MAP})'P(y-X\hat{\beta}_{MAP})}{n+2b+2}$.

From relation (13), we have:

$$\begin{aligned}T&=E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})\right]-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2}{\hat{\sigma}^2_{MAP}}\right]-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{(\beta-\hat{\beta}_{MAP})'X'PX(\beta-\hat{\beta}_{MAP})}{\hat{\sigma}^2_{MAP}}\right]\\&=E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})\right]-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2}{\hat{\sigma}^2_{MAP}}\right]\\&\quad-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{(\beta-\hat{\beta}_{MAP})'(X'PX+V^{-1})(\beta-\hat{\beta}_{MAP})}{\hat{\sigma}^2_{MAP}}\right]+\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{(\beta-\hat{\beta}_{MAP})'V^{-1}(\beta-\hat{\beta}_{MAP})}{\hat{\sigma}^2_{MAP}}\right]\end{aligned}$$

The preceding expression will be evaluated in order to ascertain the value of the last three terms, utilizing the properties of estimators for the case of informative a priori distribution:

$$\begin{aligned}-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{n\sigma^2}{\hat{\sigma}^2_{MAP}}\right]&=-\frac{n}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{\sigma^2}{\hat{\sigma}^2_{MAP}}\right]\\&=-\frac{n}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{1}{\frac{\hat{\sigma}^2_{MAP}}{\sigma^2}}\right]\\&=-n(b^*+1)E_{\hat{\theta}_{MAP}(x)}\left[\frac{1}{2(b^*+1)\frac{\hat{\sigma}^2_{MAP}}{\sigma^2}}\right]\\&=-\frac{n(b^*+1)}{2b^*-2}\end{aligned}$$

Furthermore:

$$-\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{\left(\beta-\hat{\beta}_{MAP}\right)'(X'PX+V^{-1})\left(\beta-\hat{\beta}_{MAP}\right)}{\hat{\sigma}^2_{MAP}}\right] = -\frac{p(b^*+1)}{2b^*}E_{\hat{\theta}_{MAP}(x)}\left[\frac{b^*\left(\beta-\hat{\beta}_{MAP}\right)'(X'PX+V^{-1})\left(\beta-\hat{\beta}_{MAP}\right)}{p(b^*+1)\hat{\sigma}^2_{MAP}}\right]$$

$$= \frac{p(b^*+1)}{2b^*}\frac{2b^*}{2b^*-2}$$

$$= -\frac{p(b^*+1)}{2(b^*-1)}$$

Finally:

$$\frac{1}{2}E_{\hat{\theta}_{MAP}(x)}\left[\frac{\left(\beta-\hat{\beta}_{MAP}\right)'V^{-1}\left(\beta-\hat{\beta}_{MAP}\right)}{\hat{\sigma}^2_{MAP}}\right] = E_{\hat{\beta}_{MAP}}\left[\left(\beta-\hat{\beta}_{MAP}\right)'V^{-1}\left(\beta-\hat{\beta}_{MAP}\right)\right]E_{\hat{\sigma}^2_{MAP}}\left[\frac{1}{\hat{\sigma}^2_{MAP}}\right]$$

$$= \frac{1}{2\hat{\sigma}^2_{MAP}}E_{\hat{\beta}_{MAP}}\left[tr\left(\left(\beta-\hat{\beta}_{MAP}\right)'V^{-1}\left(\beta-\hat{\beta}_{MAP}\right)\right)\right]$$

$$= \frac{1}{2\hat{\sigma}^2_{MAP}}tr\left[V^{-1}E_{\hat{\beta}_{MAP}}\left[\left(\beta-\hat{\beta}_{MAP}\right)'\left(\beta-\hat{\beta}_{MAP}\right)\right]\right]$$

$$= \frac{1}{2\hat{\sigma}^2_{MAP}}tr\left[V^{-1}E_{\hat{\beta}_{MAP}}[(\beta-E(\beta|y))'(\beta-E(\beta|y))]\right]$$

$$= \frac{1}{2\hat{\sigma}^2_{MAP}}tr\left[V^{-1}\left(\frac{b^*+1}{b^*-1}\right)\hat{\sigma}^2_{MAP}(X'PX+V^{-1})^{-1}\right]$$

