In System Identification, Interval (and Fuzzy) Estimates Can Lead to Much Better Accuracy than the Traditional Statistical Ones: General Algorithm and Case Study

Sergey I. Kumkov[1], Vladik Kreinovich[2], Andrzej Pownuk[2]

[1]Institute of Mathematics and Mechanics, Ural Branch
Russian Academy of Sciences, and Ural Federal University
Ekaterinburg, Russia, kumkov@imm.uran.ru
[2]Computational Science Program, University of Texas at El Paso
El Paso, TX 79968, USA, ampownuk@utep.edu, vladik@utep.edu

1. System Identification: A General Problem

- Often, we are interested in a quantity y which is difficult (or even impossible) to measure directly.
- This difficulty and/or impossibility may be technical:
  - while we can directly measure the distance between the two buildings by simply walking there,
  - there is no easy way to measure the distance to a nearby star by flying there.
- Impossibility may come from predictions – today, we cannot measure tomorrow's temperature.

2. System Identification (cont-d)

- A natural idea is to find easier-to-measure quantities $x_1, \ldots, x_n$ that are related to y by a known dependence

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

- Then, we can use the results $\widetilde{x}_i$ of measuring these auxiliary quantities to estimate y as $\widetilde{y} \stackrel{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- *Example:* we can find the distance to a nearby star by measuring the direction to this star in two seasons:
  - when the Earth is at different sides of the Sun, and
  - the angle is thus slightly different.
- To predict tomorrow's temperature T:
  - we can measure the temperature and wind speed and direction at different locations today, and
  - use this data to predict T.

3. System Identification (final)

- In some cases, we know the dependence

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

- In other cases, we only know the general form of this dependence

$$y = f(a_1, \ldots, a_m, x_1, \ldots, x_n).$$

- The values a_i must be estimated based on measurement results.
- We have the results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$ of measuring y and x_i in several situations $k = 1, \ldots, K$.
- Estimating a_i is called *system identification.*

4. Need to Take Measurement Uncertainty into Account

- Measurements are not 100% accurate.
- In general, the measurement result $\widetilde{x}$ is different from the actual (unknown) value x: $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x \neq 0$; thus,
  - while for the (unknown) actual values y_k and x_{ki}, we have $y_k = f(a_1, \ldots, a_m, x_{k1}, \ldots, x_{kn})$,
  - the relation between measurement results $\widetilde{y}_k \approx y_k$ and $\widetilde{x}_{ki} \approx x_{ki}$ is approximate:

$$\widetilde{y}_k \approx f(a_1, \ldots, a_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}).$$

- It is therefore important to take this uncertainty into account when estimating the values $a_1, \ldots, a_m$.

5. How Can We Describe Uncertainty?

- In all the cases, we should know the bound Δ on the absolute value of the measurement error: $|\Delta x| \leq \Delta$.
- This means that only values Δx from the interval $[-\Delta, \Delta]$ are possible.
- If this is the only information we have then:
  - based on the measurement result $\widetilde{x}$,
  - the only information that we have about the actual value x is that $x \in [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- Processing data under such interval uncertainty is known as *interval computations*.

6. How Can We Describe Uncertainty (cont-d)

- Ideally, it is also desirable to know how frequent are different values Δx within this interval.
- In other words, it is desirable to know the probabilities of different values $\Delta x \in [-\Delta, \Delta]$.
- The measurement uncertainty Δx often comes from many different independent sources.
- Thus, due to the Central Limit Theorem, the distribution of Δx is close to Gaussian.
- This explains the usual engineering practice of using normal distributions.

7. Two Approximations, Two Options

- Gaussian distribution is that it is *not* located on any interval.
- The probability of measurement error Δx to be in any interval – no matter how far away from Δ – is non-zero.
- From this viewpoint, the assumption that the distribution is Gaussian is an approximation.
- It seems like a very good approximation, since for normal distribution with mean 0 and st. dev. σ:
  - the probability to be outside the 3σ interval $[-3\sigma, 3\sigma]$ is very small, approximately 0.1%, and
  - the probability for it to be outside the 6σ interval is about 10^{-8}, practically negligible.
- Since the difference is small, this should not affect system identification.

