How Accurately Can We Determine the Coefficients: Case of Interval Uncertainty

Michal Cerny and Vladik Kreinovich

Abstract In many practical situations, we need to estimate the parameters of a linear (or more general) dependence based on measurement results. To do that, it is useful, before we start the actual measurements, to estimate how accurately we can, in principle, determine the desired coefficients: if the resulting accuracy is not sufficient, then should not waste time trying and resources and instead, we should invest in more accurate measuring instruments. This is the problem that we analyze in this paper.

1 Formulation of the Problem

Need to determine the dependence between different quantities. One of the main objectives of science is to find the dependencies $y = f(x_1, \ldots, x_n)$ between values of different quantities at different moments of time and at different locations.

Once we know such dependencies, we can then use them to predict the future values of different quantities.

For example, Newton's laws describe how the acceleration y of a celestial body depends on the current location and masses of this and other bodies $x_1, \ldots, x_n$ – and thus, these laws enable us to predict how these bodies will move.

Another important case is when we want to estimate the value of a quantity y which is difficult to directly measure. In such cases, it is often possible to find easier-to-measure quantities $x_1, \ldots, x_n$ knowing which we can determine y. For example, it is difficult to directly measure the distance y between two faraway locations on the

Michal Cerny
Faculty of Informatics and Statistics, University of Economics, nam. W. Churchilla 4
13067 Prague, Czech Republic, e-mail: cernym@vse.cz

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: vladik@utep.edu

Earth, but we can determine this distance if we use astronomical observations – or, nowadays, signals from the GPS satellites – to find the exact coordinates of each of the two locations.

How can we determine this dependence. In some cases, we can use the known physical laws to derive the desired dependence. However, in most other cases, this dependence needs to be determined empirically:

- we measure the values $x_1, \ldots, x_n$, and y in different situations, and then
- we use the measurement results to find the desired dependence.

Often, we know the general form of the dependence, we just need to find the coefficients. In many cases, we know the general form of the desired dependence, i.e., we know that $y = F(x_1, \ldots, x_n, c_0, c_1, \ldots, c_m)$, where F is known function, and the coefficients c_i need to be determined.

For example, we may know that the dependence is linear, i.e., that

$$y = c_0 + c_1 \cdot x_1 + \ldots + c_n \cdot x_n.$$

This is a typical situation when the values x_i have a narrow range $[\underline{X}_i, \overline{X}_i]$ and thus, we can expand the function $f(x_1, \ldots, x_n)$ in Taylor series over $x_i - \underline{X}_i$ and ignore quadratic (and higher order) terms in this expansion.

Need to take uncertainty into account – in particular, interval uncertainty. Measurements are never absolutely accurate: the measurement result $\widetilde{x}$ is, in general, different from the actual (unknown) value x. In many practical situations, the only information that we have about the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is the upper bound Δ on its absolute value: $|\Delta x| \leq \Delta$; see, e.g., [5].

In this case, after each measurement, the only information that we have about the actual value x is that this value is somewhere in the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$. Because of this fact, this case is known as the case of *interval uncertainty*. There exist many algorithms for dealing with such uncertainty; see, e.g., [1, 3, 4].

Measurement uncertainty leads to uncertainty in coefficients. Since we can only measure the values x_i and y with some uncertainty, we can therefore only determine the coefficients c_i with some uncertainty.

It is therefore important to determine how accurate are the values c_i that we get as a result of these measurements.

Which uncertainty should be taken into account. Strictly speaking, there are measurement uncertainties both when we measure easier-to-measure quantities $x_1, \ldots, x_n$, and when we measure the desired difficult-to-measure quantity y. However, usually, because of the very fact that y is much more difficult to measure than x_i, the measurement errors Δy corresponding to measuring y are much larger than the measurement errors of measuring x_i – so much larger that we can usually safely ignore the measurement errors of measuring x_i and assume that these values are known exactly.

Thus, in the linear case, we can safely assume that for each measurement k, we know the exact values $x_1^{(k)}, \ldots, x_n^{(k)}$, but we only know $y^{(k)}$ with uncertainty – i.e.,

based on the measurement result $\widetilde{y}^{(k)}$ and the known accuracy $\Delta > 0$, we know that the actual value $y^{(k)} = c_0 + \sum\limits_{i=1}^{n} c_i \cdot x_i^{(k)}$ is between $\underline{y}^{(k)} = \widetilde{y}^{(k)} - \Delta$ and $\overline{y}^{(k)} = \widetilde{y}^{(k)} + \Delta$.

