11-1-2021

How Probabilistic Methods for Data Fitting Deal with Interval Uncertainty: A More Realistic Analysis

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Sergey P. Shary
Novosibirsk University, shary@ict.nsc.ru

Comments:
Technical Report: UTEP-CS-21-102

How Probabilistic Methods for Data Fitting Deal with Interval Uncertainty: A More Realistic Analysis*

Vladik Kreinovich[1] and Sergey P. Shary[2]
[1]Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA
[2]Novosibirsk University, Novosibirsk, Russia
vladik@utep.edu,shary@ict.nsc.ru

Abstract

In our previous paper, we showed that a simplified probabilistic approach to interval uncertainty leads to the known notion of a united solution set. In this paper, we show that a more realistic probabilistic analysis of data fitting under interval uncertainty leads to another known notion – the notion of a tolerable solution set. Thus, the notion of a tolerance solution set also has a clear probabilistic interpretation. Good news is that, in contrast to the united solution set whose computation is, in general, NP-hard, the tolerable solution set can be computed by a feasible algorithm.

Keywords: Interval uncertainty, united solution set, tolerable solution set, probabilistic uncertainty
AMS subject classifications: 65G20, 65G30, 65G40

1 General motivation

When processing data, most practitioners use probabilistic methods. It is therefore desirable to study how, for the case of interval uncertainty, these methods compare with interval techniques; see, e.g., [1, 5, 6, 7].

2 Data fitting problem

In many situations:

- we know the general form $y = F(x, c)$ of the dependence of a quantity y on quantities $x = (x_1, \ldots, x_n)$, but
- we do not know the exact values of the parameters $c = (c_1, \ldots, c_m)$.

*Submitted: November 30, 2021; Revised: ?; Accepted: ?.

Let us have a few example.

- We may have a linear dependence

$$y = c_1 \cdot x_1 + \ldots + c_n \cdot x_n + c_{n+1}.$$

- We may have a general quadratic dependence.
- For a radioactive delay, we have a linear combination of exponentially decreasing terms

$$y = \sum_{i=1}^{p} c_{2i-1} \cdot \exp(-c_{2i} \cdot t),$$

  etc.

In all these cases, the values c_i must be determined from the measurement results.

For this purpose, several (K) times, we measure x_i and y. Based on the measurement results $\widetilde{x}_k = (\widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn})$ and $\widetilde{y}_k$, we need to estimate the values of the parameters that fit the data. This problem is also called *problem of parameter estimation.*

3 Need to take measurement uncertainty into account

Measurements are never absolutely accurate. Because of this, we need to take into account that the measurement results $\widetilde{v}$ are, in general, different from the actual (unknown) values of the corresponding quantity v, i.e., that there is a non-zero measurement error $\Delta v := \widetilde{v} - v$; see, e.g., [7].

It is important to take the corresponding measurement uncertainty into account when estimating the values of the parameters c_i.

4 Situations when we know the probability distributions

In many cases, we know the probability distributions $f_i(\Delta x_i)$ and $f(\Delta y)$ of the measurement errors, and the measurement errors corresponding to different distributions are independent.

In this case, we can use the Maximum Likelihood (ML) approach; see, e.g., [9]. This means that we select the *most probable* values c (and x_{ki}), i.e., the values for which the corresponding probability – which is equal to the product

$$\prod_{k=1}^{K} \Big(f(\widetilde{y}_k - F(x_k, c)) \cdot \prod_{i=1}^{n} f_i(\widetilde{x}_{ki} - x_{ki}) \Big),$$

attains its largest possible value.

Usually, instead of maximizing the likelihood, we solve the equivalent problem of maximizing the logarithm of the likelihood – which is known as *log-likelihood.* This reduction often simplifies the computations – e.g., for the Gaussian distribution, logarithm is an easy-to-maximize quadratic function.

5 Interval uncertainty

In many practical situations, we do not know the probability distributions, all we know is that the measurement errors Δv are located on the given interval $[-\Delta_v, \Delta_v]$; see, e.g., [1, 5, 6, 7].

In such situations, a usual probabilistic approach is to select, on this interval, the distribution with maximal entropy. This turns out to be the uniform distribution; see, e.g., [2].

6 Simplest case

The simplest – and rather frequent – case is when the values x_i are measured very accurately. In this case, we can safely ignore the corresponding measurement errors and conclude that $\widetilde{x}_{ik} = x_{ik}$ for all i and k. In this case, the ML approach selects the folllowing set:

The set of all possible values c for which, for all k, we have
$F(x_k, c) \in [\widetilde{y}_k - \Delta_y, \widetilde{y}_k + \Delta_y]$.