8. Two Approximations, Two Options (cont-d)

- At first glance, if we keep the bounds but ignore probabilities, we will do much worse.
- Our results show that the opposite is true:
  - if we ignore the probabilistic information and use only interval (or fuzzy) information,
  - we get much more accurate estimates for a_j than in the statistical case.
- This is not fully surprising: theory shows that asymptotically, interval bounds are better.
- However, the drastic improvement in accuracy was somewhat unexpected.

9. System Identification: Interval Case

- For each pattern $k = 1, \ldots, K$:
  - we know the measurement results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$, and
  - we know the accuracies Δ_k and Δ_{ki} of the corresponding measurements.
- Thus, we know that:
  - the actual (unknown) value y_k belongs to the interval $[\underline{y}_k, \overline{y}_k] = [\widetilde{y}_k - \Delta_k, \widetilde{y}_k + \Delta_k]$; and
  - the actual (unknown) value x_{ki} belongs to the interval $[\underline{x}_{ki}, \overline{x}_{ki}] = [\widetilde{x}_{ki} - \Delta_{ki}, \widetilde{x}_{ki} + \Delta_{ki}]$.
- We need to find $a_1, \ldots, a_m$ for which, for every k, for some $x_{ki} \in [\underline{x}_{ki}, \overline{x}_{ki}]$,

$$f(a_1, \ldots, a_m, x_{k1}, \ldots, x_{kn}) \in [\underline{y}_k, \overline{y}_k].$$

- Specifically, for each j from 1 to m, we would like to find the range $[\underline{a}_j, \overline{a}_j]$ of all possible values of a_j.

10. Analysis of the Problem

- In the statistical case, we use the Least Squares method and find $\widetilde{a}_1, \ldots, \widetilde{a}_m$ that minimize the sum:

$$\sum_{k=1}^{K} (\widetilde{y}_k - f(a_1, \ldots, a_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}))^2 \to \min_{a_1, \ldots, a_m}.$$

- The measurement errors Δx_{ki} are usually small.
- Thus, the differences $\Delta a_j = \widetilde{a}_j - a_j$ are also small.
- We can keep only linear terms in the Taylor expansion: $Y_k = y_k - \sum\limits_{j=1}^{m} b_j \cdot \Delta a_j - \sum\limits_{i=1}^{n} b_{ki} \cdot \Delta x_{ki}$, where:

$$Y_k = f(\widetilde{a}_1, \ldots, \widetilde{a}_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}), \;\; b_j = \frac{\partial f}{\partial a_j}, \;\; b_{jk} = \frac{\partial f}{\partial x_{ki}}.$$

11. Analysis of the Problem

- For each Δa_j, the min and max values of Y_k are:

$$\underline{Y}_k = Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki};$$

$$\overline{Y}_k = Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}.$$

- We want some values $Y_k \in [\underline{Y}_k, \overline{Y}_k]$ to be in $[\underline{y}_k, \overline{y}_k]$, i.e., that $[\underline{Y}_k, \overline{Y}_k] \cap [\underline{y}_k, \overline{y}_k] \neq \emptyset$.
- This is equivalent to $\underline{y}_k \leq \overline{Y}_k$ and $\underline{Y}_k \leq \overline{y}_k$.
- Thus, we need to optimize a linear expression under linear inequalities.
- For such *linear programming* (LP) problems, there are efficient algorithms.

