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[1] and Vladik Kreinovich[2]
[1]International College of Digital Innovation, Chiang Mai University,
Chiang Mai 50200, Thailand somsak_ch@cmu.ac.th
[2]Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, Texas 79968, USA, vladik@utep.edu

1. 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*.

2. Examples

- The simplest example is:
  - 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:*
  - let r denote the stock's rate of return,
  - let r_f denote the risk-free interest rate,
  - let r_m denote the overall market's rate of return.
  - we may want to know the parameter β that describes how $r - r_f$ depends on $r_m - r_f$:

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

3. General Problem: Usual Case

- Usually, we have several (K) cases $k = 1, \ldots, K$.
- In each case, 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.
- In this method, we find the values $c_1, \ldots, c_m$ of the parameters that minimize the expression

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

- Alternatively, we can minimize the sum of the absolute values of the differences (or other expression):

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

4. 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 $x_{i1}^{(k)}, \ldots, x_{iv_i}^{(k)}$;
  - these values are measured, in general, by several different measuring instruments.
- For each k:
  - instead of a single result $y^{(k)}$ of measuring y, we have several values $y_1^{(k)}, \ldots, y_v^{(k)}$;
  - these values are measured, in general, by several different measuring instruments.

5. Several Measurement Results (cont-d)

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

6. 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.
- This price 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.

7. Econometric Example (cont-d)

- However, 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 $v_i = 4$.

8. Econometric Example (cont-d)

- 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.
- 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 a quantity $x_i^{(k)}$,
  - as a result of the measurements, we get an interval $\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right]$ that contains the actual value $x_i^{(k)}$.

9. We Can Naturally Combine Measurement Results When We Know the Accuracies

- In many practical situations, we know the accuracy of different measuring instruments:
- For each input i and for each instrument $j = 1, \ldots, v_i$ used to measure x_i, we know the st. dev. σ_{ij}.
- For each instrument $j = 1, \ldots, v$ used to measure y, we know the corresponding standard deviation σ_j.

10. We Can Naturally Combine Measurement Results When We Know the Accuracies (cont-d)

- 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

$$x_i^{(k)} = \sum_{j=1}^{v_i} w_{ij} \cdot x_{ij}^{(k)}, \text{ where } w_{ij} = \frac{\sigma_{ij}^{-2}}{\sum\limits_{j'=1}^{v_i} \sigma_{ij'}^{-2}}.$$

11. We Can Naturally Combine Measurement Results When We Know the Accuracies (cont-d)

- 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

$$y^{(k)} = \sum_{j=1}^{v} w_j \cdot y_j^{(k)}, \text{ where } w_j = \frac{\sigma_j^{-2}}{\sum\limits_{j'=1}^{v} \sigma_{j'}^{-2}}.$$

- In both cases, the coefficients w add to 1:

$$\sum_{j=1}^{v_i} w_{ij} = 1 \text{ and } \sum_{j=1}^{v} w_j = 1.$$

12. 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 talk.
- Specifically, we describe two natural general solutions.
- We also 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.

13. First Approach: Laplace's Indeterminacy Principle

- 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.
- Let us apply this idea to our problem.
- For each i, we have several unknown values σ_{ij}.
- We have no reason to believe that one of these values is larger.
- So, we conclude that all these values are equal to each other: $\sigma_{i1} = \sigma_{i2} = \ldots$

14. Indeterminacy Principle (cont-d)

- In this case, the above formula leads to $w_{ij} = \frac{1}{v_i}$ and to the arithmetic mean $x_i^{(k)} = \frac{1}{v_i} \cdot \sum_{j=1}^{v_i} x_{ij}^{(k)}$.
- We have no reason to believe that one of the values σ_j is larger.
- So, we conclude that all these values are equal to each other: $\sigma_1 = \sigma_2 = \ldots$
- In this case, $w_j = \frac{1}{v}$, and we get the arithmetic mean

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

15. Interval Case

- Interval case is when the two estimates are the two endpoints of the interval.
- In this case, the above formulas 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 x_i and y,
  - this approach recommends applying the regression technique to midpoints

$$x_i^{(k)} = \frac{x_{i1}^{(k)} + x_{i2}^{(k)}}{2} \text{ and } y^{(k)} = \frac{y_1^{(k)} + y_2^{(k)}}{2}.$$

- The use of midpoints is exactly what was proposed by Billard and Diday in 2002.
- Thus, our analysis provides a theoretical explanation for this semi-heuristic method.

16. Second Approach: Using the Known Dependence Between x_i and y

- 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.
- 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

$$f\left(x_1^{(k)}, \ldots, x_n^{(k)}, c_1, \ldots, c_m\right).$$

17. Second Approach (cont-d)

- We find c_i, w_{ij}, and w_j that minimizes the selected objective function.
- For least squares, the minimized objective function takes the form

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

18. Interval Case

- The interval case is when we only know the intervals

$$\left[x_{i1}^{(k)}, x_{i2}^{(k)}\right] \text{ and } \left[y_1^{(k)}, y_2^{(k)}\right].$$

- In this case, the idea is to select appropriate convex combinations

$$x_i^{(k)} = w_{i1} \cdot x_{i1}^{(k)} + (1 - w_{i1}) \cdot x_{i2}^{(k)} \text{ and}$$

$$y^{(k)} = w_1 \cdot y_1^{(k)} + (1 - w_1) \cdot y_2^{(k)}.$$

- In other words, we select the coefficients for which the following expression is the smallest possible:

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

19. Interval Case (cont-d)

- This idea of using convex combinations has indeed been proposed and successfully used.
- Thus, our analysis provides a theoretical explanation for this semi-heuristic idea as well.

20. 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);
- 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.

21. Bibliography

- 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

- 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.

22. Bibliography (cont-d)

- 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
- E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
- D. J. Sheskin, *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC, London, UK, 2011.