Research & Reviews: Journal of of Statistics and Mathematical Sciences

The Robust Bayesian Approach to the Model Selection Algorithm

Krzysztof W Fornalski[1]* and Ludwik Dobrzyński[2]

[1]PGE EJ 1, Warszawa, Poland

[2]National Centre for Nuclear Research (NCBJ), Otwock-Świerk, Poland

Research Article

Received date: 22/06/2015

Accepted date: 04/08/2015

Published date: 26/10/2015

***For Correspondence**

Krzysztof W. Fornalski, PGE EJ 1, ul. Mysia 2, 00-496 Warszawa, Poland

E-mail: krzysztof.fornalski@gmail.com

Keywords: Model selection, Bayesian, Robust analysis, Regression, Bayesian analysis.

ABSTRACT

The paper presents the robust Bayesian approach to the experimental data analysis. Firstly, the algorithm of the curve fitting to data points is introduced. The method is based on the robust Bayesian regression analysis, which substantially reduces the role of outlying data points (outliers). In the second part the Bayesian model selection algorithm was presented. Finally, the exemplary applications of both methods are discussed.

INTRODUCTION

The nature of experimental data consists of several elements, e.g. the data points are subjected to large uncertainties and/or large scatter. Therefore finding the most reliable curve describing data points may be not a trivial task. However, once such a goal is achieved, the scientists can come up with general conclusions or theories about the examined object or phenomenon.

The form of the fitted curve corresponds as a rule to the theory aimed at describing the data. However, the scatter and uncertainties of experimental data points imply that many potential models can describe those data equally well. Thus one has to decide which model (amongst other good ones) is the most reliable one in light of the analyzed data.

This paper discusses the robust Bayesian approach to two important elements of data analysis: fitting of a curve to the data (regression analysis) and the model selection algorithm.

ROBUST CURVE FITTING

The robust data fit (or robust regression analysis) is a part of what is known as the robust statistics, representing statistical methods that are not unduly affected by outliers [1]. In this case an outlier is a data point which significantly deviates from the main trend of points, e.g. due to a strong systematic error. The main goal of all robust analyses is to find the best fit of the curve to existed data points, which is least sensitive to such potential outliers.

Simple comparison between the results obtained using robust Bayesian and classical least squares methods [2] is presented in **Figure 1.** One can clearly see that outliers in classical least squares method would result in very misleading conclusion, while Bayesian fit clearly reproduces the main trend as if the outliers were not present in the data. This results from the fact that each *i*-th data point appears with the probability whose probability density function (PDF), P_i, is composed of a proper Gaussian distribution (so called likelihood function) around its expected value as well as the prior function for its probability σ_i [3]. If one suspects that the real uncertainty may be larger than originally quoted one, σ_{0i} , one can choose such a prior as σ_{0i} / σ_i^2 , which finally results in:

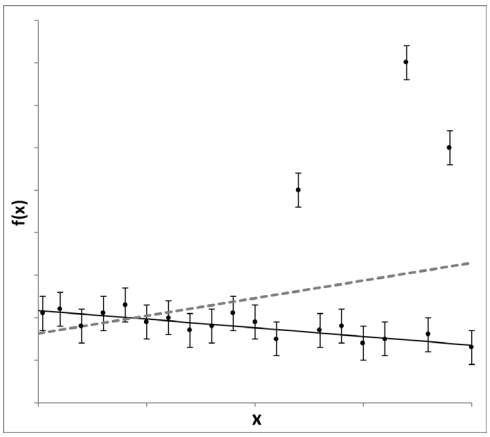

Figure 1. The example of robust Bayesian (black solid line) and least squares (grey dashed line) fits to some virtual experimental data with three outliers (outstanding points) [4,6].

$$P_i = \int_{\sigma_{0i}}^{\infty} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(M_i - D_i)^2}{2\sigma_i^2}} \times \frac{\sigma_{0i}}{\sigma_i^2} d\sigma_i \tag{1}$$

