Processing Quantities with Heavy-Tailed Distribution of Measurement Uncertainty: How to Estimate the Tails of the Results of Data Processing

Michal Holčapek[1] and Vladik Kreinovich[2]
[1]Centre of Excellence IT4Innovations, University of Ostrava,
Institute for Research and Applications of Fuzzy Modeling,
Ostrava, Czech Republic, michal.holcapek@osu.cz
[2]University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

1. Need for Data Processing

- We are often interested in the values of a quantity y which is not easy to measure directly, e.g.:
  - tomorrow's weather,
  - distance to a faraway planet,
  - amount of oil in an oil well.
- In such situations in which we cannot measure y *directly*, we can often measure y *indirectly*, i.e.:
  - measure auxiliary quantities $x_1, \ldots, x_n$ related to the desired quantity y by a known relation
$$y = f(x_1, \ldots, x_n);$$
  - use the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring the quantities x_i to compute the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- The process of computing $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is known as *data processing*.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

2. Need to Estimate Uncertainty of the Result of Data Processing

- Measurements are never 100% accurate.
- In general, the measurement results $\widetilde{x}_i$ are somewhat different from the actual values x_i: $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i \neq 0$.
- Since $\widetilde{x}_i \neq x_i$, the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$.
- Often, there is additional difference since the dependence between y and x_i is only approximately known.
- It is therefore important to gauge how much the actual value y can differ from this estimate.
- In other words, we need to gauge the uncertainty of the result of data processing.

3. Traditional Statistical Approach

- Usually, there are many different (and independent) factors which contribute to the measurement error.
- Due to Central Limit Theorem, the distr. of the joint effect of numerous independent factors is $\approx$ normal.
- To describe a normal distribution, it is sufficient to know the mean μ and the standard deviation σ.
- If $\mu \neq 0$, we can compensate for this bias, so each Δx_i is normally distributed with mean 0 and st. dev. σ_i.
- The measurement errors Δx_i are usually small, so

$$\Delta y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \Delta x_i,$$

- Thus, Δy is normally distributed with 0 mean and variance $\sigma^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \cdot \sigma_i^2$.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

4. Heavy-Tailed Distributions

- In practice, the probability distribution of the measurement error is often different from normal.
- In many such situations, the variance is infinite.
- Such distributions are called *heavy-tailed*.
- Mandelbrot (of fractal fame) found that price fluctuations follows the Pareto power-law

$$\rho(x) = A \cdot x^{-\alpha}, \quad \alpha \approx 2.7.$$

- For this empirical value α, variance is infinite.
- We need to estimate Δy for the case when distributions for Δx_i are heavy-tailed.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

5. Result for Addition $y = f(x_1, x_2) = x_1 + x_2$

- For addition, $\Delta y = \Delta x_1 + \Delta x_2$.
- Let us assume that:
  - the measurement error Δx_1 of the first input has a tail with asymptotics $\rho_1(\Delta x_1) \sim A_1 \cdot |\Delta x_1|^{-\alpha_1}$;
  - the measurement error Δx_2 of the second input has a tail with asymptotics $\rho_2(\Delta x_2) \sim A_2 \cdot |\Delta x_2|^{-\alpha_2}$,
- Then the tail for Δy has the asymptotics

$$\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha}, \text{ with } \alpha = \min(\alpha_1, \alpha_2).$$

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

6. Case of a General Linear Combination

- Let us assume that:
  - $y = a_0 + \sum_{i=1}^{m} a_i \cdot x_i$; and
  - the measurement error Δx_i of the i-th input has a tail with asymptotics

$$\rho_i(\Delta x_i) \sim A_i \cdot |\Delta x_i|^{-\alpha_i}.$$

- Then the tail for Δy has the asymptotics

$$\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha} \text{ with } \alpha = \min(\alpha_1, \ldots, \alpha_m).$$

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

7. Case of the Product $y = f(x_1, x_2) = x_1 \cdot x_2$: Result

- Let us assume that:
  - the measurement error Δx_1 of the first input has a tail with asymptotics $\rho_1(\Delta x_1) \sim A_1 \cdot |\Delta x_1|^{-\alpha_1}$;
  - the measurement error Δx_2 of the second input has a tail with asymptotics $\rho_2(\Delta x_2) \sim A_2 \cdot |\Delta x_2|^{-\alpha_2}$.
- Then, $\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha}$ with $\alpha = \min(\alpha_1, \alpha_2)$.
- Similar formulas hold for an arbitrary combination $y = a_0 \cdot \prod_{i=1}^{m} x_i^{a_i}$:
  - when the meas. error Δx_i of the the i-th input has a tail with asymptotics $\rho_i(\Delta x_i) \sim A_i \cdot |\Delta x_i|^{-\alpha_i}$,
  - then the tail for Δy has the asymptotics $\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha}$ with $\alpha = \min(\alpha_1, \ldots, \alpha_m)$.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

