Scientific Research Publishing

Open Journal of Applied Sciences, 2024, 14, 1775-1787
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Comparative Study of Probabilistic and Least-Squares Methods for Developing Predictive Models

Boribo Kikunda Philippe[1,2,3], Thierry Nsabimana[1], Jules Raymond Kala[2], Jeremie Ndikumagenge[1], Longin Ndayisaba[1]

[1]Centre de Recherche en Infrastructure Environnement et Technologie (CRIET), Faculty of Engineering Sciences, Université du Burundi, Bujumbura, Burundi
[2]Faculty of Sciences, Université Catholique de Bukavu (UCB), Bukavu, DR Congo
[3]Management Computer Department, Institut supérieur Pédagogique de bukavu (ISP/Bukavu), Bukavu, DR Congo
Email: kikunda.boribo@ucbukavu.ac.cd, thierry.nsabimana@ub.edu.bi, jeremie.ndikumagenge@ub.edu.bi, longin.ndayisaba@ub.edu.bi

How to cite this paper: Philippe, B.K., Nsabimana, T., Kala, J.R., Ndikumagenge, J. and Ndayisaba, L. (2024) Comparative Study of Probabilistic and Least-Squares Methods for Developing Predictive Models. *Open Journal of Applied Sciences*, **14**, 1775-1787.
https://doi.org/10.4236/ojapps.2024.147116

Received: May 6, 2024
Accepted: July 16, 2024
Published: July 19, 2024

Abstract

This article explores the comparison between the probability method and the least squares method in the design of linear predictive models. It points out that these two approaches have distinct theoretical foundations and can lead to varied or similar results in terms of precision and performance under certain assumptions. The article underlines the importance of comparing these two approaches to choose the one best suited to the context, available data and modeling objectives.

Keywords

Predictive Models, Least Squares, Bayesian Estimation Methods

1. Introduction

Predictive modeling is an essential step in data analysis and data science, aimed at developing models capable of predicting future events or behaviors based on historical data [1]. In other words, it represents the process by which a mathematical model synthesizing a reality is found. This model can then be used to make decisions in the face of new data, which would not normally have been used to develop it [2].

It is used in many fields such as finance, healthcare, marketing, education and many others.

The choice of an efficient algorithm in the process of designing a model can sometimes be very time-consuming. Knowledge of the general properties of algorithms and of problems can be a major asset in choosing the right algorithm. Today, there are several algorithms available for setting up models. These can be grouped into two approaches especially in the case of linear models: the maximum likelihood approach and the Bayesian approach. The choice of a linear model is often motivated by the need to understand the relationship between the factors influencing a phenomenon and the phenomenon itself [3].

These two approaches offer distinct theoretical frameworks for modeling and predicting phenomena based on observed data. This can often lead to debate as to which is more effective at predicting future outcomes. The probabilistic method relies on concepts of statistics and probability to estimate the parameters of a model [4]. In this perspective, the parameter is considered as a random variable whose probability distribution must be estimated [5], whereas the least-squares method is based on minimizing the differences between observed and predicted values. Some authors, such as Legendre, describe it as an algebraic method for solving an incompatible system of n equations with m unknowns [6].

The comparison between the two approaches will enable us to highlight the assumptions that would lead to the choice of one algorithm over another to solve a given problem. In this article, we will explore these two approaches in detail, highlighting their theoretical foundations, strengths, limitations and application in the design of predictive models to enable researchers and practitioners to choose the most appropriate method for their specific needs.

2. Least Squares Method

The method of least squares was introduced by Karl Gauss in 1809. It is used to estimate the parameters of a mathematical model by minimizing the sum of the deviations between the observed values and the values predicted by the model [7]. In other words, it seeks to find the best approximation to a linear relationship between variables by adjusting the model coefficients so that the mean square error is minimal.