Interestingly, in this case, the probabilistic approach leads to the same answer as the interval techniques – for which this set is called the *united solution set.*

7 General case

In general, we also know the values x_{ki} with interval uncertainty. Then the ML approach selects the following set:

The set of all the values c for which $F(x_k, c) \in \boldsymbol{y}_k = [\widetilde{y}_k - \Delta_y, \widetilde{y}_k + \Delta_y]$ for some *values $x_{ki} \in \boldsymbol{x}_{ki} = [\widetilde{x}_{ki} - \Delta_{x_i}, \widetilde{x}_{ki} + \Delta_{x_i}]$.*

This is also exactly the *united solution set* to the interval equation system constructed from interval data. Thus, the united solution set has a natural probabilistic meaning; see, e.g., [4].

8 A more realistic description of the practical problem

Often, when we get a measurement result, this does not mean that there was only one measurement. It means that there were several different measurements leading to the same result – e.g., same intervals. Let us give a few examples.

- When a patient's blood pressure is measured at the doctor's office, usually, the device performs three measurements and – if they coincide – combines them into a single measurement result.
- This is also how super-precise atomic clocks work – each of them consists of several independent clocks, whose results are returned to the user if most of their readings coincide.

- This is how new values are measured – be it a more accurate value of the distance to the Moon or a new values of an element's atomic weight. With a single measurement, the result is not fully reliable, so, to make it reliable, several measurements are performed and if they all coincide, the joint result is accepted.

9 How probabilistic techniques deal with this situation

For each k, instead of a single combination x_k, we have several $x_{k\ell}$ for different ℓ. For each combination of values $x_{k\ell i} \in \boldsymbol{x}_{ki}$, we can form the log-likelihood

$$\sum_{k=1}^{K} \sum_{\ell} \sum_{i=1}^{n} \ln(f_i(\widetilde{y}_k - F(x_{k\ell}, c))). \tag{1}$$

We do not know the actual values $x_{k\ell i}$. Following the maximum entropy idea, we assume that they are uniformly distributed on the corresponding intervals $\boldsymbol{x}_{ki}$.

For a reasonably large number of constituent measurement ℓ, the sample average of any quantity – i.e., the arithmetic average over ℓ – is very close to its expected value; see, e.g, [9]. Thus, the sum over ℓ in the formula (1) – which is proportional to the sample average – is proportional to the expected value.

Multiplying the objective function by a proportionality constant does not change the location of its maxima. Thus, maximizing the original expression (1) for the likelihood (1) is equivalent to maximizing the expected value of the log-likelihood

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \ln(f_i(\widetilde{y}_k - F(x_{k\ell}, c)))$$

over these uniform distributions.

10 What is the resulting estimate

Result. Let us show that, as a result, we return the following set:

The set of all the values c for which $f(x_k, c) \in \boldsymbol{y}_k$ for all $x_{ki} \in \boldsymbol{x}_{ki}$.

Proof. Indeed, if the condition $f(x_k, c) \in \boldsymbol{y}_k$ is not satisfied for some $x_{ki} \in \boldsymbol{x}_{ki}$, then, for a continuous function $f(x, c)$, there is a whole subrange of the interval $\boldsymbol{x}_{ki}$ on which this condition is not satisfied. On this subrange, the likelihood will be equal to 0. Thus, on this subrange, the log-likelihood is equal to $\ln(0) = -\infty$; hence, the expected value of log-likelihood is equal to $-\infty$ – so it cannot be the largest. Thus, for all the tuples c selected by the Maximum Likelihood approach, we indeed have $f(x_k, c) \in \boldsymbol{y}_k$ for *all* $x_{ki} \in \boldsymbol{x}_{ki}$.

Since we consider uniform distributions, for each probability distribution, all non-zero values are the same. Thus, for all such tuples c, we will have the exact same values of the expected log-likelihood. So, all such tuples c will be selected by the Maximum Likelihood approach.

This is exactly the tolerable solution set. The above formula is exactly the *tolerable solution set* to the interval equation system constructed from data; see, e.g., [8].

So, the tolerable solution set also makes sense in the probabilistic setting.

Unexpected consequence: a more realistic analysis makes the data fitting problem easier to solve. Good news is that:

- in contrast to the united solution set – whose computation is, in general, NP-hard even when the expression $f(x,c)$ linearly depends on c_i (see, e.g., [3]),
- computation of the tolerable solution set can be, for the case when $f(x,c)$ is linear in c_i, reduced to linear programming and is, thus, feasible; see, e.g., [3, 8].

Acknowledgments

This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science);
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 19th International Symposium on Scientific Computing, Computer Arithmetic, and Verified Numerical Computation SCAN'2021 (Szeged, Hungary, September 13–15, 2021) for valuable suggestions.

References

[1] L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2001.

[2] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[3] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

[4] V. Kreinovich and S. P. Shary, "Interval methods for data fitting under uncertainty: a probabilistic treatment", *Reliable Computing*, 2016, Vol. 23, pp. 105–141.

[5] G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.

[6] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[7] S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.

[8] S. P. Shary, "Weak and strong compatibility in data fitting problems under interval uncertainty", *Advances in Data Science and Adaptive Analysis*, 2020, Vol. 12, No. 1, Paper 2050002.

[9] D. J. Sheskin, *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC, London, UK, 2011.