11-2020

How to Find the Dependence Based on Measurements with Unknown Accuracy: Towards a Theoretical Justification for Midpoint and Convex-Combination Interval Techniques and Their Generalizations

Somsak Chanaim
Chiang Mai University, somsak_ch@cmu.ac.th

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-20-110

How to Find the Dependence Based on Measurements with Unknown Accuracy: Towards a Theoretical Justification for Midpoint and Convex-Combination Interval Techniques and Their Generalizations

Somsak Chanaim and Vladik Kreinovich

Abstract In practice, we often need to find regression parameters in situations when for some of the values, we have several results of measuring this same value. If we know the accuracy of each of these measurements, then we can use the usual statistical techniques to combine the measurement results into a single estimate for the corresponding value. In some cases, however, we do not know these accuracies, so what can we do? In this paper, we describe two natural approaches to solving this problem. In addition to describing general techniques, our results also provide a theoretical explanation for several semi-heuristic ideas proposed for solving an important particular case of this problem – the case when we deal with interval uncertainty.

1 Formulation of the Problem

General problem. In many practical situations:

- we know that the general form of the dependence of a quantity y on quantities $x_1, \ldots, x_n$, i.e., we know that $y = f(x_1, \ldots, x_n, c_1, \ldots, c_m)$ for some known function $f(x_1, \ldots, x_n, c_1, \ldots, c_m)$, but
- we do not know the values of the parameters $c_1, \ldots, c_m$; these values need to be determined empirically, from the known results of observations and measurements.

This general situation is known as *regression*.

Somsak Chanaim
International College of Digital Innovation, Chiang Mai University, Chiang Mai 50200, Thailand
e-mail: somsak_ch@cmu.ac.th

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

A simple example. The simplest example if when $n = 1$, and y is simply proportional to x_1, with an unknown coefficient of proportionality c_1, so that $y = c_1 \cdot x_1$. In this case, we have $m = 1$ parameter c_i, and $f(x_1, c_1) = c_1 \cdot x_1$.

Econometric example. We may want to know the parameter β that describes, for a given stock, how the difference $r - r_f$ between the stock's rate of return r and the risk-free interest rate r_f depends on the difference $r_m - r_f$ between the overall market's rate of return r_m and the value r_f:

$$r - r_f = \beta \cdot (r_m - r_f).$$

General problem: usual case. Usually, we have several (K) cases $k = 1, \ldots, K$ in each of which we measure x_i and y, resulting in the values $x_1^{(k)}, \ldots, x_n^{(k)}$ and $y^{(k)}$. In this case, to find the values of the parameters c_i, a reasonable idea is to apply the Least Squares method (see, e.g., [5]), i.e., to find the values $c_1, \ldots, c_m$ of the parameters that minimize the expression

$$\sum_{k=1}^{K} \left(y^{(k)} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}, c_1, \ldots, c_m \right) \right)^2. \tag{1}$$

Alternatively, we can minimize the sum of the absolute values of the differences $y^{(k)} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}, c_1, \ldots, c_m \right)$, or any other appropriate objective function.

What if for each case, we have several measurement results? Sometimes, in each case k, we have several different measurement results of each of the variables:

- for each k and i, instead of a single measurement result $x_i^{(k)}$, we have several values $x_{i1}^{(k)}, \ldots, x_{iv_i}^{(k)}$ measured, in general, by several different measuring instruments, and
- for each k, instead of a single result $y^{(k)}$ of measuring y, we have several values $y_1^{(k)}, \ldots, y_v^{(k)}$ measured, in general, by several different measuring instruments.

In such situation, a natural idea is to do the following:

- first, for each k and for each i, we use all the results $x_{i1}^{(k)}, \ldots, x_{iv_i}^{(k)}$ of measuring x_i to come up with a single estimate $x_i^{(k)}$;
- then, for each k, we use all the results $y_1^{(k)}, \ldots, y_v^{(k)}$ of measuring y to come up with a single estimate $y^{(k)}$;
- then, we find the values of the parameters $c_1, \ldots, c_m$ that minimize the objective function (1) – or the corresponding alternative objective function.

To implement this idea, we need to be able to combine several estimates into a single one.

Econometric example. The stock price fluctuates during the day. The usual economic assumption is that:

- on any day, there is the fair price of the stock – the price that reflects its current value and its prospects;
- this fair price changes rarely – definitely rarely several times a day, it only changes based on the new information;
- on the other hand, the observed minute-by-minute price changes all the time, because it is obtained by adding some random fluctuations to the fair price.