2.1. Principle

First, let's remember that a model is said to be linear if it is linear with respect to its parameters [4]. This means it can be represented either by a straight line or by a curve. This is the case, for example, with a polynomial model, which is classified as a linear model. In this case, the general form of a linear model is given by the following formula:

$$g(z,w)=\sum_{i=1}^{p} w_i f_i(z) \tag{1}$$

where the functions $f_i(z)$ are non-parametric functions of the variables. We can replace them with x_i and we obtain the form:

$$g(x,w)=w\cdot x \tag{2}$$

In the formula (2), w represents the parameter vector and x represents a row of data. If we have n rows of data, we can define the observation matrix X, which will have n rows and p columns. Thus, the linear model becomes:

$$g(X,w)=X\cdot w \tag{3}$$

Notation Given the data matrix X and y vector of data representing the phenomenon to be modeled, the problem therefore comes down to estimating the parameters w of the linear model of the g function. Least squares proposes the function to be minimized

$$J(w)=\sum_{k=1}^{N_A}\left(y_k^p-g(x_k,w)\right)^2$$

où N_A is the data sample size, x_k is the vector of variables for the data k, and y_k^p is the value of the quantity to be modeled for the data k. To find parameter values w for which this function is minimal, it suffices to write that its first derivative is zero:

$$\nabla_w J=\frac{\partial J(w)}{dw}=0$$

which take a set of p equations, whose p unknowns are the parameters w_i, $i=1?p$. These equations are called normal equations. We show that

$$w_{mc}=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y^p \tag{4}$$

When X has a large dimension, recursive least squares is used [8]. The numerical method to be used can be either singular value separation or gradient descent [9].

2.2. Geometric Interpretation of the Least-Squares Solution

This solution can be obtained using a geometric method, which is an elegant interpretation of the method of least squares [10]. Determining the w coefficients of formula 3 is equivalent to solving the system $X\cdot w=y$. This system is incompatible because it has more equations than unknowns. Since a solution cannot be found because $y\notin ImX$, we're forced to settle for finding a w' such that $X\cdot w'=y'$ with y' as close as possible to y. This means that $\forall y'',\exists y'$ of ImX, $\left\|Xw'-y''\right\|>\left\|Xw'-y'\right\|$. From a geometric point of view, y' closest to y is that which is the orthogonal project of y in the image of the matrix X. Therefore, all column vectors of X are orthogonal to the vector $y-y'$. Thus,

$$X^{\mathrm{T}}\left(y-y'\right)=0$$

$$X^{\mathrm{T}}\left(Xw'-y'\right)=0$$

$$X^{\mathrm{T}}Xw'-X^{\mathrm{T}}y'=0$$

$$X^{\mathrm{T}}Xw'=X^{\mathrm{T}}y'$$

$$w'=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y \tag{5}$$

We note that the solution of the formula 4 is the same as that obtained by the formula 5.

However, most of the time, researchers manipulate very limited data in the form of a sample, tainted by noise. The fact that a sample is drawn at random means that the solution may vary with respect to the sample drawn. Hence the need to formulate the problem of estimating the parameters of a model from a probabilistic point of view.

3. Probabilistic Approach

Probability is defined according to two schools of thought [11]:

- The frequentist approach, according to which the probability of an event is the relative frequency with which this event occurs when an experiment is repeated a large number of times under identical conditions.
- The Bayesian approach, in which probability is seen as a measure of subjective belief or uncertainty about an event. It represents the degree of belief associated with the occurrence of an event.

3.1. Parameter Estimation Using the Frequentist Approach

The frequentist method for estimating model parameters is maximum likelihood estimation [12]. It seeks to find the parameter values that make the observed data more probable within the framework of the model under consideration.

Thus the probability density of y knowing x is given by:

Let ε_k be the deviation between the predicted value of the model for data k in the sample and the observed value of the measurement to be modeled y_k i.e $\varepsilon_k = y_k - g(w,x)$. Suppose that ε_k follows a normal distribution with mean 0 and variance δ^2 *i.e.* $\varepsilon \sim N(0,\delta^2)$ et ε are independent.

This implies that y knowing x follows a normal distribution with mean $g(x,w)$ and variance σ^2.

$$P\left(y/x,w,\sigma^2\right) = N\left(g(x,w),\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-g(x,w)\right)^2}{2\sigma^2}\right] \tag{6}$$

The maximum likelihood function L for n independent observations following a normal distribution of $\mu = g(w,x)$ with variance σ^2 will be written as the product of the individual probability density of each observation.

$$L\left(y/X,w,\sigma^2\right) = \prod_{k=1}^{N}\left[N\left(g(x_k,w),\sigma^2\right)\right] \tag{7}$$

