Need for Techniques Intermediate Between Interval and Probabilistic Ones*

Olga Kosheleva[0000−0003−2587−4209] and
Vladik Kreinovich[0000−0002−1244−1650]

University of Texas at El Paso, El Paso, Texas 79968, USA
{olgak,vladik}@utep.edu

Abstract. In high performance computing, when we process a large amount of data, we do not have much information about the dependence between measurement errors corresponding to different inputs. To gauge the uncertainty of the result of data processing, the two usual approaches are: the interval approach, when we consider the worst-case scenario in which all measurement errors are strongly correlated, and the probabilistic approach, when we assume that all these errors are independent. The problem is that usually, the interval approach leads to too pessimistic, too large uncertainty estimates, while the probabilistic approach often underestimates the resulting uncertainty. To get realistic estimates, it is therefore desirable to have techniques intermediate between interval and probabilistic ones. In this paper, we propose such techniques based on the assumption that, in each practical situation, there is an upper bound $b \in [0,1]$ on the absolute value of all correlations – the bound that needs to be experimentally determined. For $b = 0$, we get probabilistic estimates, for $b = 1$, we get interval estimates, and for intermediate values b, we get the desired intermediate techniques. We also provide efficient algorithms for implementing the new techniques.

Keywords: Interval uncertainty· Probabilistic uncertainty · High performance computing.

1 Formulation of the Problem

Need to take uncertainty into account in high-performance computing. One of the main applications of high performance computing is estimating the values of some quantities y based on the inputs $x_1, \ldots, x_n$. For example, in weather prediction, we estimate tomorrow's temperature y at some location

* 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), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology. 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, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

based on the results x_i of meteorological measurements in the vicinity of this location.

The problem is that even when the data processing algorithm

$$y = f(x_1, \ldots, x_n)$$

describes the exact relation between y and x_i, the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ – that we obtain by processing measurement results $\widetilde{x}_i$ – is not exact: since the measurement results $\widetilde{x}_i$ are, in general, different from the actual (unknown) values x_i of the corresponding quantities. Because of the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the result $\widetilde{y}$ of data processing is, in general, different from the desired value y. It is important to provide an estimate for the resulting uncertainty $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$; see, e.g., [6].

What do we usually know and what we usually do not know about the measurement errors Δx_i. For each measuring instrument, we know the upper bound Δ_i on the absolute value of the measurement error, i.e., a value for which $|\Delta x_i| \le \Delta_i$. Indeed, if no such bound is guaranteed, this would mean that for any measurement result, the actual value can be anything – this would be a wild guess, not a measuring instrument.

In many practical applications, each measuring instrument is calibrated: before using this instrument, we several times compare its results with the results of a much more accurate instrument; thus, if the mean value of the measurement error was not 0, we can find this mean value (known as *bias*) and correct for it by subtracting this mean value from all the measurement results. Thus, we can safely assume that for each instrument, the mean value of the measurement error is 0.

In most applications, we can also safely assume that the measurement errors are relatively small. So we can safely ignore terms which are quadratic or higher order in terms of these errors. For example, even if the relative measurement error is 10%, its square is 1%, which can be safely ignored in comparison with 10%.

This is often all we know. Ideally, we should also know the probability distributions of all the measurement errors and all the correlations between them. In simple computations, when the number n of inputs is small, it is possible to extract this information for all n instruments and all $n^2/2$ pairs of instruments. So, for simple computations, this information is sometimes available. However, for high-performance computing, when n is large, it is not feasible to extract all this information, so this information is usually not available.

Possibility of linearization. By definition of the measurement errors, we have $x_i = \widetilde{x}_i - \Delta x_i$, thus

$$\Delta y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n).$$

Since the measurement errors Δx_i are small, we can expand the expression $f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n)$ in Taylor series in terms of Δx_i and keep only linear

terms in this expansion. As a result, we get

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i, \tag{1}$$

where

$$c_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i}_{|x_1=\tilde{x}_1,\ldots,x_n=\tilde{x}_n}. \tag{2}$$

How Δy is estimated now: first technique. Since we have no information about the correlation between the measurement errors, a natural idea is to consider all possible correlations. In general, since $|a+b| \le |a|+|b|$ and $|a\cdot b| = |a|\cdot|b|$, from the formula (1), we get

$$|\Delta y| \le \sum_{i=1}^{n} |c_i| \cdot |\Delta x_i|.$$

Since $|\Delta x_i| \le \Delta_i$, we get

$$|\Delta y| \le \Delta_{\text{int}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \Delta_i. \tag{3}$$

This value Δ_{int} is the exact upper bound, in the sense that it is possible to have $|\Delta y| = \Delta_{\text{int}}$ with probability 1. Indeed, this happens when:

- with probability $1/2$, we have $\Delta x_i = \Delta_i \cdot \text{sign}(c_i)$, where, as usual, $\text{sign}(x) = +1$ for $x > 0$ and $\text{sign}(x) = -1$ for $x < 0$; and
- with probability $1/2$, we have $\Delta x_i = -\Delta_i \cdot \text{sign}(c_i)$.

