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

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

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.

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
$$c = (c_1, \ldots, c_m).$$
- 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.
- This problem is also called *problem of parameter estimation*.

3. Data fitting problem (cont-d)

- 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 $\Delta v := \widetilde{v} - v$.

4. Known probability distributions

- In many cases, we know the probability distributions $f_i(\Delta x_i)$ and $f(\Delta y)$ of the measurement errors.
- In this case, we can use the Maximum Likelihood (ML) approach.
- This means that we select the *most probable* values c (and x_{ki}) for which the product $\prod\limits_{k=1}^{K} \left(f(\widetilde{y}_k - F(x_k, c)) \cdot \prod\limits_{i=1}^{n} f_i(\widetilde{x}_{ki} - x_{ki}) \right)$ is the largest.
- Usually, the logarithm of this product, known as *log-likelihood*, is maximized for computational convenience.

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 $[-\Delta_v, \Delta_v]$.
- 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.

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 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.

7. General case

- In general, we also know the values x_{ki} with interval uncertainty.
- Then the ML approach selects the set of all the values c for which

$$F(x_k, c) \in \boldsymbol{y}_k = [\widetilde{y}_k - \Delta_y, \widetilde{y}_k + \Delta_y]$$

$$\text{for some values } x_{ki} \in \boldsymbol{x}_{ki} = [\widetilde{x}_{ki} - \Delta_{x_i}, \widetilde{x}_{ki} + \Delta_{x_i}].$$

- This is exactly the *united solution set* to the interval equation system constructed from interval data.
- Thus, the united solution set has a natural probabilistic meaning.

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.

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 $\sum\limits_{k=1}^{K} \sum\limits_{\ell} \sum\limits_{i=1}^{n} \ln(f_i(\widetilde{y}_k - F(x_{k\ell}, c)))$.
- 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 large number of constituent measurement ℓ, the sum over ℓ is proportional to the expected value.
- Thus, a reasonable idea is to maximize the expected value of the log-likelihood over this uniform distribution.

10. What is the resulting estimate

- We show that, as a result, we return all values of c for which

$$f(x_k, c) \in \boldsymbol{y}_k \text{ for } \textit{all } x_{ki} \in \boldsymbol{x}_{ki}.$$

- Indeed, otherwise, on a subrange:
  - the likelihood is 0;
  - thus, the log-likelihood is $\ln(0) = -\infty$;
  - hence, its mean value is $-\infty$ – so it cannot be the largest.
- This is exactly the *tolerable solution set* to the interval equation system constructed from data.
- So, the tolerable solution set also makes sense in the probabilistic setting.

11. 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.

12. Bibliography

- V. Kreinovich and S. P. Shary, "Interval methods for data fitting under uncertainty: a probabilistic treatment", *Reliable Computing*, 23 (2016), 105–141.
- S. P. Shary, "Weak and strong compatibility in data fitting problems under interval uncertainty", *Advances in Data Science and Adaptive Analysis*, 12 (2020), No. 1, Paper 2050002.