Once we perform the measurements, we can feasibly find the accuracy. One we have the measurement results, we can find the bounds on each of the coefficients c_i (and, similarly, the bounds on any linear combination of c_i) by solving the following linear programming problems (see, e.g., [7]): minimize (maximize) c_i under the constraints that

$$\underline{y}^{(k)} \le c_0 + \sum_{i=1}^{n} c_i \cdot x_i^{(k)} \le \overline{y}^{(k)}$$

for all the measurements $k = 1, \ldots, K$.

Remaining question. But before we start spending our resources on measurements, it is desirable to check how accurately we can, in principle, determine the coefficients c_i.

This checking is important: If the resulting accuracy is not enough for us – then we should not waste time performing the measurements, and instead we should invest in a more accurate y-measuring instrument.

Of course, we can answer the above question by simulating measurement errors, but it would be great to have simple analytical expressions that would not require extensive simulation-related computations.

What we do in this paper. In this paper, we provide such expressions for the linear case.

2 Definitions and Results

Discussion. The range of each physical quantity is usually bounded:

- coordinates of Earth locations are bounded by the Earth's size,
- velocities are bounded by the speed of light, etc.

Thus, we can safely assume that for each variable x_i, we know the interval $[\underline{X}_i, \overline{X}_i]$ of its possible values.

Thus, we arrive at the following formulation of the problem.

Definition 1. *Let us assume that we are given the value $\Delta > 0$ and n intervals $[\underline{X}_i, \overline{X}_i]$, $i = 1, 2, \ldots, n$. We say that a tuple $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ is* within the possible uncertainty *if for each tuple $(c_0, c_1, \ldots, c_n)$ and for each combination of values $x_i \in [\underline{X}_i, \overline{X}_i]$, we have $|y' - y| \le \Delta$, where:*

- $y \stackrel{\text{def}}{=} c_0 + \sum\limits_{i=1}^{n} c_i \cdot x_i$ *and*
- $y' \stackrel{\text{def}}{=} c_0' + \sum\limits_{i=1}^{n} c_i' \cdot x_i$, *where* $c_i' \stackrel{\text{def}}{=} c_i + \Delta c_i$.

Comment. Because of the measurement uncertainty, after the measurement, the range of possible values of the corresponding quantity x is $[\widetilde{x}-\Delta,\widetilde{x}+\Delta]$. It may be therefore convenient to represent the intervals $[\underline{X}_i,\overline{X}_i]$ in the same form, as

$$[\underline{X}_i,\overline{X}_i] = [\widetilde{X}_i-\Delta_i,\widetilde{X}_i+\Delta_i].$$

For this, we need to take $\widetilde{X}_i = \dfrac{\underline{X}_i+\overline{X}_i}{2}$ and $\Delta_i = \dfrac{\overline{X}_i-\underline{X}_i}{2}$.

Proposition 1. *For each Δ and $[\underline{X}_i,\overline{X}_i]$, a tuple $(\Delta c_0,\Delta c_1,\ldots,\Delta c_n)$ is within the possible uncertainty if and only if*

$$|\Delta c_0'| + \sum_{i=1}^{n} |\Delta c_i|\cdot\Delta_i \le \Delta, \tag{1}$$

where $\Delta c_0' \stackrel{\text{def}}{=} \Delta c_0 + \sum\limits_{i=1}^{n} \Delta c_i\cdot\widetilde{X}_i$.

Proof of Proposition 1. One can easily see that, since the dependence of y on c_i is linear, the difference $\Delta y \stackrel{\text{def}}{=} y'-y$ is equal to $\Delta y = \Delta c_0 + \sum\limits_{i=1}^{n} \Delta c_i\cdot x_i$.

Each of the variables x_i independently runs over its own interval

$$[\underline{X}_i,\overline{X}_i] = [\widetilde{X}_i-\Delta_i,\widetilde{X}_i+\Delta_i].$$

Thus, each value x_i from this interval can be represented as $\widetilde{X}_i+\Delta x_i$, where $\Delta x_i \stackrel{\text{def}}{=} x_i-\widetilde{X}_i$ takes all possible values from the interval $[-\Delta_i,\Delta_i]$.

Substituting this expression for x_i into the above formula for Δy, we conclude that

$$\Delta y = \Delta c_0' + \sum_{i=1}^{n} \Delta c_i\cdot\widetilde{X}_i + \sum_{i=1}^{n} \Delta c_i\cdot\Delta x_i. \tag{2}$$

To make sure that always $|\Delta y|\le\Delta$, i.e., that always $-\Delta\le\Delta y\le\Delta$, it is sufficient to make sure that

$$-\Delta \le \underline{\Delta} \text{ and } \overline{\Delta} \le \Delta,$$

where:

- $\underline{\Delta}$ is the smallest possible value of the expression (2), while
- $\overline{\Delta}$ is the largest possible value of the expression (2).