$$= \frac{1}{2}\left(\frac{b^*+1}{b^*-1}\right)tr\left[\left((X'PX+V^{-1})V\right)^{-1}\right]$$

$$= \frac{1}{2}\left(\frac{b^*+1}{b^*-1}\right)tr[(X'PXV+I)^{-1}]$$

The previous results allow us to write:

$$T = E_{\hat{\theta}_{MAP}(x)}\left[-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})+\frac{1}{2}\ln(\det(P))\right]-\left(\frac{b^*+1}{b^*-1}\right)\left[\frac{n+p}{2}-\frac{1}{2}tr[(X'PXV+I)^{-1}]\right]$$

$$= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2_{MAP})+\frac{1}{2}\ln(\det(P))-\left(\frac{b^*+1}{b^*-1}\right)\left[\frac{n+p}{2}-\frac{1}{2}tr[(X'PXV+I)^{-1}]\right]$$

Therefore, we can conclude that, $-2T = n\ln(2\pi)+n\ln(\hat{\sigma}^2_{MAP})-\ln(\det(P))+\left(\frac{b^*+1}{b^*-1}\right)\left[n+p-tr[(X'PXV+I)^{-1}]\right]$.

Similarly, as with a non-informative a priori law, a criterion for selecting a multiple linear regression model exists in the event that the a priori law is informative:

$$BKLIC = n\ln(2\pi)+n\ln(\hat{\sigma}^2_{MAP})-\ln(\det(P))+\left(\frac{b^*+1}{b^*-1}\right)\left[n+p-tr[(X'PXV+I)^{-1}]\right].$$

2.4. Discussions

In the present subsection we will utilize a database designated as “eucalyptus” for our analysis. The data set comprises the height and circumference of 1,429 eucalyptus trees. The dependent variable is their height and the explanatory variables are their circumference and its square root. In this context, we will evaluate the quality of the models $y = \beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$ and $y = \alpha_0+\alpha_1X_1+\epsilon$. Calculating the AIC, BIC and BKLIC, we obtain the results on the table below:

Table 1. Criteria values for the two candidate models

	Model 1: $y = \beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$	Model 2: $y = \alpha_0+\alpha_1X_1+\epsilon$
AIC BIC	4429.074 4446.133	4578.454 4594.248
BKLIC$_1$ BKLIC$_2$	3710.088 4460.712	3862.964 4608.43

where BKLIC$_1$ is the new criterion for non-informative a priori law and BKLIC$_2$ for the informative a priori law. These four criteria lead us to choose model 1, i.e. the adapted forecasting model seems to be the multiple regression model. After estimating the parameters of the chosen model, we obtain the following results in the case where the a priori law is informative:

```
                                beta_0       beta_1     beta_2        sigma^2
Maximum a posteriori: -24.3446729  -0.4828339   9.9850919   1.3220238
```

The resulting model is written as follows: $y = -24.34 - 0.48X_1 + 9.99X_2$.

Comment: We can see that these criteria select the same model. This allows us to say that these new criteria can be used within the framework of Bayesian statistics, more precisely in the case of Bayesian regression: one for the estimate whose a priori distribution is uninformative and the other for the a priori distribution that is informative. Furthermore, this criterion can be represented graphically as a function of the number of explanatory variables. A sample of size $n = 20$ was taken, and the various values of this criterion were obtained according to the number of independent variables (explanatory variables k) for the two cases of a priori laws (see Table 2).

Table 2. Various values of the two criteria as a function of explanatory variables

Number of explanatory variables	BKLIC$_1$	BKLIC$_2$
1	57.0925	70.17104
2	-1261.5	21.25829
3	-1258.951	22.47774
4	-1255.976	22.72266

A minimum is topically obtained for a specific value of k. This minimum provides the optimal model, as illustrated in Fig. 1 below:

