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, USA
[2] Novosibirsk University, Novosibirsk, Russia
vladik@utep.edu, shary@ict.nsc.ru

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.

Data fiting 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 $c = (c_1, \ldots, c_m)$. These values 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*.

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

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 — i.e., select the *most probable* values c (and x_{ki}) for which the product $\prod_{k=1}^{K} \Big(f(\widetilde{y}_k - F(x_k, c)) \cdot \prod_{i=1}^{n} f_i(\widetilde{x}_{ki} - x_{ki}) \Big)$ is the largest. Usually, the logarithm of this product, known as *log-likelihood*, is maximized for computational convenience.

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 – which turns out to be the uniform distribution.

Simplest case. The simplest – and rather frequent – case is when the values x_i are measured very accurately, so 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 $F(x_k, c) \in [\widetilde{y}_k - \Delta_y, \widetilde{y}_k + \Delta_y]$; see, e.g., [1]. Interestingly, in this case, the probabilistic approach leads to the same answer as the interval techniques.

General case. If we also know the values x_{ki} with interval uncertainty, then the ML approach selects the set of all the values c for which $F(x_k, c) \in \boldsymbol{y}_k = [\widetilde{y}_k - \Delta_y, \widetilde{y}_k + \Delta_y]$ 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 [1, 2]. Thus, the united solution set has a natural probabilistic meaning.

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.

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.

What is the resulting estimate. We show that, as a result, we return all values of c for which $f(x_k, c) \in \boldsymbol{y}_k$ for *all* $x_{ki} \in \boldsymbol{x}_{ki}$ — which is exactly the *tolerable solution set* to the interval equation system constructed from data, a solution set that has many useful properties; see, e.g., [2]. So, the tolerable solution set also makes sense in the probabilistic setting.

References

[1] V. Kreinovich, S. P. Shary: Interval methods for data fitting under uncertainty: a probabilistic treatment, *Reliable Computing*, 23 (2016), 105–141.

[2] 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.