Let us find these smallest and largest values.

Each of the variables Δx_i independently runs over its own interval $[-\Delta_i,\Delta_i]$. Thus, the smallest possible value of (1) is attained when each of the terms in the sum (2) is the smallest.

- For $\Delta c_i\ge 0$, the term $\Delta c_i\cdot\Delta x_i$ is increasing with Δx_i, so its smallest value if when x_i is the largest: $\Delta x_i = -\Delta_i$. In this case, the value is equal to $-\Delta c_i\cdot\Delta_i$.
- For $\Delta c_i\le 0$, the term $\Delta c_i\cdot\Delta x_i$ is decreasing with Δx_i, so its smallest value if when x_i is the largest: $\Delta x_i = \Delta_i$. In this case, the value is equal to $\Delta c_i\cdot\Delta_i$.

We can describe both terms by a single formula $-|\Delta c_i| \cdot \Delta_i$. Thus, the smallest possible value $\underline{\Delta}$ of Δy is equal to $\underline{\Delta} = \Delta c_0' - \sum_{i=1}^{n} |\Delta c_i| \cdot \Delta_i$, and the condition $-\Delta \le \underline{\Delta}$ is equivalent to

$$-\Delta c_0' + \sum_{i=1}^{n} |\Delta c_i| \cdot \Delta_i \le \Delta. \tag{3}$$

Similarly, the largest possible value of each term $\Delta c_i \cdot \Delta x_i$ is equal to $|\Delta c_i| \cdot \Delta_i$, thus

$$\overline{\Delta} = \Delta c_0' + \sum_{i=1}^{m} |\Delta c_i| \cdot \Delta_i,$$

and the condition $\overline{\Delta} \le \Delta$ can be described as

$$\Delta c_0' + \sum_{i=1}^{n} |\Delta c_i| \cdot \Delta_i \le \Delta. \tag{4}$$

Inequalities (3) and (4) are equivalent to requiring that the largest of the two left-hand sides is smaller than or equal to Δ, i.e., to the desired inequality. The proposition is proven.

Discussion. Based on Proposition 1, we can find bounds on each of the coefficient $\Delta c_1, \ldots, \Delta c_n$:

Proposition 2. *For each i from 1 to n, among all possible tuples which are within the possible uncertainty, the corresponding values of Δc_i form the interval*

$$\left[-\frac{\Delta}{\Delta_i}, \frac{\Delta}{\Delta_i}\right].$$

Comments. Thus, if we can measure y with accuracy Δ, and we can use any value x_i from the interval $[\widetilde{X}_i - \Delta_i, \widetilde{X}_i + \Delta_i]$, then we can determine the coefficient c_i that describes the dependence of y on x_i with accuracy $\dfrac{\Delta}{\Delta_i}$.

It is worth mentioning that the accuracy $\dfrac{\Delta}{\Delta_i}$ is what we can *guarantee* if we perform sufficiently many measurements. However, even with a primitive y-measuring device, for which the measurement accuracy Δ is high, we can get lucky and get much more accurate – even absolutely accurate – values of c_i.

Indeed, let us assume that for each tuple $\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)$ of the x-values, for which the actual value of y is $y^{(k)} = c_0 + \sum_{i=1}^{n} c_i \cdot x_i^{(k)}$, we perform two y-measurements:

- in the first measurement, we get $\widetilde{y}^{(k)} = y^{(k)} + \Delta$ and thus, based on this measurement result, we conclude that the actual value of $y^{(k)}$ belongs to the interval

$$[\widetilde{y}^{(k)} - \Delta, \widetilde{y}^{(k)} + \Delta] = [y^{(k)}, y^{(k)} + 2\Delta];$$

- in the second measurement, we get $\widetilde{y}^{(k)} = y^{(k)} - \Delta$ and thus, based on this measurement result, we conclude that the actual value of $y^{(k)}$ belongs to the interval

$$[\widetilde{y}^{(k)} - \Delta, \widetilde{y}^{(k)} + \Delta] = [y^{(k)} - 2\Delta, y^{(k)}].$$