In this example, we do not know the fair daily price of the stock x_i, but we can measure several characteristics that provide an approximate description of this fair price: the smallest daily price $x_{i1}^{(k)}$, the largest daily price $x_{i2}^{(k)}$, the closing price $x_{i3}^{(k)}$, the starting price $x_{i4}^{(k)}$, etc. If we limit ourselves to these four characteristics, then we have $v_i = 4$.

Instead of these four measurement results, we can use only two: the smallest daily price $x_{i1}^{(k)}$ and the largest daily price $x_{i2}^{(k)}$. In this case, what we know is an *interval* $\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right]$ that contains the actual (unknown) fair price $x_i^{(k)}$ on day k.

Comment. There are other practical examples where, as a result of measurements, we get a lower bound $x_{i1}^{(k)}$ and an upper bound $x_{i2}^{(k)}$ for the desired quantity $x_i^{(k)}$, i.e., where, as a result of the measurements, we get an interval $\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right]$ that contains the actual (unknown) value $x_i^{(k)}$.

We can naturally combine measurement results when we know the accuracy of each measurement. In many practical situations, we know the accuracy of different measuring instruments. For example:

- for each input i and for each instrument $j = 1, \ldots, v_i$ used to measure x_i, we know the corresponding standard deviation σ_{ij}, and
- for each instrument $j = 1, \ldots, v$ used to measure y, we know the corresponding standard deviation σ_j.

In this case, a natural idea for estimating $x_i^{(k)}$ is to use the least squares approach, i.e., to minimize the sum

$$\sum_{j=1}^{v_i} \frac{\left(x_i^{(k)} - x_{ij}^{(k)}\right)^2}{\sigma_{ij}^2}.$$

This minimization results in the estimate

$$x_i^{(k)} = \sum_{j=1}^{v_i} w_{ij} \cdot x_{ij}^{(k)}, \tag{2}$$

where

$$w_{ij} = \frac{\sigma_{ij}^{-2}}{\sum\limits_{j'=1}^{v_i} \sigma_{ij'}^{-2}}. \tag{3}$$

Similarly, a natural idea for estimating $y^{(k)}$ is to use the least squares approach, i.e., to minimize the sum

$$\sum_{j=1}^{v} \frac{\left(y^{(k)} - y_j^{(k)}\right)^2}{\sigma_j^2}.$$

This minimization results in the estimate

$$y^{(k)} = \sum_{j=1}^{v} w_j \cdot y_j^{(k)}, \tag{4}$$

where

$$w_j = \frac{\sigma_j^{-2}}{\sum\limits_{j'=1}^{v} \sigma_{j'}^{-2}}. \tag{5}$$

Comment. In both cases, the coefficients w add to 1: $\sum\limits_{j=1}^{v_i} w_{ij} = 1$ and $\sum\limits_{j=1}^{v} w_j = 1$.

Remaining problem. In some cases – e.g., in the econometric example – we do not know the corresponding accuracies. What shall we do?

This is a problem that we consider in this paper. Specifically, we describe two natural general solutions – and we explain how each of them is related to previously proposed methods. It turns out that this way, several previous proposed semi-empirical methods can be theoretically justified.

2 First Approach: Laplace's Indeterminacy Principle

Main idea. In its most general form, Laplace's Indeterminacy Principle states that if we have no reason to assume that one quantity is smaller or larger than the other one, then it is reasonable to assume that these two quantities are equal to each other; see, e.g., [4].

Let us apply this idea to our problem. For each i, we have several unknown values σ_{ij}. Since we have no reason to believe that one of these values is larger, we conclude that all these values are equal to each other: $\sigma_{i1} = \sigma_{i2} = \ldots$ In this case, formula (3) leads to $w_{ij} = \dfrac{1}{v_i}$, and the estimate (2) becomes simply the arithmetic mean

$$x_i^{(k)} = \frac{1}{v_i} \cdot \sum_{j=1}^{v_i} x_{ij}^{(k)}. \tag{6}$$

Similarly, since we have no reason to believe that one of the values σ_j is larger, we conclude that all these values are equal to each other: $\sigma_1 = \sigma_2 = \ldots$ In this case, formula (5) leads to $w_j = \dfrac{1}{v}$, and the estimate (4) becomes the arithmetic mean

$$y^{(k)} = \frac{1}{v} \cdot \sum_{j=1}^{v} y_j^{(k)}. \tag{7}$$

Interval case. In the case when the two estimates are the two endpoints of the interval, formulas (6)-(7) result in a midpoint of this interval. Thus, in situations when we only know the intervals $\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right]$ and $\left[y_1^{(k)}, y_2^{(k)}\right]$ containing the desired values x_i and y, this approach recommends applying the regression technique to midpoints $x_i^{(k)} = \dfrac{x_{i1}^{(k)} + x_{i2}^{(k)}}{2}$ and $y^{(k)} = \dfrac{y_1^{(k)} + y_2^{(k)}}{2}$ of these intervals.