12. Algorithm for the Interval Case

- We know the expression $f(a_1, \ldots, a_m, x_1, \ldots, x_n)$.
- We know the measurement results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$, and accuracies Δ_k and Δ_{ki}.
- First, we use Least Squares to find $\widetilde{a}_1, \ldots, \widetilde{a}_m$.
- Then, we compute $\underline{y}_k = \widetilde{y}_k - \Delta_k$, $\overline{y}_k = \widetilde{y}_k + \Delta_k$, and the partial derivatives b_j and b_{ki}.
- $\underline{a}_{j_0}$ ($\overline{a}_{j_0}$) is the solution to the following LP problem: minimize (maximize) a_{j_0} under the constraints

$$\underline{y}_k \le Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}, \quad 1 \le k \le K;$$

$$Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki} \le \overline{y}_k, 1 \le k \le K.$$

13. How to Use These Formulas to Estimate y?

- What if we now need to predict the value y corresponding to given values $x_1, \ldots, x_m$?
- In this case, $y = f(a_1, \ldots, a_m, x_1, \ldots, x_n) =$

$$f(\widetilde{a}_1 - \Delta a_1, \ldots, \widetilde{a}_m - \Delta a_m, x_1, \ldots, x_n) = \widetilde{y} - \sum_{j=1}^{M} B_j \cdot \Delta a_j,$$

where $\widetilde{y} = f(\widetilde{a}_1, \ldots, \widetilde{a}_m, x_1, \ldots, x_n)$, $B_j \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial a_j}_{|a_k = \widetilde{a}_k, x_i}$.

- The smallest possible value $\underline{y}$ of y can be found by minimizing $\widetilde{y} - \sum_{j=1}^{m} B_j \cdot \Delta a_j$ under the same constraints.
- The largest possible value $\overline{y}$ of y can be found by maximizing the expression $\widetilde{y} - \sum_{j=1}^{m} B_j \cdot \Delta a_j$.

14. What if We Underestimated the Measurement Inaccuracy?

- In practice, the constraints were often inconsistent.
- So, we underestimated the measurement inaccuracy.
- Since measuring y is the most difficult part, most probably we underestimated the accuracies of measuring y.
- Let's denote the ignored part of y-error by ε.
- Then, we should have $|\Delta y_k| \le \Delta_k + \varepsilon$.
- It's reasonable to look for the smallest $\varepsilon > 0$ s.t. constraints are consistent, i.e., minimize $\varepsilon > 0$ under:

$$\widetilde{y}_k - \Delta_k - \varepsilon \le Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki},$$

$$Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki} \le \widetilde{y}_k + \Delta_k + \varepsilon.$$

15. Simplest Case: Linear Dependence on One Variable $y = a \cdot x + b$

- Let's consider the case $a > 0$ ($a < 0$ is similar).
- In this case, the range of $a \cdot x + b$ is $[a \cdot \underline{x}_k + b, a \cdot \overline{x}_k + b]$.
- This interval intersects with $[\underline{y}_k, \overline{y}_k]$ if

$$a \cdot \underline{x}_k + b \leq \overline{y}_k \text{ and } \underline{y}_k \leq a \cdot \overline{x}_k + b.$$

- So, once we know a, we have the following lower bounds and upper bounds for b:

$$\underline{y}_k - a \cdot \overline{x}_k \leq b \text{ and } b \leq \overline{y}_k - a \cdot \underline{x}_k.$$

- Such a value b exists if and only if every lower bound for b is $\leq$ every upper bound for b:

$$\underline{y}_k - a \cdot \overline{x}_k \leq \overline{y}_\ell - a \cdot \underline{x}_\ell \text{ for all } k \text{ and } \ell.$$

- This is equivalent to $\overline{y}_\ell - \underline{y}_k \geq a \cdot (\underline{x}_\ell - \overline{x}_k)$.