To simplify the calculation, we introduce the logarithm into the likelihood function, and obtain:

$$\begin{aligned}\ln L\left(y/X,w,\sigma^2\right) &= \ln\prod_{k=1}^{N}\left[N\left(g(x_k,w),\delta^2\right)\right] \\ &= \sum_{k=1}^{N}\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y_k - wx_k}{2\sigma^2}\right)\right]\end{aligned} \tag{8}$$

Let $\beta = \frac{1}{\sigma^2}$

$$\ln L\left(y/X, w, \sigma^2\right) = \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\beta) - \frac{\beta}{2}\sum_{k=1} N\left[\left(y - f\left(x_k, w\right)\right)^2\right]$$

$$-\ln L\left(y/X, w, \sigma^2\right) = \frac{\beta}{2}\sum_{k=1} N\left[\left(y - f\left(x_k, w\right)\right)^2\right] - \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) \quad (9)$$

Since $\ln(2\pi)$ and $\ln(\beta)$ are constant, formula (9) can be written as:

$$-\ln L\left(y/X, w, \sigma^2\right) = \frac{\beta}{2}\sum_{k=1}^{N}\left[\left(y - f\left(x_k, w\right)\right)^2\right] + \text{constant} \quad (10)$$

Now, maximizing (-*y*) is equivalent to minimizing (*y*). This being the case, formula (10) corresponds exactly to the minimization of deviations obtained by the algebraic approach to formula (4). So the least squares method is the result of the frequentist approach to solving the problem of estimating the parameters of a linear model under the assumption of a normal noise distribution [12]. But maximum likelihood estimation of the variance of noise $sigma^2$ is often underestimated or overestimated [13]. If the number of parameters and data evolve simultaneously, or if the density function is not convex, this phenomenon is at the root of overlearning or underlearning, a key property of this approach.

Having determined the optimal value of the parameters w and β, we can now make a prediction for a new value of x. The probability distribution of y knowing x, w, β can then be formulated as follows:

$$P\left(t \mid x, w^*, \beta^*\right) = N\left(t \mid y\left(x, w^*\right), \beta^{-1}\right) \quad (11)$$

This approach provides good results when we have a large quantity of data and the relationship between the variables x and y are linear. But when the sample size is very small, it can lead to overlearning and poor prediction results [1]. That's why another approach is possible in such circumstances.

In practice, we'll be using python's scikit-learn package for least-squares parameter estimation.

3.2. Hypothesis and Limitations of Using Least Squares

The least-squares approach produces good results when the properties contained in the data exist, including linearity, independence of observations, constant variance of the error term and normality of the latter. However, this method can be ineffective if the data sample contains outliers, multicollinearity between variables and small data size. The latter characteristic results in overfitting.

3.3. Parameter Estimation Using the Bayesian Approach

We can also take a Bayesian approach to parameter estimation. In this case, it is important to define an a priori probability for the parameters and an a posteriori probability obtained by Bayes' theorem. The maximum a posteriori method will be used to incorporate the a priori distribution into parameter estimation.

The a priori probability distribution on the parameters is the probability on

the parameters before having observed the data. It will be denoted by

$$P(w \mid \text{alpha}) = N\left(0, \alpha^{-1} I\right).$$

The a posteriori distribution is an update of the beliefs or parameter distribution after taking into account the a priori distribution and the observed data. Using Bayes' theorem, we can express it as follows:

$$P\left(w \mid X, \alpha, \beta^{-1}\right) = \frac{L\left(Y/X, w, \beta^{-1}\right) \cdot P(w \mid \alpha)}{P(Y/X)}$$

The a posteriori distribution is directly proportional to the a priori distribution and likelihood.

$$P\left(w \mid X, \alpha, \beta^{-1}\right) \propto L\left(Y/X, w, \beta^{-1}\right) \cdot P(w \mid \alpha)$$

We can thus determine w by looking for the most probable value of w given the observed data, *i.e.* by maximum a posteriori (MAP). By introducing the logarithm into formula (11), we obtain:

$$w^* = \ln\left[L\left(Y \mid X, w, \beta^{-1}\right)\right] P(w \mid \alpha)$$

$$w^* = \ln\left[L\left(Y \mid X, w, \beta^{-1}\right)\right] N\left(0, \alpha^{-1} I\right)$$

$$w^* = \ln\left[L\left(Y \mid X, w, \beta^{-1}\right)\right] + \ln\left(N\left(0, \alpha^{-1} I\right)\right)$$

$$w^* = -\frac{\beta}{2}\sum_{k=1}^{N}\left(y - g\left(x_k, w\right)\right)^2 + \frac{\alpha}{2} w^{\mathrm{T}} w$$