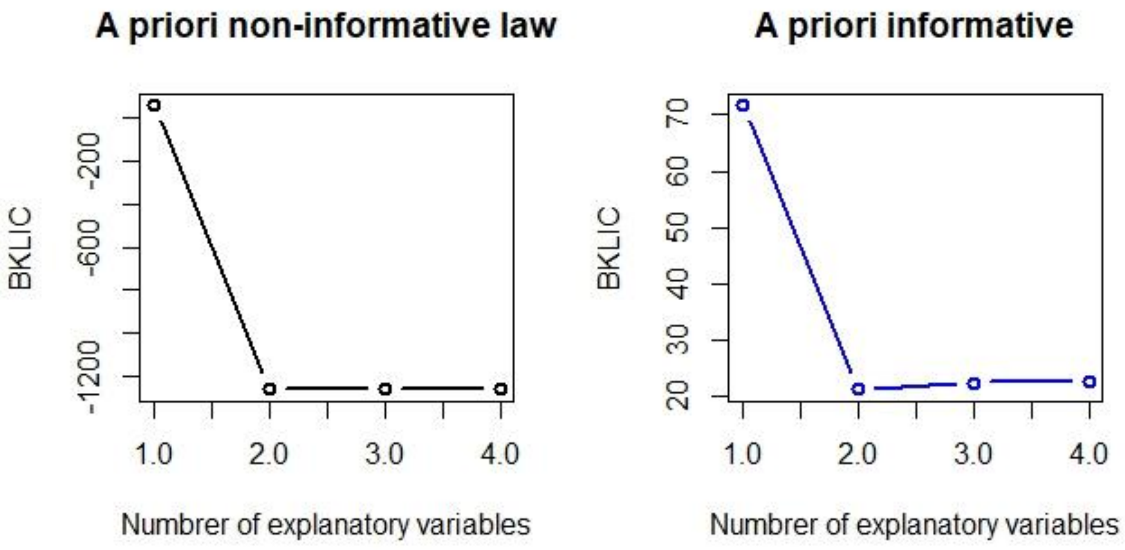

Figure1: BKLIC curves

III. CONCLUSION

In this article. We put forth two Bayesian information criteria for the purpose of selecting a linear regression model. The two criteria provide an accessible tool for selecting a Bayesian linear regression model when estimators are obtained by maximum a posteriori. These criteria do not adhere to a single a priori case; rather, they consider two cases simultaneously. Moreover, experimental results demonstrate the efficacy of these two criteria in identifying a reliable model among the candidates. From an explanatory standpoint, the results produced by these criteria align with those of the classical criteria, particularly in the context of small sizes. This validates the efficacy of our criterion in the domain of Bayesian linear regression model inference. However, further development is necessary. In the future, we intend to extend these two criteria to the Bayesian analysis of time series models, including AR(p) and ARMA(p,q) models.

REFERENCES

[1]. T. Bayes, *An essay towards solving a problem in the doctrine of chances*, Philosophical Transactions Royal Society London, 53, 1763, 370-418.

[2]. G. Celeux, J. M. Marin and C. Robert, Sélection bayésienne de variables en régression linéaire, Journal de la société française de statistique, tome 147, num 1, 2006, 59-79.

[3]. H. Jeffreys, Theory of probability (New York: Oxford University Press, 1939).

[4]. J. D. Hamilton, Times series analysis (Princeton University Press, 41 William St., Princeton, New Jersey 08540, 1954).

[5]. R.Kass andA. Raftery, Bayes factors, Journal of the American Statistical Association, 90, 1995, 773-795.

[6]. K. R. Koch, Introduction to Bayesian statistics (Second Edition, Springer, 2007).

[7]. S. Konishi, T. Ando and S. Imoto, Bayesian information criteria and smoothing parameter selection in radial basis function networks, Biometrika, 91, 2004, 27-43.

[8]. S. Kullback and R. Leibler, On information and sufficiency, Annals of Mathematical Statistics, 22, 1951, 79-86.

[9]. P. S. Laplace, Essai philosophique sur les probabilités (Epistémé. Christian Bourgeois, Paris. Reprinted in 1986, 1795).

[10]. S. D. Permai and H. Tanty, Linear regression model using Bayesian approach for energy performance of residential building, Procedia Computer Science, 135, 2018, 671-677.

[11]. G. Schwarz, Estimating the dimension of a model, Annals of Statistics, 6, 1978, 461-464.