The symbol D_i corresponds to the *i*-th experimental data value, where M_i is a model (theoretical) curve, e.g. the polynomial $M_i=\lambda_1+\lambda_2 x_i+\lambda_3 x_i^2+...+\lambda_n x_i^{n-1}$, where x_i is the *i*-th independent variable, and λ_i denotes *i*-th parameter. The right-side prior function for σ_i in eq. (1) assumes that the *i*-th analyzed probability σ_i lies between the original one (σ_{0i}) and infinity [3,4], as mentioned earlier. Its particular choice presented above leads to closed form solution. As shown by Sivia and Skilling [3] this form may be different, however, it does not affect final conclusions. The procedure makes the weights of all outliers insignificantly low in calculation of the posterior probability distribution *P* for measuring of all *N* points, where, according to the maximum likelihood method [4,5] one can also use a sum instead of a product:

$$P = \prod P_i \Leftrightarrow S = \sum \ln P_i \tag{2}$$

where P_i is a result of the integration of eq. (1) for single point *i*:

$$P_i = \frac{\sigma_{0i}}{(M_i - D_i)^2 \sqrt{2\pi}} \left[1 - \exp\left(\frac{-(M_i - D_i)^2}{2\sigma_{0i}^2} \right) \right] \tag{3}$$

After the differentiation of logarithmic probability S over all *n* fitting parameters $\lambda=\{\lambda_1, \lambda_2, ..., \lambda_n\}$, one can find the final and general form of a Bayesian fitting equation [3]:

$$\frac{dS}{d\lambda} = \sum_{i=1}^{N} g_i (M_i - D_i) \frac{dM_i}{d\lambda} \equiv 0 \tag{4}$$

where aforementioned weights g_i of the points are:

$$g_i = \frac{1}{(M_i - D_i)^2} \left[2 - \frac{(M_i - D_i)^2}{\sigma_{0i}^2} \times \frac{1}{\exp\left(\frac{(M_i - D_i)^2}{2\sigma_{0i}^2} \right) - 1} \right] \tag{5}$$

The equation (4) can be implemented directly into the computational algorithm to find the best Bayesian fit to all *N* experimental data points (x_i,D_i) with vertical uncertainties σ_{0i} each. As mentioned earlier, the M_i means the theoretical value predicted by the assumed model and the best fitted *n* parameters, $M_i(\lambda_1, \lambda_2, ..., \lambda_n)$. Just this technique was used in fitting of a linear function $M_i=\lambda_1+\lambda_2 x_i$ presented in **Figure 1**.

This method, presented by Sivia and Skilling [3] and not often used in the studies of e.g. epidemiological data, was extensively used in our papers [4-10].

MODEL SELECTION ALGORITHM

The Bayesian analysis allows one to assess relative reliability of two chosen models that can describe the data (the classical methods of model selection, like AIC[1], or other Bayesian ones, like BIC[2], will be omitted here). Examples of the use of such analysis can be found [4-10].

[1]Akaike Information Criterion
[2]Bayesian Information Criterion

The Bayes theorem connects the probabilities of *P(Model|Data)~P(Data|Model)*, which can be used to estimate the relative reliability of two models, *M*s, in the case of the same data, *D*. The reliability of a model *M* with the fitting parameter *λ*, using the marginalization procedure can be written as [3]:

$$P(M|D) \propto P(D|M) = \int P(D,\lambda|M)\,d\lambda = \int P(D|\lambda,M) \times P(\lambda|M)\,d\lambda \tag{6}$$

The *P(D|λ,M)* corresponds to the likelihood function, represented by the Gaussian distribution around the expected value $\lambda_0 \pm \sigma_\lambda$ with maximum probability of likelihood function equals $P(D|\lambda_0, M)$. The prior probability *P(λ|M)* can be assumed as a uniform distribution $U(\lambda_{min}, \lambda_{max})$. Because such form of *P(λ|M)* is independent of *λ*, the integral (6) can be written as [3]:

$$P(D|M) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} P(D|\lambda_0, M)\, e^{-\frac{(\lambda-\lambda_0)^2}{2\sigma_\lambda^2}}\, d\lambda \tag{7}$$