If we refer to Equation (1), Formula (13) becomes:

$$-\frac{\beta}{2}\sum_{k=1}^{N}\left(y - w^{\mathrm{T}}\phi\left(x_k\right)\right)^2 + \frac{\alpha}{2} w^{\mathrm{T}} w$$

With $\phi(x)$ a basis function that transforms the initial sample data so that a linear model can be used in the problem formulation. We also note that maximizing the posterior probability is equivalent to minimizing the sum of squared regularized errors.

This has the effect of reducing the phenomenon of overfitting observed in the frequentist approach. Probabilities thus prove to be a powerful tool for estimating the parameters of a linear model.

Formula (15) is used to make a point estimate of the w parameter values. It corresponds to the Ridge regression of L_2. When the error distribution follows a Laplace distribution in formula 14, we obtain the Ridge regression of L_1 [3]. But if we want to make an estimate in terms of an interval, we'll use Bayes' approach in all its completeness, taking into account the multiplication and sum rule in probability.

This involves representing (15) as a probability distribution. This distribution is the predictive distribution. Consider X and $\mathbf{y}$ the training data. Given a new point x, the task is to predict the value y. We need to formulate and calculate the expression $P(y \mid x, X, \mathbf{y})$. The parameters α and β are known in advance, as they can be calculated from the data.

The predictive probability distribution is given by applying Bayes' theorem using the multiplication and sum rule in probability theory:

$$P(y \mid x, X, \mathbf{y}) = \int P(y \mid x, w) P(w \mid X, \mathbf{y}) \mathrm{d}w \tag{12}$$

This expression can be calculated analytically by:

$$P(y \mid x, X, \mathbf{y}) = N\left(y \mid m(x), s^2(x)\right) \tag{13}$$

The probability distribution $P(y \mid x, X, \mathbf{y})$ is equivalent to $P(w \mid y)$.

$$m(x) = \beta \phi(x)^{\mathrm{T}} S \sum_{n=1}^{N} \phi(x_n) y_n \tag{14}$$

$$s^2(x) = \beta^{-1} + \phi(x)^{\mathrm{T}} S \phi(x) \tag{15}$$

With the matrix S equal to:

$$S^{-1} = \alpha I + \beta \sum_{n=1}^{N} \phi(x_n) \phi(x_n)^{\mathrm{T}} \tag{16}$$

Applying Bayes' approach to estimating the parameters of a linear model yields the same results as minimizing the variance of the regularized errors. But the Bayes approach has the advantage of using only the training sample [1]. Given a small sample size and prior knowledge of the parameter distribution, it gives excellent results. Minimization by the regularized error function requires another validation sample for hyper-parameter optimization [14].

In practice, the Monte Carlo Chain Markov algorithm and the variational inference algorithm, which we won't be presenting here, are available for estimating the probability a posteriori. These two algorithms can be used to approximate the integral of formula (16). In Python, they are contained in the pycm3 and Bayepy packages respectively.

3.4. Hypothesis and Limitations of Using the Bayesian Approach

The following assumptions must be met in order to use the Bayes approach:

- Probabilities are considered as measures of subjective belief or uncertainty.
- Before observing the data, the assumption of an a priori distribution on the parameters of the model.
- The assumption of a likelihood function that describes the probability of observing the data given the parameters.
- The assumption of an a posteriori distribution that incorporates both initial beliefs and information from the data.
- The assumption of updating beliefs: unlike the frequentist approach, which treats parameters as unknown constants, the Bayesian approach allows initial beliefs about model parameters to be updated in the light of observed data.

The results of the Bayesian approach can be sensitive to the choice of a priori distributions. Inappropriate a priori specifications can lead to biased estimates. This approach is also influenced by outliers if the a priori specifications are not well chosen.

3.5. Comparative Summary of the Two Approaches

Table 1 presents the advantages and disadvantages of the two approaches, after having elucidated the characteristics of each of them.

Table 1. Comparative summary of the two approaches.