16. Case When $y = a \cdot x + b$ (cont-d)

- We have $\overline{y}_\ell - \underline{y}_k \ge a \cdot (\underline{x}_\ell - \overline{x}_k)$.
- If $\underline{x}_\ell - \overline{x}_k > 0$, $a \le \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}$; if $\underline{x}_\ell - \overline{x}_k < 0$, $a \ge \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}$.
- Thus, the range $[\underline{a}, \overline{a}]$ for a goes from the largest of the lower bounds to the smallest of the upper bounds:

$$\underline{a} = \max_{k,\ell:\ \underline{x}_\ell < \overline{x}_k} \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}; \quad \overline{a} = \min_{k,\ell:\ \underline{x}_\ell > \overline{x}_k} \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}.$$

- Similarly, $a \cdot \underline{x}_k + b \le \overline{y}_k$ and $\underline{y}_k \le a \cdot \overline{x}_k + b$ is equivalent to: $a \cdot x_k \le \overline{y}_k - b$ and $\overline{y}_k - b \le a \cdot \overline{x}_k$.
- If $\underline{x}_k > 0$, $a \le \dfrac{\overline{y}_k}{\underline{x}_k} - \dfrac{1}{\underline{x}_k} \cdot b$; if $\underline{x}_k < 0$, $\dfrac{\overline{y}_k}{\underline{x}_k} - \dfrac{1}{\underline{x}_k} \cdot b \le a$.
- If $\overline{x}_k > 0$, $\dfrac{\underline{y}_k}{\overline{x}_k} - \dfrac{1}{\overline{x}_k} \cdot b \le a$; if $\overline{x}_k > 0$, $a \le \dfrac{\underline{y}_k}{\overline{x}_k} - \dfrac{1}{\overline{x}_k} \cdot b$.

17. Case When $y = a \cdot x + b$ (cont-d)

- Inequalities $A_p + B_p \cdot b \le a$, $a \le C_q + D_q \cdot b$ are consistent if every lower bound $\le$ every upper bound:

$$A_p + B_p \cdot b \le C_q + D_q \cdot b \Leftrightarrow (D_q - B_p) \cdot b \ge A_p - C_q.$$

- So, similarly to the a-case, we get:

$$\underline{b} = \max_{p,q:\ D_q > B_p} \frac{A_p - C_q}{D_q - B_p}; \quad \overline{b} = \max_{p,q:\ D_q < B_p} \frac{A_p - C_q}{D_q - B_p}.$$

- If we underestimated the measurement inaccuracy, we get the new bounds $\underline{y}_k - \varepsilon$ and $\overline{y}_k + \varepsilon$.

- So, if $\underline{x}_\ell > \overline{x}_k$, we get $a \le \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \dfrac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon$, else $a \ge \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \dfrac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon.$

18. What If We Underestimate Measurement Uncertainty

- If $\underline{x}_\ell > \overline{x}_k$, we get $a \le \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \dfrac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon$, else $a \ge \dfrac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \dfrac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon$.
- Inequalities $A_p + B_p \cdot \varepsilon \le a$ and $a \le C_q + D_q \cdot \varepsilon$ are consistent if every lower bound $\le$ every upper bound:

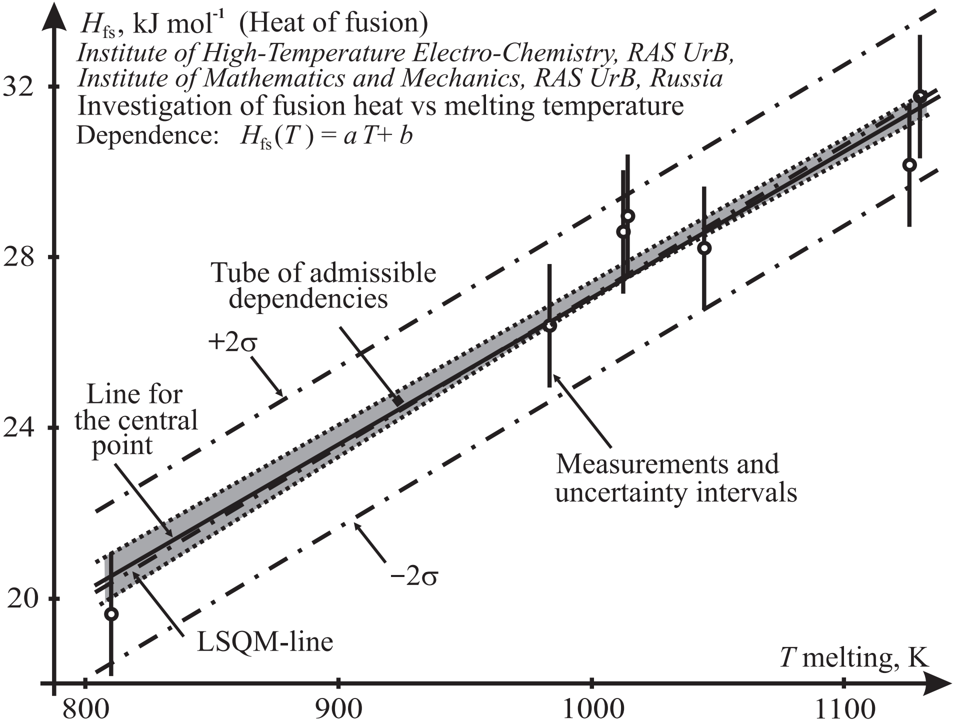
$$A_p + B_p \cdot \varepsilon \le C_q + D_q \cdot \varepsilon \Leftrightarrow (D_q - B_p) \cdot \varepsilon \ge A_p - C_q.$$