Since the value $y^{(k)}$ belongs to both intervals $[y^{(k)}, y^{(k)} + 2\Delta]$ and $[y^{(k)} - 2\Delta, y^{(k)}]$, it belongs to their intersection – and this intersection consists of the single point $y^{(k)}$. Thus, in this lucky case, we get the exact value of each y – and thus, after $n+1$ measurement, determine the exact values of all $n+1$ coefficients $c_0, 1c_1, \ldots, c_n$ by solving the corresponding system of linear equations

$$c_0 + \sum_{i=1}^{n} c_i \cdot x^{(k)} = y^{(k)}, \;\; k = 1, \ldots, n+1.$$

Proof of Proposition 2. If Δc_i is a part of the tuple which is within the possible uncertainty, then from the inequality (1), we can conclude that $|\Delta c_i| \cdot \Delta_i \le \Delta$, hence that

$$|\Delta c_i| \le \frac{\Delta}{\Delta_i}. \tag{5}$$

Vice verse, for each value Δc_i that satisfies the inequality (5), we can take $\Delta c_1 = \ldots = \Delta c_{i-1} = \Delta c_{i+1} = \ldots \Delta c_n = 0$ and choose $\Delta c_0 = -\Delta x_i \cdot \widetilde{X}_i$, then $\Delta c_0' = 0$ and thus, the inequality (1) is satisfied.

The proposition is proven.

Proposition 3. *When 0 is a possible value of each variable x_i, then among all possible tuples which are within the possible uncertainty, the corresponding values of Δc_0 form the interval $[-\Delta, \Delta]$.*

Comment. Thus, if we can measure y with accuracy Δ, and we can use any value x_i from the interval $[\widetilde{X}_i - \Delta_i, \widetilde{X}_i + \Delta_i]$ containing 0, then we can determine the free term c_0 in the dependence of y on $x_1, \ldots, x_n$ with accuracy Δ.

Proof of Proposition 3. If a tuples $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ is within the possible uncertainty, then for possible value $x_1 = \ldots = x_n = 0$, we get $|\Delta c_0| \le \Delta$.

Vice versa, if we have a value Δc_0 for which $|\Delta c_0| \le \Delta$, them, by taking $\Delta c_1 = \ldots = \Delta c_n = 0$, we get a tuple that, as one can easily see, satisfies the desired inequality for all x_i and is, thus, within the possible uncertainty.

The proposition is proven.

Discussion. When for some i, $0 \notin [\underline{X}_i, \overline{X}_i]$, then all values $\Delta c_0 \in [-\Delta, \Delta]$ are still possible, but some values outside this interval are possible too.

Proposition 4. *For every value $\Delta c_0 \in [-\Delta, \Delta]$, there exists a tuple $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ which is within the possible uncertainty.*

Proof: this was, in effect, already proven in the proof of Proposition 3.

Proposition 5. *For $n = 1$, the range of possible values of Δc_0 is $[-\Delta', \Delta']$, where $\Delta' = \Delta + \dfrac{\Delta}{\Delta_1} \cdot m_1$ and:*

- $m_1 = 0$ *if* $0 \in [\underline{X}_1, \overline{X}_1]$*;*
- $m_1 = \underline{X}_1$ *is* $\underline{X}_1 > 0$*, and*
- $m_1 = |\overline{X}_1|$ *when* $\overline{X}_1 < 0$*.*

Proof. We have already proven this result for the case when $0 \in [\underline{X}_1, \overline{X}_1]$. Without losing generality, let us consider the case when $\underline{X}_1 > 0$; the case when $\overline{X}_1 < 0$ is proven similarly.

In this case, on the one hand, for $x_1 = \underline{X}_1$, we have $|\Delta c_0 + \Delta c_1 \cdot \underline{X}_1| \le \Delta$, hence

$$|\Delta c_0| \le |\Delta c_0 + \Delta c_1 \cdot \underline{X}_1| + |-\Delta c_1 \cdot \underline{x}_1| \le \Delta + |\Delta c_1| \cdot \underline{X}_1.$$

By Proposition 1, we have $|\Delta c_1| \le \dfrac{\Delta}{\Delta_1}$, hence indeed $|\Delta c_0| \le \Delta'$.

On the other hand, let us prove that the value $\Delta c_0 = \Delta'$ is possible. Then, by swapping the signs of all Δc_i, we can prove that the value $-\Delta'$ is also possible. The inequalities $|\Delta c_0 + \Delta c_1 \cdot x_1| \le \Delta$ that describe the set of possible tuples is an intersection of convex sets and is, thus, itself convex. So, with Δ' and $-\Delta'$, any convex combination of them is also possible – i.e., all the values from the interval $[-\Delta', \Delta']$.

Hence, it is sufficient to prove that the value $\Delta c_0 = \Delta'$ is possible. Indeed, we will prove that it is possible if we take $\Delta c_1 = -\dfrac{\Delta}{\Delta_1}$. We then need to prove that for these values Δc_i, we have $|\Delta c_0 + \Delta c_1 \cdot x_1| \le \Delta$ for all $x_1 \in [\underline{X}_1, \overline{X}_1]$.