The result of the integral (7) can be approximated by $P(D|\lambda_0, M)\sigma_\lambda\sqrt{2\pi}$ because "*the sharp cut-offs at λ_{min} and λ_{max} do not cause a significant truncate on of the Gaussian*" probability distribution from eq. (7) [3]. Because λ_0 corresponds to the parameter found by the robust Bayesian best fit method for model *M* (see eq. (4)), the maximum value of likelihood function $P(D|\lambda_0, M)$ can be replaced by the set of P_i given by eq. (1) or (3) and the final form of the reliability function can be approximated by [7-8]:

$$P(M|D) \propto P(D|M) \approx \sum P_i \times \frac{\sigma_\lambda\sqrt{2\pi}}{\lambda_{max} - \lambda_{min}} \tag{8}$$

The right-hand term in eq. (8) is called an Ockham factor. Equation (8) corresponds to the situation, where model *M* has only one (*n=1*) fitting parameter, $\lambda_0 \pm \sigma_\lambda$. In the case of a model which contains no parameters the Ockham factor equals 1 and model *M* is just a constant value (M_i=*const*). Thus, if such a model describes the data equally well as a model containing one parameter, the Ockham factor will always favor the former one.

In the case of *n* fitting parameters $\lambda=\{\lambda_1, \lambda_2, ..., \lambda_n\}$ with their estimated uncertainties $\sigma_\lambda=\{\sigma_1, \sigma_2, ..., \sigma_n\}$, the most general form of eq. (8) can be presented as [6,8]:

$$P(M|D) \propto \sum_{i=1}^{N} \frac{1}{(M_i - D_i)^2}\left[1 - \exp\left(-\frac{(M_i - D_i)^2}{2\sigma_{0i}^2}\right)\right] \times \prod_{j=1}^{n} \frac{\sigma_{\lambda_j}\sqrt{2\pi}}{(\lambda_{max} - \lambda_{min})_j} \tag{9}$$

Let us recall that *N* represents the number of experimental points (x_i, D_i) with "vertical" uncertainties σ_{0i} each, to which model *M* is fitted using *n* fitting parameters $\lambda \pm \sigma_\lambda$. The most problematic is the choice of values λ_{min} and λ_{max}, for all *λ*s. In the simplest case they can be taken as minimum/maximum possible values of the considered parameter *λ* using the largest span that can be tolerated by the data. In order not to extend the range of *λ* in the case of a substantial scatter of data, not more than e.g. three points are allowed to lie outside the range [8].

In the final step of analysis one can calculate the relative value of each two models, say *A* and *B*, to check which of them is more likely to describe the data:

$$W_M = \frac{P(M = A \mid D)}{P(M = B \mid D)} \tag{10}$$

When W_M is greater than 1, model *A* wins over *B*. When $W_M \approx 1$, both models have the same degree of belief. In general, W_M can quantify the preference of one model with respect to the other one. In practice, the real values of W_M may show that the plausibility of models can differ by orders of magnitude Fornalski and Dobrzyński [8].

DISCUSSION AND PRACTICAL APPLICATIONS

The Bayesian model selection algorithm may be applied in all cases when the clear distinction between the models is hard to obtain or even impossible. A good example of that process is presented in **Figure 2** where both linear and parabolic models seem to be appropriate ones. In considered situation the eq. (10) is described by [5] :

$$W_m = \frac{\sum_{i=1}^{N} \frac{1}{R_{Ai}^2}\left[1 - \exp\left(-\frac{R_{Ai}^2}{2\sigma_{0i}^2}\right)\right]}{\sum_{i=1}^{N} \frac{1}{R_{Bi}^2}\left[1 - \exp\left(-\frac{R_{Bi}^2}{2\sigma_{0i}^2}\right)\right]} \cdot \frac{\frac{a_{\max}^{(B)} - a_{\min}^{(B)}}{\sigma_a^{(B)}\sqrt{2\pi}} \cdot \frac{b_{\max}^{(B)} - b_{\min}^{(B)}}{\sigma_b^{(B)}\sqrt{2\pi}} \cdot \frac{c_{\max}^{(B)} - c_{\min}^{(B)}}{\sigma_c^{(B)}\sqrt{2\pi}}}{\frac{a_{\max}^{(A)} - a_{\min}^{(A)}}{\sigma_a^{(A)}\sqrt{2\pi}} \cdot \frac{b_{\max}^{(A)} - b_{\min}^{(A)}}{\sigma_b^{(A)}\sqrt{2\pi}}} \tag{11}$$

where the straight line and parabola are implemented as $R_{Ai}=a^{(A)}x_i+b^{(A)} - D_i$ and $R_{Bi}=a^{(B)}x_i^2+b^{(B)}x_i+c^{(B)} - D_i$, respectively. In this ituation the value of $W_m \approx 30$, which means that model *A* (line) is approximately 30 times more likely than *B* (parabola) **(Figure 2)**.