Comment. The use of midpoints is exactly what was proposed in [1]. Thus, our analysis provides a theoretical explanation for this semi-heuristic method.

3 Second Approach: Using the Known Dependence Between x_i and y

Alternative idea. We consider the case when do not know the measurement accuracies σ_{ij} and σ_j, so we cannot use these accuracies to find the coefficients w_{ij} and w_j. In other words, we do not know which linear combinations of the measurement results most adequately represent the actual values $x_i^{(k)}$ and $y^{(k)}$.

A natural idea is to take into account that the actual (unknown) values x_i and y should satisfy the formula $y = f(x_1, \ldots, x_n, c_1, \ldots, c_m)$. Thus, it is reasonable to select the coefficients w_{ij} and w_j for which the resulting linear combination $y^{(k)}$ is as close as possible to the value $f\left(x_1^{(k)}, \ldots, x_n^{(k)}, c_1, \ldots, c_m\right)$. To be more precise, we find the parameters $c_1, \ldots, c_m$ *and* the coefficients w_{ij} and w_j from the condition that the expression (1) (or any other selected objective function) attains its smallest possible value, where $x_i^{(k)}$ and $y^{(k)}$ are determined by the formulas (2) and (4).

In this case, the minimized objective function (1) takes the form

$$\sum_{k=1}^{K} \left(\sum_{j=1}^{v} w_j \cdot y_j^{(k)} - f\left(\sum_{j=1}^{v_1} w_{1j} \cdot x_{1j}^{(k)}, \ldots, \sum_{j=1}^{v_n} w_{nj} \cdot x_{nj}^{(k)}, c_1, \ldots, c_m \right) \right)^2.$$

Interval case. In the interval case, when we know the intervals $\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right]$ and $\left[y_1^{(k)}, y_2^{(k)}\right]$, the idea is to select appropriate convex combinations $x_i^{(k)} = w_{i1} \cdot x_{i1}^{(k)} + (1 - w_{i1}) \cdot x_{i2}^{(k)}$ and $y^{(k)} = w_1 \cdot y_1^{(k)} + (1 - w_1) \cdot y_2^{(k)}$, i.e., coefficients for which the following expression is the smallest possible:

$$\sum_{k=1}^{K} \left(w_1 \cdot y_1^{(k)} + (1 - w_1) \cdot y_2^{(k)} - f\left(w_{11} \cdot x_{11}^{(k)} + (1 - w_{11}) \cdot x_{12}^{(k)}, \ldots, c_1, \ldots \right) \right)^2.$$

Comment. This idea of using convex combinations has indeed been proposed and successfully used; see, e.g., [2, 3]. Thus, our analysis provides a theoretical explanation for this semi-heuristic idea as well.

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), and 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.

This work was also supported by the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand.

References

1. L. Billard and E. Diday, "Symbolic Regression Analysis", In: K. Jajuga, A. Sokolowski, and H. H. Bock (eds.), *Classification, Clustering, and Data Analysis*, Springer, Berlin, Heidelberg, 2002, pp. 281–288, https://doi.org/10.1007/978-3-642-56181-8_31
2. S. Chanaim, S. Sriboonchitta, and C. Rungruang, "A convex combination method for linear regression with interval data", In: *Proceeding of the 5th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making IUKM'2016*, Springer, 2016, pp. 469–480.
3. S. Chanaim, C. Khiewngamdee, S. Sriboonchitta, and C. Rungruang, "A convex combination method for quantile regression with interval data", In: Ly H. Anh, Le Si Dong, V. Kreinovich, and Nguyen Ngoc Thach (eds.) *Econometrics for Financial Applications*, Springer, Cham, Switzeerland, 2018, pp. 440–449, https://doi.org/10.1007/978-3-319-73150-6_35
4. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
5. D. J. Sheskin, *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC, London, UK, 2011.