Measurement 55 (2014) 564–570

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Bayesian estimate of the degree of a polynomial given a noisy data sample

G. Mana *, P.A. Giuliano Albo, S. Lago

INRIM – Istituto Nazionale di Ricerca Metrologica, Str. delle Cacce 91, 10135 Torino, Italy

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 5 May 2014
Accepted 23 May 2014
Available online 19 June 2014

Keywords:
Data analysis
Regression analysis
Uncertainty budget
Fitting by polynomials
Uncertainty in measurments

ABSTRACT

Regression analysis is widely used to create continuous representations of discrete datasets. When the regression model is not based on the physics underlying the data, heuristic models play a crucial role and the model choice affects the data analysis. This paper identifies the most appropriate model in terms of Bayesian selection. The result is applied to two practical examples, one of which is taken from a problem of chemical thermodynamics.

1. Introduction

A problem of regression analysis is to determine how many basis functions to include in the data model. Examples are finding calibration curves and polynomial representations of thermodynamic equations of state [1–3]. Any set of basis functions can be considered; when they are polynomials, the problem is determining the degree of the regression.

A maximum likelihood approach leads to the highest number of basis functions and is not the right choice. Many authors considered this problem in different statistical settings; their investigations led to a number of proposals [4–8]. An information theoretic solution is the Akaike criterion. It minimises the Kullback–Leibler distance between the model and the process that generated the data and carries out a trade-off between the data likelihood and the number of free parameters [9,10]. A tutorial paper on Bayesian reasoning by Gull [11] hides an original and undeservedly neglected proposal, where the idea is to calculate and to compare the odds that each model is true, given the data and any available prior information.

In order to bring the Gull's result to the metrologist's attention, we reassess his work and make clear its usefulness in selecting among different models. Although the main idea and tools are not new, we built on the Gull's work and deliver three additional results. Firstly, by slightly changing the parametrisation, we obtain an exact expression of the model evidence – the basic ingredient to calculate the model odds. Secondly, we explicitly demonstrate the evidence invariance and asymptotic properties and that our exact expression reduces to the Gull's approximate one. The role of the data zero-offset is also clarified. Thirdly, we use the evidence to consistently include the model uncertainty in the error budget.

This results help to solve partial differential equations by polynomials [2,3]. In this case, different choices of the polynomial degree lead to different sets of coefficients and, consequently, to different solutions. The availability of an exact criterion based on the probability calculus allows arbitrary choices – driven by the residuals analysis – to be avoided. To illustrate these concepts, we show how

* Corresponding author. Tel.: +39 011 3919728.
E-mail addresses: g.mana@inrim.it (G. Mana), a.albo@inrim.it (P.A. Giuliano Albo), s.lago@inrim.it (S. Lago).

http://dx.doi.org/10.1016/j.measurement.2014.05.037
0263-2241/© 2014 Elsevier Ltd. All rights reserved.

to determine the set of basis functions that best fits the measured values of the speed of sound in acetone, as a function of the temperature and pressure.

2. Problem statement

We represent the $\boldsymbol{y} = [y_1, y_2, \ldots y_N]^{\mathrm{T}}$ measurement results by the linear model

$$\boldsymbol{y} = W\boldsymbol{a} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots \epsilon_N]^{\mathrm{T}}$ are additive uncorrelated Gaussian errors having unknown variance σ^2 and zero mean, $\boldsymbol{a} = [a_0, a_1, \ldots a_{l-1}]^{\mathrm{T}}$ are model parameters, W is a $N \times l$ matrix explaining the data, $W_{nm} = w_m(x_n)$, and $\{w_0(x), w_1(x), \ldots w_{l-1}(x)\}$ is a set of l basis functions. The basis functions may be polynomials, for instance, $w_m(x) = x^m$, but, in general, they are any set of linearly independent functions. The problem is to find the set of basis functions most supported by the data; when they are polynomials, this corresponds to find the optimal degree of the regression. The interpretative model of the data is summarised by the matrix W; therefore, the problem is equivalent to find – within a set of matrices explaining the data – the one most supported by the data.

Following [11], we assume that $\langle \boldsymbol{y} \rangle = 0$, where the angle brackets indicate the unconditioned mean of the data, and introduce an additional model parameter by writing $C_{yy} = \beta^2 \mathbb{1}$, where C_{yy} is the unconditioned covariance of the data, $\mathbb{1}$ is the unit matrix, and $\beta > \sigma$. As shown in Appendix A, these assumptions are central to assign pre-data probabilities to the possible values of the model parameters. It is worth to notice that unconditioned means that $\langle \boldsymbol{y} \rangle$ and $C_{yy} = \langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \rangle$ are relevant to the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{a}$. In addition, we assume that all the Ws have equal a priori probability and that there is no prior information about the standard deviations σ and β.

