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, TX 79968, USA, vladik@utep.edu

Abstract. Measurements are never absolutely accurate; so, it is important to estimate how the measurement uncertainty affects the result of data processing. Traditionally, this problem is solved under the assumption that the probability distributions of measurement errors are normal – or at least are concentrated, with high certainty, on a reasonably small interval. In practice, the distribution of measurement errors is sometimes heavy-tailed, when very large values have a reasonable probability. In this paper, we analyze the corresponding problem of estimating the tail of the result of data processing in such situations.

1 Formulation of the Problem

Need for data processing. In many practical situations, we are interested in the values of a quantity y which is not easy (or even impossible) to measure directly: for example, we may be interested in tomorrow's weather, in the distance to a faraway planet, in the amount of oil in an oil well, etc. In such situations in which we cannot measure y *directly*, we can often measure y *indirectly*, i.e.:

- measure the values of auxiliary quantities $x_1, \ldots, x_n$ which are related to the desired quantity y by a known relation $y = f(x_1, \ldots, x_n)$, and then
- use the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring the quantities x_i and the known dependence to compute the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for y.

The process of computing $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is known as *data processing.*

Need to estimating uncertainty of the result of data processing. Measurements are never 100% accurate; so, in general, the measurement results $\widetilde{x}_i$ are somewhat different from the actual values x_i of the corresponding quantities. Because of these measurement errors, the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the desired value $y = f(x_1, \ldots, x_n)$ (often, there is an additional difference cause by the fact that the dependence between y and x_i is only approximately

known). It is therefore important not just to generate an estimate $\widetilde{y}$, but also to gauge how much the actual value y can differ from this estimate, i.e., what is the uncertainty of the result of data processing; see, e.g., [7].

Estimating uncertainty of the result of data processing: traditional statistical approach. Usually, there are many different (and independent) factors which contribute to the measurement error. In many such situations, it is possible to apply the Central Limit Theorem (see, e.g., [9]), according to which, under reasonable conditions, the distribution of the joint effect of numerous independent factors is close to normal. In such situations, it is therefore reasonable to assume that all the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ are independent and normally distributed.

To describe a normal distribution, it is sufficient to know the mean μ and the standard deviation σ. Thus, under the normality assumption, to gauge the distribution of each measurement error Δx_i, we must know the mean μ_i and the standard deviation σ_i of this measurement error. If the known mean is different from 0, this means that this measuring instrument has a bias; we can always compensate for this bias by subtracting the value μ_i from all the measured values. After this subtraction, the mean error will become 0. Thus, without losing generality, we can assume that each measurement error is normally distributed with mean 0 and known standard deviation σ_i.

The traditional way of estimating the resulting uncertainty $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ in y is based on this assumption. Specifically, since the measurement errors Δx_i are usually relatively small, we can expand the expression

$$\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) =$$

$$f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n)$$

in Taylor series in Δx_i, ignore quadratic and higher order terms, and keep only terms in Δx_i in this dependence. As a result, we get an expression

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i,$$

where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$. Based on this expression, we conclude that the linear combination Δy of n independent normally distributed random variables is also normally distributed, its mean value of is 0, and its variance σ^2 is equal to: $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$ (see, e.g., [7]).

Heavy-tailed distributions. There are many practical situations in which the probability distribution for the measurement error is drastically different from normal. In many such situations, the variance is infinite; such distributions are called *heavy-tailed*. Since then, similar heavy-tailed distributions have been empirically found in many other application areas; see, e.g., [1, 8].

These distributions surfaced in the 1960s, when Benoit Mandelbrot, the author of fractal theory, empirically studied the fluctuations and showed [4] that large-scale fluctuations follow the Pareto power-law distribution, where for some x_0, for all $x \geq x_0$, the probability density function has the form $\rho(x) = A \cdot x^{-\alpha}$, for some empirical constants $A > 0$ and $\alpha \approx 2.7$. For this empirical value α, variance is infinite. The above empirical result, together with similar empirical discovery of heavy-tailed laws in other application areas, has led to the formulation of *fractal theory*; see, e.g., [5, 6].

Comments.

- Please note that Mandelbrot's empirical observations only describe the probability density $\rho(x)$ for values $x \geq x_0$; the values $\rho(x)$ for $x < x_0$ can be different.
- In general, the condition that $\int \rho(x)\,dx = 1$ implies that $\alpha > 1$.
- One can easily check that the variance $\int x^2 \cdot \rho(x)\,dx$ is infinite when $\alpha \leq 3$.

Problem. If the measurement errors Δx_i of the inputs x_i are distributed according to the heavy-tailed distributions, then what can we conclude about Δy?

What we do in this paper. In this paper, we provide an answer to the above question for the simplest cases when data processing consists of applying a single arithmetic operation: addition, subtraction, multiplication, or division.