	Advantages	Disadvantages
Bayesian approach	When data size is reduced [10]; When a priori knowledge of parameter distributions is available [11]; To quantify the uncertainty of estimates [8]; When data is heterogeneous or has complex structures; Suitable for e-learning that doesn't require full data storage [13].	Model inference can be time-consuming; The Bayesian method can lead to poor results if a wrong choice has been made on the a priori distribution.
Frequentist approach	When data size is large [9]; Suitable for batch learning [13]; When errors follow a normal distribution.	Overfitting in the case of small amounts of data; Based solely on data and not on other sources of knowledge; Makes a point estimate and does not quantify uncertainties in the estimated values.

4. Méthodologie

Predictive analysis is the final step in a modeling study. To get there, there are several other preliminary steps, such as cleaning, transformations, visualization, etc., which are carried out on the data [14]. For this reason, we will use a simulation example to randomly generate a sample of data with two input variables x_i and one output variable y. The two variables must follow different laws of probability for them to be decorrelated.

The values of y will be obtained from the assumption of a linear model existing between the latter and the input variables, to which a small random value will be added.

```
[language = Python]
np.random.seed (0)
X1 = np.linspace (0, 10, 500)
X2 = np.random.normal (2, 10, size = 500)
true_beta0 = 4
true_beta1 = 2
true_beta2 = 5
Y = true_beta0 + true_beta1 * X1 + true_beta2 * X2+ np.random.normal (0, 1, size = 500)
```

X = pandas. DataFrame ('X1': X1, 'X2':X2, columns = ['X1', 'X2'])

Then we'll proceed in two stages:

- First, we'll consider a small sample size and estimate the model parameters using both approaches;
- Secondly, we will consider a large sample size and estimate the model parameters using both approaches.

Each time, the calculated parameter values will be compared with the true known values, which are assumed to have generated the sample.

Consider the following model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \tag{17}$$

with $\beta_0 = 4, \beta_1 = 2, \beta_2 = 5$, $\varepsilon \sim \mathcal{N}(0,1)$.

Formula (21) can be implemented using previous python code and generates a sample of the data.

On the other hand, we'll take the same approach by considering the estimation of the function $\sin x + \varepsilon$ by a polynomial function, which is also a linear form with respect to its parameters. This example can also be found in [13].

5. Comparison of the Two Approaches

5.1. Results Presentation

We're going to estimate the parameters of a linear model using least squares on the one hand, and the Bayesian approach on the other. To do this, we have chosen to use the Python language. The sklearn package will be used to implement least squares, while the pymc3 and Bayespy packages will be used to implement the Bayesian approach.

Let's consider formula (21) and estimate the parameters. Here are the results:

- **Case 1, small sample size, with** $\beta_0 = 4$, $\beta_1 = 2$, $\beta_2 = 7$.

Table 2. Comparison results with reduced size.

	Least squares			Regression Ridge			MCMC			BayesPy		
Size	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$N = 5$	4.63	1.88	6.98	4.89	1.85	6.97	4.55	1.90	6.98	4.63	1.88	6.98
$N = 15$	3.83	2.01	7.02	3.98	1.99	7.01	3.84	2.01	7.02	3.83	2.01	7.02
$N = 30$	4.15	2.02	7.00	4.20	2.01	7.00	4.14	2.02	7.00	4.15	2.02	7.00

- **2nd case, large sample size, with** $\beta_0 = 4$, $\beta_1 = 2$, $\beta_2 = 7$.

Table 3. Comparison results with large size.

	Least square			Regression Ridge			MCMC			BayesPy		
Taille	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$N = 75$	4.04	1.99	7.01	4.06	1.98	7.00	4.04	1.99	7.01	4.04	1.99	7.01
$N = 120$	4.10	1.97	6.99	4.11	1.97	6.99	4.10	1.97	6.99	4.10	1.97	6.99
$N = 300$	4.18	1.97	6.99	4.18	1.97	6.98	4.18	1.97	6.99	4.18	1.97	4.18

Let's now consider the function $\sin x + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.3)$ having generated the data sample. This function can be estimated using a polynomial function of degree 3. This degree is obtained after a model selection step such as cross-validation, etc., which we won't go into here. We will now evaluate the different methods on the function $a_0 + a_1 x + a_2 x^2 + a_3 x^3$.