8. Epistemic vs Aleatory Uncertainty

- The main objective of this paper is to deal with measurement (epistemic) uncertainty.
- However, the same formula can be used if we have *aleatory* uncertainty.
- For example, we can use these formulas to analyze what happens if:
  - we have a population of two-job individuals;
  - we know the distribution $\rho_1(x_1)$ of first salaries;
  - we know the distribution $\rho_2(x_2)$ of second salaries;
  - we know that these distributions are independent, and
  - we want to find the distribution of the total salary

$$y = x_1 + x_2.$$

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

9. General Asymptotics Remains a Challenge

- For a normal distribution, $\text{Prob}(|\Delta x_i| > 6\sigma_i) \approx 10^{-6}\%$.
- Such large deviations can be safely ignored, so measurement errors Δx_i are small.
- So, we can approximate the dependence $y = f(x_1, \ldots, x_n)$ by linear terms in its Taylor expansion.
- For $\rho(\Delta x) \approx A \cdot |\Delta x|^{-\alpha}$ with $\alpha = 2$, the probability of Δx exceeding 6σ is $\approx 6^{-2} \approx 3\%$: quite probable.
- Even deviations of size 100σ are possible: they occur once every 10,000 trials.
- For such large deviations, we can no longer ignore quadratic or higher order terms.
- So, we can no longer reduce any smooth function to its linear approximation.
- Each smooth function has to be treated separately.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

10. Need to Go from Asymptotics to a Complete Description

- So far, we only found the asymptotics of the probability distribution for the approximation error $\Delta y = \widetilde{y} - y$.
- It is desirable to find the whole distribution for Δy.
- For that, in addition to the exponent α, we also need to find the following:
  - the coefficient A at the asymptotic expression
$$\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha};$$
  - the threshold Δ_0 after which this asymptotic holds; and
  - the probability density $\rho(\Delta y)$ on $[-\Delta_0, \Delta_0]$.
- Once we have this info for Δx_1 and Δx_2, we can use the formula
$$\rho(\Delta y) = \int \rho_1(\Delta x_1) \cdot \rho_2(\Delta y - \Delta x_1)\, d(\Delta x_1).$$

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

11. What If We Only Have Partial Information about the Distributions of Δx_i

- In practice, we only have partial information about the probability distributions $\rho_i(\Delta x_i)$. So:
  - instead of the exact values of the corresponding cumulative distribution functions
$$F(x) \stackrel{\text{def}}{=} \text{Prob}(X \le x),$$
  - we only know an interval $[\underline{F}(x), \overline{F}(x)]$ of possible values of $F(x)$.
- The corresponding interval-valued function $[\underline{F}(x), \overline{F}(x)]$ is known as a *probability box* (or *p-box*, for short).
- Several algorithms are known for propagating p-boxes via data processing.
- It is desirable to extend these algorithms to also cover interval uncertainty about the values A, α, and Δ_0.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

12. Acknowledgments

This work was supported in part:

- by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070),
- by the National Science Foundation grants HRD-0734825, HRD-1242122, DUE-0926721,
- by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and
- by a grant N62909-12-1-7039 from the Office of Naval Research.

13. Fractals: Reminder

- Mandelbrot studied not only the local price fluctuations.
- He also studied the global geometry of the curves describing the dependence of price on time.
- It turned out that this analysis is closely related to the notion of dimension.
- For each $\varepsilon > 0$, we can *ε-approximate* the set S by a finite set $S' = \{s_1, \ldots, s_n\}$, in the sense that:
  - every point s from the set S is ε-close to some point $s_i \in S'$, and
  - vice versa, every point $s_i \in S'$ is ε-close to some point $s \in S$.
- For each set S, we can have ε-approximating sets S' with different number of elements.

14. Fractals (cont-d)

- For each ε, we can find the number of elements $N_\varepsilon(S)$ in the *smallest* ε-approximating finite set.
- For a 1-D smooth curve S of length L:
  - the smallest number $N_\varepsilon(S)$ is attained
  - if we take the points $s_1, \ldots, s_n \in S$ located at equal distance $\approx 2\varepsilon$ from each other.
  - the number of such points is asymptotically equal to $N_\varepsilon(S) \sim \text{const} \cdot \dfrac{L}{\varepsilon}$.
- For a 2-D smooth surface S of area A:
  - the smallest number $N_\varepsilon(S)$ is attained
  - if we take the points on a rectangular 2-D grid with linear step $\approx \varepsilon$;
  - the number of such points is asymptotically equal to $N_\varepsilon(S) \sim \text{const} \cdot \dfrac{A}{\varepsilon^2}$.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .

15. Fractals (cont-d)

- For a 3-D body S of volume V:
  - the smallest number $N_\varepsilon(S)$ is attained
  - if we take the points on a rectangular 3-D grid with linear step $\approx \varepsilon$;
  - the number of such points is asymptotically equal to $N_\varepsilon(S) \sim \text{const} \cdot \dfrac{V}{\varepsilon^3}$.
- For the price trajectory S, we have $N_\varepsilon(S) \sim \dfrac{C}{\varepsilon^a}$ for a fractional (non-integer) a.
- By analogy with the smooth sets, the value a is called a *dimension* of the trajectory S.
- Thus, the trajectory S is a set of a fractal dimension.
- Mandelbrot called such sets *fractals*.

Need for Data Processing
Need to Estimate . . .
Traditional Statistical . . .
Heavy-Tailed . . .
Result for Addition . . .
Case of a General . . .
Case of the Product . . .
General Asymptotics . . .
What If We Only Have . . .