In this case:

- with probability $1/2$, we have $\Delta y = \Delta_{\text{int}}$, and
- with probability $1/2$, we have $\Delta y = -\Delta_{\text{int}}$.

This worst-case estimate (3) is known as the *interval estimate*, since this is the only estimate that we can guarantee based on the available information – that all measurement errors Δx_i are located within the corresponding interval $[-\Delta_i, \Delta_i]$; see, e.g., [2, 4, 5].

Interval technique: limitation. The main problem with this approach is that the resulting worst-case estimates are too pessimistic. In most practical situations, the actual value Δy is much smaller than Δ_{int}.

How can we explain this limitation. The above limitation can be easily explained. Indeed:

- In the arrangement that leads to $\Delta y = \Delta_{\text{int}}$, all measurement errors are highly correlated, with correlation coefficients ± 1.

– In practice, it is possible that common factors affect several measurement instruments, but there are also usually other factors which affect only one measuring instrument, so the correlation is usually larger than -1 and smaller than 1.

How Δy is estimated now: second technique. Another idea is that since we have no reason to prefer negative or positive correlation, it is reasonable to assume that the correlation is 0, and, more generally, that different measurement errors are independent.

This is also what follows from the Maximum Entropy approach [3], when out of all possible joint distributions $\rho(\Delta x_1, \ldots, \Delta x_n)$ for which mean of each variable is 0 and which are located on the given intervals $[-\Delta_i, \Delta_i]$, we select the distribution with the largest value of entropy

$$S \stackrel{\text{def}}{=} -\int \rho(\Delta x_1, \ldots, \Delta x_n) \cdot \ln(\rho(\Delta x_1, \ldots, \Delta x_n))\, d\Delta x_1 \ldots d\Delta x_n.$$

Independence means that for each $i \neq j$, the expected value $E[\Delta x_i \cdot \Delta x_j]$ of the product $\Delta x_i \cdot \Delta x_j$ is equal to the product of expected values

$$E[\Delta x_i \cdot \Delta x_j] = E[\Delta x_i] \cdot E[\Delta x_j],$$

i.e., since the mean value of each measurement error is 0, to

$$E[\Delta x_i \cdot \Delta x_j] = 0.$$

In this case, the expected value of $(\Delta y)^2$ is equal to

$$E[(\Delta y)^2] = \sum_{i=1}^{n} c_i^2 \cdot V_i,$$

where by

$$V_i \stackrel{\text{def}}{=} E[(\Delta x_i - E[\Delta x_i])^2] = E[(\Delta x_i)^2],$$

we denoted the variance of the i-th measurement error.

As is well known in statistics, for large n, the deviation from this average is small – since it grows with n as $\sqrt{n}$, while the expected value itself grows as n [7], so we conclude that the actual value $(\Delta y)^2$ is, with high accuracy, equal to this expected value:

$$(\Delta y)^2 \approx \sum_{i=1}^{n} c_i^2 \cdot V_i.$$

We do not know the variances V_i, but, since $|\Delta x_i| \leq \Delta_i$, we have $(\Delta x_i)^2 \leq \Delta_i^2$. Thus, the expected value V_i of the square $(\Delta x_i)^2$ is also bounded by the same bound Δ_i^2:

$$V_i \leq \Delta_i^2.$$

This upper bound on the variance V_i is the best we can have – it is attained if:

- we have $\Delta x_i = \Delta_i$ with probability $1/2$, and
- we have $\Delta x_i = -\Delta_i$ with probability $1/2$.

Thus, we conclude that

$$(\Delta y)^2 \le \sum_{i=1}^{n} c_i^2 \cdot \Delta_i^2,$$

i.e., that

$$|\Delta y| \le \Delta_{\text{prob}} \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n} c_i^2 \cdot \Delta_i^2}. \tag{4}$$

Probabilistic technique: limitation. The main problem with this probabilistic technique is that it is too optimistic, it often drastically decreases the approximation error Δy.

How can we explain this limitation. The above limitation can be easily explained. Indeed:

- This technique assumes that all the measurement errors are independent.
- However, as we have mentioned, in reality, there may be common factors affecting several instruments, and thus, there is correlation.

Need for intermediate techniques. Since the interval techniques are too pessimistic and the probability techniques are too optimistic, it is desirable to have intermediate techniques that would provide more realistic estimates.