The left-hand side of the inequality is a convex function of x_1, so it is sufficient to check this inequality for the endpoints $x_1 = \underline{X}_1$ and $x_1 = \overline{X}_1$. For $x_1 = \underline{X}_1$, we have

$$\Delta c_0 + \Delta c_1 \cdot \underline{X}_1 = \Delta + \frac{\Delta}{\Delta_1} \cdot \underline{X}_1 - \frac{\Delta}{\Delta_1} \cdot \underline{X}_1 = \Delta,$$

and for $x_1 = \overline{X}_1$, we get

$$\Delta c_0 + \Delta c_1 \cdot \overline{X}_1 = \Delta + \frac{\Delta}{\Delta_1} \cdot \underline{X}_1 - \frac{\Delta}{\Delta_1} \cdot \overline{X}_1 =$$

$$\Delta - \frac{\Delta}{\Delta_1} \cdot (\overline{X}_1 - \underline{X}_1) = \Delta - \frac{\Delta}{\Delta_1} \cdot 2\Delta_1 = \Delta - 2\Delta = -\Delta.$$

In both cases, we have $|\Delta c_0 + \Delta c_1 \cdot x_1| \le \Delta$. Thus, the proposition is proven.

3 Discussion

What if we have probabilistic uncertainty. In the above text, we considered the case when we only know the upper bound on the measurement errors – i.e., when we only know the interval of possible values of the measurement error. In many practical situations, however, in addition to this upper bound, we also have some information about the probability of different values from this interval.

In such cases, it is convenient to represent the measurement error as the same of two components:

- its mean, which is called *systematic error*, and
- the difference between the measurement error and its mean, which is called the *random error*.

Usually, we know the upper bound Δ_i on the absolute value of the systematic error, and we know some characteristics of the random error; see, e.g., [5]. With what accuracy can we then determine c_i?

Interestingly, we get the same answer as in the interval case. Indeed, if for the same example, we measure y several times, the arithmetic average of the measurement results tends to its mean value, i.e., to the actual value y plus the systematic error s_i; see, e.g., [6]. Thus, in measurement results obtained this way, the random error disappears and we get, in effect, the interval case.

What if we consider quadratic dependencies. In the above text, we considered the case when we could ignore quadratic and higher order terms, and thus, safely assume that the dependence of y on x_i is linear. What if we want a more accurate description and thus, consider quadratic terms as well, i.e., consider the dependence

$$y = c_0 + \sum_{i=1} c_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \cdot x_i \cdot x_j.$$

In this case, even for a single tuple

$$(\Delta c_0, \Delta c_1, \ldots, \Delta c_n, \Delta c_{11}, \Delta c_{12}, \ldots, \Delta c_{nn}),$$

it is NP-hard (= intractable) to check whether this tuple is within the accuracy, i.e., whether

$$|\Delta y| = \left| \Delta c_0 + \sum_{i=1} \Delta c_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta c_{ij} \cdot x_i \cdot x_j \right| \le \Delta$$

for all values x_i from the corresponding intervals $[\underline{X}_i, \overline{X}_i]$: indeed, finding the maximum of a quadratic function under interval uncertainty is known to be NP-hard [2, 8].

What if we have an ellipsoid. Instead of requiring that possible values of $(x_1, \ldots, x_n)$ form a box, we can consider the case when this set is an ellipsoid.

In this case, the range of a linear expression $\Delta c_0 + \sum_{i=1}^{n} \Delta c_i \cdot x_i$ can also be explicitly computed and thus, we also have an analytical expression describing tuples $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ which are within the possible uncertainty.

What if we also have relative measurement error. In our text, we assumed that the measurement accuracy Δ is the same for all y, i.e., in measurement terms, that we have an *absolute* error. In practice, we often also have *relative* error component, in which cases the upper bound $\Delta(y)$ on the y-measurement error depends on y as $\Delta(y) = \Delta_0 + c \cdot |y|$, for some $\Delta_0 > 0$ and $c > 0$.

Once we have measurement results, we can still use linear programming to find the accuracy with which we can determine the coefficients c_i, but it is not clear how to come up with an analytical expression for the tuples $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ which are within the possible uncertainty.

Acknowledgements This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence). M. Cerny acknowledges the support of the Czech Science Foundation (project 19-02773S).

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. V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.
3. G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.
4. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.
5. S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.
6. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
7. R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Springer, New York, 2014.
8. S. A. Vavasis, *Nonlinear Optimization: Complexity Issues*, Oxford University Press, New York, 1991.