- **Case 1, small sample size**

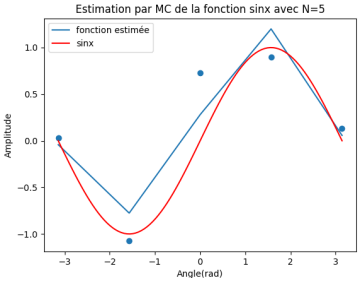
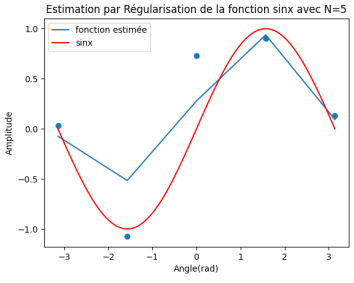
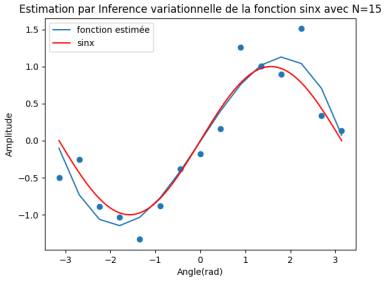
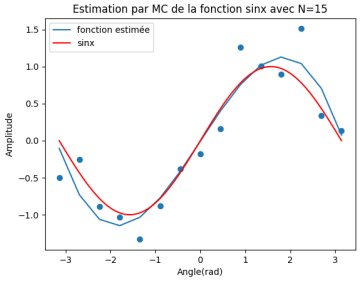
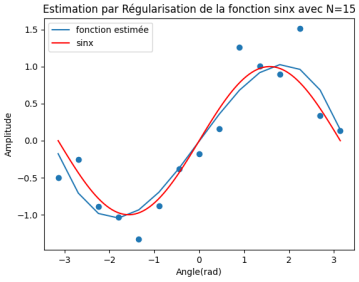
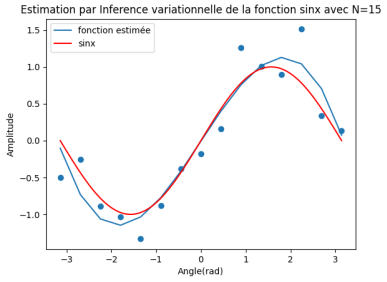
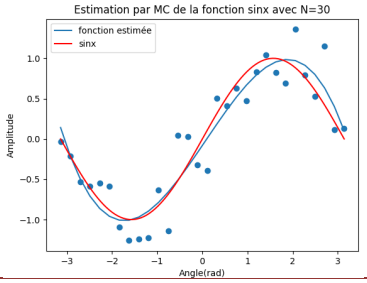
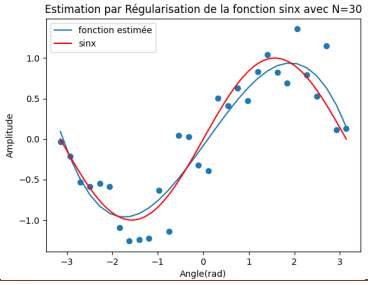
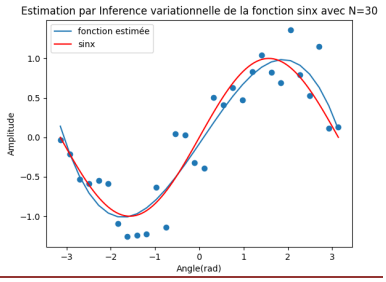
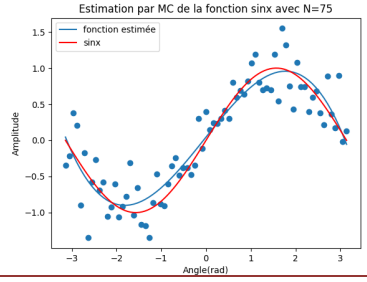
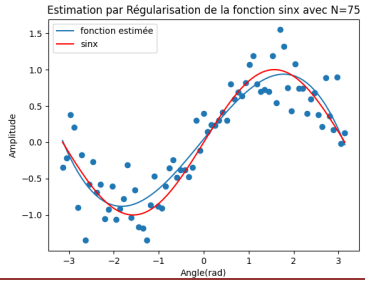
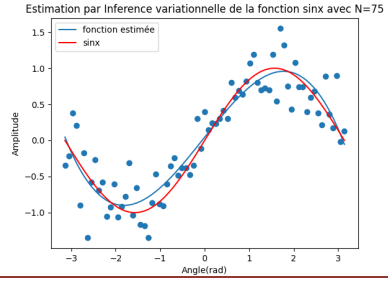
Size	Least square	Regression Ridge	BayesPy
$N = 5$			
$N = 15$			
$N = 30$			

Figure 1. Estimation of $\sin x$ by different approaches with reduced size.