- So, the desired lower bound for ε for b is equal to the largest of the lower bounds:

$$\varepsilon = \max_{p,q:\ D_q > B_p} \frac{A_p - C_q}{D_q - B_p}.$$

19. Case Study

- One of the important engineering problems is the problem of storing energy:
  - solar power and wind turbines provide access to large amounts of renewable energy,
  - but this energy is not always available – the sun goes down, the wind dies,
  - and storing it is difficult.
- Similarly, electric cars are clean, but we spend a lot of weight on the batteries.
- We want batteries with high energy density.
- One of the most promising directions is using molten salt batteries, including liquid metal batteries.
- Melting energy E linearly depends on temperature T: $E = a \cdot T + b$. What are a and b?



H_{fs}, kJ mol^{-1} (Heat of fusion)
Institute of High-Temperature Electro-Chemistry, RAS UrB,
Institute of Mathematics and Mechanics, RAS UrB, Russia
Investigation of fusion heat vs melting temperature
Dependence: $H_{fs}(T) = a\,T + b$
32
28
24
20
Tube of admissible dependencies
+2σ
Line for the central point
Measurements and uncertainty intervals
−2σ
LSQM-line
T melting, K
800
900
1000
1100

Figure 1:

20. Results of Our Analysis

- We generated two different bounds on y:
  - bounds based on interval estimates, and
  - 2σ-bounds coming from the traditional statistical analysis.
- It turned out that the interval results are an order of magnitude smaller that the statistical ones.
- A similar improvement was observed in other applications ranging from catalysis and to mechanics.

21. Conclusions

- Traditional engineering techniques assume that the measurement errors are normally distributed.
- In practice, the distribution of measurement errors is indeed often close to normal.
- Often, however, we also have an additional information about measurement uncertainty.
- Namely, we also know the upper bounds Δ on the corresponding measurement errors.
- Based on the measurement result $\widetilde{x}$, the actual value x is in the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- We can use interval computations techniques to estimate the accuracy of the result of data processing.
- Example: for linear models, we can use linear programming techniques to compute the corr. bounds.

22. Conclusions (cont-d)

- Which approaches leads to more accurate estimates:
  - the traditional approach, when we ignore the upper bounds and only consider the probabilities, or
  - the interval approach, we only take into account the bounds and ignore probabilities?
- When the number of measurements n increases, the interval estimates become more accurate.
- We show that interval techniques indeed lead to much more accurate estimates.
- So, we recommend to try interval techniques: they may lead to more accurate estimates.
- For linear interval models, we also provide a faster algorithm.

23. Acknowledgments

This work was supported in part:

- by the Russian Foundation for Basic Research grant 15-01-07909,
- by the National Science Foundation grants HRD-0734825, HRD-1242122, and DUE-0926721,
- by an award from Prudential Foundation.