In fact, the r.h.s. of eq. (11) could be also multiplied by $W_0=P_0(A)/P_0(B)$ – the ratio of prior beliefs in a model *A* with respect

to the model *B*. For example, if one knows that linear relation should not describe the data, such extra factor stemming from Bayesian analysis will change aforementioned conclusion. However, in the example above it was assumed, that $W_0=1$.

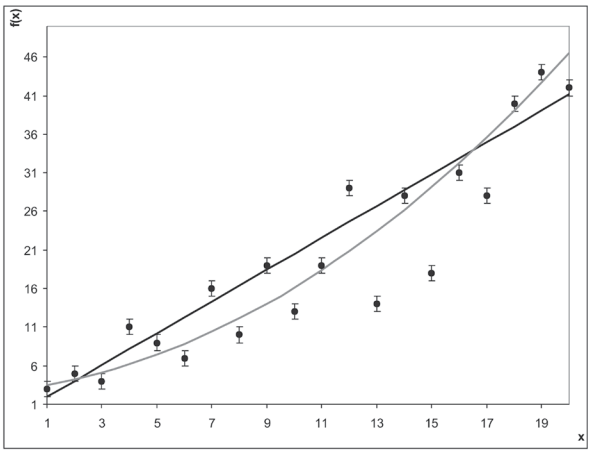

Figure 2. The fitting of two models: linear (black line) and parabolic one (grey line) to widely scattered data. The model selection algorithm prefers the linear model, $W_m \approx 30$ [5].

Presented example shows a study popular in various sciences. As mentioned earlier, the model selection fitting the existing data is a matter of great importance. However, the most common approach concerns fitting the model to one's assumptions, irrespective of its real plausibility. For instance the parabolic relation may be the outcome of some physical or biological theory, tested by the experiment (data source). On the other hand, in the absence of a strong information, giving no preference to any model ($W_0=1$) is sensible. In the example given in **Figure 2** the linear model turns out to be sufficient for such the description of widely scattered data.

Similar case is presented in **Figure 3**, where one can observe the straight line and the part of the sinusoid. The sinusoid function is some theory's outcome (e.g. mechanical oscillations) where the linear fitting is the simplest approximation of widely scattered data in some very limited range. In this example the robust statistics with $W_0=1$ results in the linear relation as the most probable fitting. This is rather not consistent with the assumed theory and shows the deficiency of not using prior preference to the sinusoidal model.

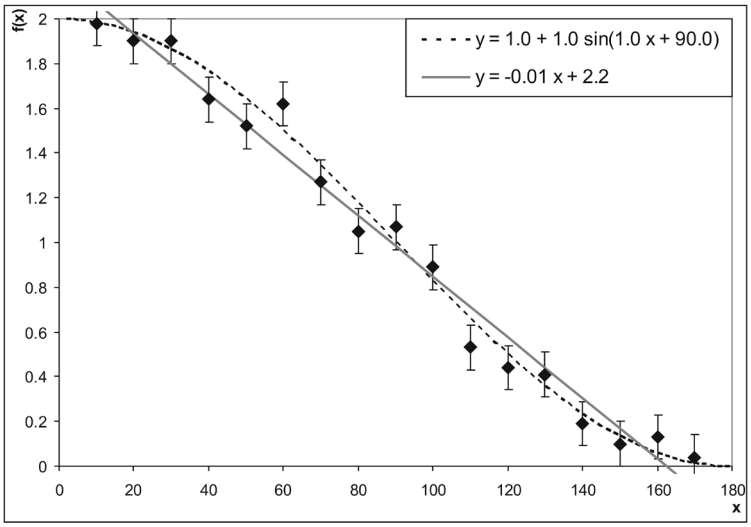

Figure 3. The linear fit (solid line) as a good approximation of sinusoid relationship (dashed line) in the light of scattered data, $W_m \approx 5$ [5].