- **2nd case, large sample size**

Size	Least square	Regression Ridge	BayesPy
$N = 75$			

Continued

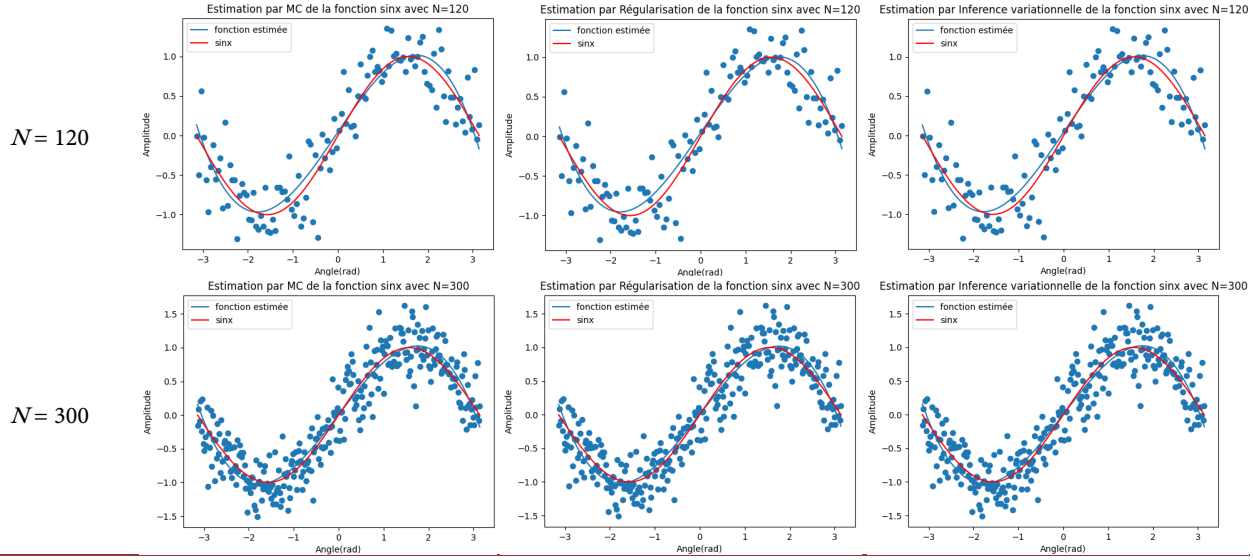


Figure 2. Estimation of $\sin x$ by different approaches with large size.

5.2. Interpretation of Results

- Looking at Table 2 and Figure 1, we see that as the data increases, the estimated values converge towards the true values for both approaches. But infinitely increasing the data, as shown in Table 3, does not significantly improve the model found.
- In the case of linear estimation, few data are needed to obtain a model close to reality. Indeed, for a problem involving the estimation of a line in a two-dimensional space, two points are needed to estimate the line. This means that at least three points would be sufficient to limit the overfitting effect.
- All methods have more or less similar results when it comes to linear estimation. In fact, a linear form does not have to be checked for variations in order to derive its true representative form from the data; hence the problem of overfitting will not be frequent in this case. We note in Table 2 that even with a size of 15 for our case, we reach the true values.

So the Bayesian approach doesn't seem to be very effective compared with least squares when faced with reduced data, since only the generated data doesn't have a complex structure. By contraposition, this may therefore confirm the hypothesis that the Bayesian approach is efficient in the face of reduced data that have a complex structure within them.

- In addition, when the linear model starts to become complex (piecewise linear), as is the case with the parameter linear model, $\sin x$ for example, the indefinite increase in data, as shown in Figure 2, tends to limit the ovefitting effect and improve the estimated model towards the true model.