3. Bayesian inferences

The Bayesian approach assigns a prior probability distribution to the parameters of each model and prior odds to the models. This allows a joint distribution of the data, parameters, and models to be written by embedding the separate distributions within a single one. By assigning the same chance to each, the probability of each model to explain the data is proportional to the odds of the observed data given the model, no matter what the values of the model parameters may be [12–14]. In turn, it is proportional to the normalising factor of the likelihood of the model parameters times the probability distribution synthesising the information about the parameter values before the measurement results are available.

To steer the calculation, we determine the post-data probability density, $P(\boldsymbol{a}, \beta, \sigma | \boldsymbol{y}, W)$, of the parameters of each model (which parameters include the unknown standard deviations β and σ) given the data $\boldsymbol{y}$ and the data-explaining matrix W. With a somewhat incongruous use of notation, from now on we will use the same symbols to indicate both the random quantities and their possible values. The post-data probability density is found via the product rule of probabilities,

$$P(\boldsymbol{a}, \beta, \sigma | \boldsymbol{y}, W) Z(\boldsymbol{y}|W) = N_N(\boldsymbol{y}|\boldsymbol{a}, \sigma, W)\pi(\boldsymbol{a}, \beta, \sigma|W), \tag{2}$$

where the N-dimensional Gaussian function

$$N_N(\boldsymbol{y}|\boldsymbol{a}, \sigma, W) = \frac{1}{\sqrt{(2\pi)^N}\,\sigma^N} \exp\left(-\frac{|\boldsymbol{y} - W\boldsymbol{a}|^2}{2\sigma^2} \right) \tag{3}$$

is the likelihood of the $\boldsymbol{a}$ and σ parameters, $\pi(\boldsymbol{a}, \beta, \sigma|W)$ is the pre-data probability density of the model parameters, the normalisation factor of $N_N(\boldsymbol{y}|\boldsymbol{a}, \sigma, W)\pi(\boldsymbol{a}, \beta, \sigma|W)$,

$$Z(\boldsymbol{y}|W) = \int_\Gamma N_N(\boldsymbol{y}|\boldsymbol{a}, \sigma, W)\pi(\boldsymbol{a}, \beta, \sigma|W)\, \mathrm{d}\boldsymbol{a}\, \mathrm{d}\beta\, \mathrm{d}\sigma, \tag{4}$$

is named model evidence, and the integration is carried out over the hypervolume Γ associated to the possible $\boldsymbol{a}$, β, and σ values.

Next, we observe that, according to (2), $Z(\boldsymbol{y}|W)$ is the probability density of the data given W – whatever the values of $\boldsymbol{a}$ and σ may be. Hence, by applying again the product rule of probabilities to the $\{W, \boldsymbol{y}\}$ pair, the post-data model-probability is [15]

$$\mathrm{Prob}(W|\boldsymbol{y}) = \frac{Z(\boldsymbol{y}|W)}{\sum_W Z(\boldsymbol{y}|W)}, \tag{5}$$

where we assigned the same prior probability to each model, the denominator is the normalisation factor of $Z(\boldsymbol{y}|W)$, and the sum extends to all the models. Therefore, to solve the stated problem, the calculation of the evidence (4) is central.

4. Pre-data distribution

To set the pre-data distribution $\pi(\boldsymbol{a}, \beta, \sigma|W)$ we assume that $\boldsymbol{a}$ is independent of σ. Hence,

$$\pi(\boldsymbol{a}, \beta, \sigma|W) = \pi_a(\boldsymbol{a}|W, \beta)\pi_\beta(\beta|\sigma)\pi_\sigma(\sigma). \tag{6}$$

As to σ and β, since there is no prior information, we use the improper Jeffreys distributions [16]

$$\pi_\sigma(\sigma) = 1/\sigma \tag{7a}$$
$$\pi_\beta(\beta|\sigma) = \vartheta(\beta - \sigma)/\beta, \tag{7b}$$

where $\vartheta(z)$ is the Heaviside function, which are invariant for a change of the measurement unit of the data. As for the $\boldsymbol{a}$ parameters, the $\langle \boldsymbol{y} \rangle = 0$ and $C_{yy} = \beta^2 \mathbb{1}$ constraints dictate

$$\pi(\boldsymbol{a}|W, \beta) = \sqrt{\frac{\det(W^{\mathrm{T}}W)}{(2\pi)^l (\beta^2 - \sigma^2)^l}} \exp\left[-\frac{\boldsymbol{a}^{\mathrm{T}} W^{\mathrm{T}} W \boldsymbol{a}}{2(\beta^2 - \sigma^2)} \right]. \tag{8}$$

The detailed derivation of (8) is given in Appendix A.

In general, the use of improper priors – like $\pi_\sigma(\sigma)$ and $\pi_\beta(\beta|\sigma)$ – must be avoided, because, in such a case, the model evidence (4) is defined only up to unknown scale factors. However, since in this case the same factor is included in all the evidences, this does not jeopardise the model comparison.