Very important and real case is presented in **Figure 4**, where one can find the actual physical description of the observed phenomenon (here: the cancer mortality ratio among irradiated nuclear workers). The precise model concerns the human health, which is a matter of great importance. However, in many regulations, e.g. radiation protection standards, some initial assumption is taken into account. In global standards more often the linear relationship is used, which comes from mentioned assumption, that all radiation is harmful. However, from the purely statistical point of view, such statement is nothing more than non-mathematical degree-of-belief in the existing data. Thus, similarly to the previous example, the simpler model (here: the constant value, M=const) is more likely to fit than the linear relationship [9]. This is because of the fact, that the huge scatter of points makes more complicated models an assumptions only, not real effects. Analogical case can be found when cancer mortality in natural background radiation is analyzed using Bayesian reasoning [10].

The Bayesian model selection algorithm can be applied to various cases, whenever one needs to choose the most proper curve. The latest application of that method was introduced in the cytogenetic biological dosimetry [11] to find the best calibration curve for chromosomal aberration testing of incidentally irradiated people. Nevertheless, it is just another example, not the complete range of possible applications.

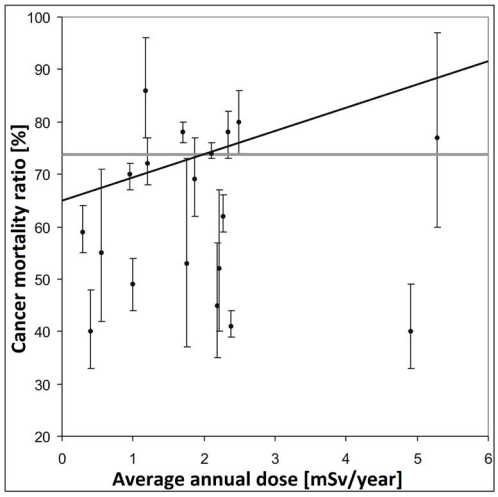

Figure 4. The algorithm' application to the cancer mortality ratio of irradiated nuclear workers [9]. The one-parametric constant model (grey horizontal line) is more likely than a linear one (black line), $W_m \approx 2$.

Concluding the paper, robust Bayesian regression method (curve to data fitting) and the model selection algorithm were described and illustrated by some examples. Both methods work well in the situation of duff data, and can be recommended for use in such cases.

REFERENCES

1. Box GEP and Tiao GC. A Bayesian approach to some outlier problems. Biometrika 1968;55:119-129.
2. Wolberg J. Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments. 2005. Springer.
3. Sivia DS and Skilling J. Data Analysis. A Bayesian Tutorial second edition. 2006. Oxford University Press.
4. Fornalski KW, et al. Application of Bayesian reasoning and the Maximum Entropy Method to some reconstruction problems. Acta Physica Polonica A. 2010;117:892-899.
5. Fornalski KW and Dobrzyński L. Zastosowania twierdzenia Bayesa do analizy niepewnych danych doświadczalnych (in Polish). Postępy Fizyki. 2010;61:178-192.
6. Fornalski KW. Alternative statistical methods for cytogenetic radiation biological dosimetry. Cornell University Library. arXiv.org/abs/1412.2048, 2014.
7. Fornalski KW. Pooled Bayesian meta-analysis of two Polish studies on radiation-induced cancers. Radiation Protection Dosimetry. 2015.
8. Fornalski KW and Dobrzyński L. Pooled Bayesian analysis of twenty-eight studies on radon induced lung cancers. Health Physics. 2011;101:265-273.
9. Fornalski KW and Dobrzyński L. Ionizing radiation and the health of nuclear industry workers. International Journal of Low Radiation. 2009;6:57-78.
10. Dobrzyński L, et al. Cancer mortality among people living in areas with various levels of natural background radiation. Dose-Response. 2015;13:1-10.
11. Pacyniak I, et al. Employment of Bayesian and Monte Carlo methods for biological dose assessment following accidental overexposures of people to nuclear reactor radiation. In proceeding of: The Second International Conference on Radiation and Dosimetry in Various Fields of Research RAD 2014, University of Nis Serbia, 27-30.05. 2014:49-52