6. Conclusions

In conclusion, it's important to recognize that both the probabilistic approach

and the least squares method are powerful and complementary tools in the design of predictive models when the relationship between features and goal is linear. The probabilistic approach takes a more theoretical and rigorous view, quantifying the uncertainty associated with predictions. It represents the theoretical framework that makes it possible to obtain algorithms for estimating the parameters of a model by initially considering certain hypotheses. Least squares, for example, are a method obtained by calculating maximum likelihood while assuming independence of observations and normality of errors between calculated and estimated model values.

From a technical point of view, the probabilistic approach requires a priori knowledge and sampling algorithms such as Monte Carlo Markov chain to calculate the a posteriori distribution. If a priori distribution is not known, these algorithms calculate the conjugates of a posteriori distribution, and this is fairly resource-intensive. In the case of linear predictive models, the least squares method stands out for its ease of application, making it a practically preferred choice. In our experience, both approaches lead to the same results when estimating the values of a linear model, assuming that we know the form of the estimated function. However, in most cases, this form will not be known initially. Other tasks may be useful beforehand, such as feature selection, data transformations, model selection, etc.

It is therefore recommended to use least squares even with little data when the form of the function to be estimated is linear and known. However, when the function to be estimated has variations, a large amount of data is required. If the form of the function is not known in advance and we have information on a priori distribution of the parameters, the probabilistic approach will be preferred. If the form of the function to be estimated is not known in advance, and the probability distribution of the parameters is not known, it will be important to consider other techniques, such as neural networks, or to carry out certain processing operations on the data beforehand, if the linear model obtained is not satisfactory.

Looking ahead, given that numerical methods for calculating a posteriori probability are based on the assumption of knowledge of the relationship between the features and the phenomenon to be predicted, it will also be interesting to study how the Bayesian approach can be integrated into the retroproputation algorithm to solve non-linear problems. On the other hand, a more in-depth study based on real data will be possible, as the artificially created data fulfilled all the assumptions for the use of least squares. This is why the two approaches led to almost similar results. This is just further evidence that the probabilistic approach provides a general framework for generating model design algorithms.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bellanger, L. and Tomassone, R. (2014) Exploration de données et Méthodes statistiques: Data analysis & data mining avec R of Références sciences. Ellipses.

[2] Samueli, J.-J. (2010) Legendre et la méthode des moindres carrés. *Bibnum. Textes fondateurs de la science*. https://doi.org/10.4000/bibnum.580

[3] Gaillard, P. (2019) Régression linéaire. *Consommation* (*GW*), **40**, 50.

[4] Maddi, A., Guessoum, A., Berkani, D. and Belkina, O. (2005) Etude de la méthode des moindres carrés récursive et application au signal de parole. 3*rd International Conference*: *Sciences of Electronic*, *Technologies of Information and Telecommunications*, Sousse, Tunisia, 27-31 March 2005.

[5] Lay, D.C. (2012) Algèbre linéaire et applications. Pearson.

[6] James, G., Witten, D., Hastie, T., Tibshirani, R., *et al*. (2013) An Introduction to Statistical Learning. Vol. 112, Springer. https://doi.org/10.1007/978-1-4614-7138-7

[7] Huber, C. and Nikulin, M.S. (1997) Remarques sur le maximum de vraisemblance. Qüestiió: Quaderns d'estadstica i investigació operativa. https://www.raco.cat/index.php/Questiio/article/download/26828/26662

[8] Lambert, M. (*2020*) Quantification de l'incertitude pour l'apprentissage automatique. *Actes de la conférence* CAID, Rennes, 18-19 November 2020, 110.

[9] Dreyfus, G. (2008) Apprentissage statistique. Editions Eyrolles.

[10] Robert, C. (2005) Le choix bayésien: Principes et pratique. Springer Science & Business Media.

[11] Kuma, J.K. (2019) Estimation par la méthode du Maximum de Vraisemblance: Éléments de Théorie et pratiques sur Logiciel. https://hal.science/cel-02189969/document

[12] Deudon, M. and Servajean, M. (2020) Régularisation et optimisation des modeles. https://www.mtpcours.fr/u/TW233MI-Optimisation-regularisation.pdf

[13] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer.

[14] Cornuéjols, A. and Miclet, L. (2011) Apprentissage artificiel: Concepts et algorithmes. Editions Eyrolles.