The main objective of this paper is to provide such estimates.

2 Main Idea and the Resulting Formula and Algorithm

Main idea. As we have mentioned, the problem with the interval technique is that it assumes that the absolute value of the correlation can be 1, while in practice, it is always smaller than 1. Similarly, the problem with the probabilistic technique is that it assumes that all correlations are 0s, while in practice, they can take non-zero values.

So, a natural idea is to assume that there is some number b between 0 and 1 that provides an upper bound for absolute values $|r_{ij}|$ of all the correlations

$$r_{ij} \stackrel{\text{def}}{=} \frac{E[\Delta x_i \cdot \Delta x_j]}{\sigma_i \cdot \sigma_j},$$

where $\sigma_i \stackrel{\text{def}}{=} \sqrt{V_i}$:

$$|r_{ij}| \le b.$$

This value can be determined empirically, by computing absolute value of the correlation for several randomly selected pairs of measuring instruments and selecting the largest of these values.

From the idea to the resulting formula. From the formula (1), we conclude that

$$(\Delta y)^2 = \sum_{i=1} c_i^2 \cdot (\Delta x_i)^2 + \sum_{i \neq j} c_i \cdot c_j \cdot \Delta x_i \cdot \Delta x_j,$$

hence

$$E[(\Delta y)^2] = \sum_{i=1}^{n} c_i^2 \cdot E[(\Delta x_i)^2] + \sum_{i \neq j} c_i \cdot c_j \cdot E[\Delta x_i \cdot \Delta x_j],$$

i.e.,

$$E[(\Delta y)^2] = \sum_{i=1}^{n} c_i^2 \cdot V_i + \sum_{i \neq j} c_i \cdot c_j \cdot r_{ij} \cdot \sigma_i \cdot \sigma_j.$$

We know that $(\Delta y)^2 \approx E[(\Delta y)^2]$, we know that $|r_{ij}| \leq b$, so we conclude that

$$(\Delta y)^2 \leq \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + \sum_{i \neq j} |c_i| \cdot |c_j| \cdot b \cdot \sigma_i \cdot \sigma_j.$$

We have mentioned that $\sigma_i \leq \Delta_i$, thus

$$(\Delta y)^2 \leq \sum_{i=1}^{n} c_i^2 \cdot \Delta_i^2 + \sum_{i \neq j} |c_i| \cdot |c_j| \cdot b \cdot \Delta_i \cdot \Delta_j. \tag{5}$$

Here,

$$I_{\text{int}}^2 = \left(\sum_{i=1}^{n} |c_i| \cdot \Delta_i \right)^2 = \sum_{i=1}^{n} c_i^2 \cdot \Delta_i^2 + \sum_{i \neq j} |c_i| \cdot |c_j| \cdot \Delta_i \cdot \Delta_j,$$

thus the formula (5) takes the form

$$(\Delta y)^2 \leq b \cdot I_{\text{int}}^2 + (1-b) \cdot \left(\sum_{i=1}^{n} c_i^2 \cdot \Delta_i^2 \right),$$

i.e., the form

$$(\Delta y)^2 \leq b \cdot I_{\text{int}}^2 + (1-b) \cdot I_{\text{prob}}^2.$$

So, we arrive at the following final formula.

Resulting formula.

$$|\Delta y| \leq I_b \stackrel{\text{def}}{=} \sqrt{b \cdot I_{\text{int}}^2 + (1-b) \cdot I_{\text{prob}}^2}. \tag{6}$$

How to compute this estimate. There exist efficient algorithms:

- for computing I_{prob} – based on Monte-Carlo simulation of normally distributed measurement errors – and
- for computing I_{prob} – based on using Cauchy distribution [1].

In both algorithms, the number of simulations depend only on the desired accuracy and does not depends on the number n of inputs.

By using these algorithms, we can efficiently compute the new estimate (6).

References

1. Kreinovich, V., Ferson, S.: A new Cauchy-based black-box technique for uncertainty in risk analysis. Reliability Engineering and Systems Safety **85**(1-3), 267–279 (2004)
2. Jaulin, L., Kiefer, M., Didrit, O., Walter, E.: Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics, Springer, London (2001)
3. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK (2003)
4. Mayer G.: Interval Analysis and Automatic Result Verification, de Gruyter, Berlin (2017)
5. Moore, R.E., Kearfott, R.B., Cloud, M.J.: Introduction to Interval Analysis, SIAM, Philadelphia (2009)
6. Rabinovich, S.G.: Measurement Errors and Uncertainty: Theory and Practice, Springer Verlag, New York (2005)
7. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman and Hall/CRC, Boca Raton, Florida (2011)