2 Main Results

Case of addition $y = f(x_1, x_2) = x_1 + x_2$. For addition, $\Delta y = \Delta x_1 + \Delta x_2$. When the measurement error Δx_1 of the first input has a tail with asymptotics $\rho_1(\Delta x_1) \sim A_1 \cdot |\Delta x_1|^{-\alpha_1}$ and the measurement error Δx_1 of the first 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 $\rho(\Delta y) \sim A \cdot |\Delta y|^{-\alpha}$ with $\alpha = \min(\alpha_1, \alpha_2)$.

Proof for the case of addition $y = f(x_1, x_2) = x_1 + x_2$. We know that $\rho(\Delta y) = \int \rho_1(\Delta x_1) \cdot \rho_2(\Delta y - \Delta x_1)\,d(\Delta x_1)$. Asymptotics mean that for any given accuracy, for sufficiently large values Δx_1 and Δx_2, we have $\rho_1(\Delta x_1) \approx A_1 \cdot |\Delta x_1|^{-\alpha_1}$ and $\rho_2(\Delta x_2) \approx A_2 \cdot |\Delta x_2|^{-\alpha_2}$. What is the asymptotic expression for the probability density $\rho(\Delta y)$ for large values Δy?

A large value of $\Delta y = \Delta x_1 + \Delta x_2$ can come from three different situations:

1) when Δx_1 is large (i.e., the asymptotic expression for $\rho_1(\Delta x_1)$ holds) and Δx_2 is not large in this sense;
2) when Δx_2 is large (i.e., the asymptotic expression for $\rho_2(\Delta x_2)$ holds) and Δx_1 is not large in this sense; and
3) when both Δx_1 and Δx_2 are large in this sense.

The first situation leads to terms proportional to $|\Delta x_1|^{-\alpha_1} = |\Delta y - \Delta x_2|^{-\alpha_1}$. Since in this case, Δx_2 is limited by the threshold after which the values become large, we have $\Delta x_2/\Delta y \to 0$ as $\Delta y \to \infty$ and thus, $|\Delta y - \Delta x_2|^{-\alpha_1} \sim |\Delta y|^{-\alpha_1}$. The second situation similarly leads to terms asymptotically equal to $|\Delta y|^{-\alpha_2}$.

In the third case, for some $K > 1$, the integral which describes $\rho(\Delta y)$ (over the whole real line) can be represented as a sum of the integral $I_{\text{in}}(\Delta y)$ over $[-K \cdot |\Delta y|, K \cdot |\Delta y|]$ and the integral $I_{\text{out}}(\Delta y)$ over the outside of this interval.

The inner integral $I_{\text{in}}(\Delta)$ is bounded by $M \cdot (2K \cdot |\Delta y|)$, where M is the maximum of the the product

$$\rho_1(\Delta x_1) \cdot \rho_2(\Delta y - \Delta x_1) = A_1 \cdot (\Delta x_1)^{-\alpha_1} \cdot A_2 \cdot (\Delta y - \Delta x_1)^{-\alpha_2}.$$

Differentiating this expression w.r.t. Δx_1 and equating derivative to 0, we conclude that $\Delta x_1 = \dfrac{\alpha_1}{\alpha_1 + \alpha_2} \cdot \Delta y$, hence the corresponding maximum is equal to $\text{const} \cdot |\Delta y|^{-(\alpha_1+\alpha_2)}$. Thus,

$$I_{\text{in}}(\Delta y) \le M \cdot (2K \cdot |\Delta y|) = \text{const} \cdot |\Delta y|^{-(\alpha_1+\alpha_2-1)}$$

for some positive constant.

Outside the interval, $|\Delta y| \le \dfrac{1}{K} \cdot |\Delta x_1|$, thus, $|\Delta y - \Delta x_1| \le \left(1 + \dfrac{1}{K}\right) \cdot |\Delta x_1|$ and so,

$$\rho_1(\Delta x_1) \cdot \rho_2(\Delta y - \Delta x_1) = A_1 \cdot |\Delta x_1|^{-\alpha_1} \cdot A_2 \cdot |\Delta y - \Delta x_1|^{-\alpha_2} \le$$

$$A_1 \cdot A_2 \cdot \left(1 + \frac{1}{K}\right)^{-\alpha_2} \cdot |\Delta x_1|^{-\alpha_1} \cdot |\Delta x_1|^{-\alpha_2} = \text{const} \cdot |\Delta x_1|^{-(\alpha_1+\alpha_2)}$$

for some positive constant. Integrating both sides of the resulting inequality, we conclude that

$$I_{\text{out}} \le \int_{-\infty}^{-K\cdot|\Delta y|} \text{const} \cdot |\Delta x_1|^{-(\alpha_1+\alpha_2)}\, d(\Delta x_1) +$$

$$\int_{K\cdot|\Delta y|}^{\infty} \text{const} \cdot |\Delta x_1|^{-(\alpha_1+\alpha_2)}\, d(\Delta x_1) = \text{const} \cdot |\Delta y|^{-(\alpha_1+\alpha_2-1)}$$

for some positive constant.

Both $I_{\text{in}}(\Delta y)$ and $I_{\text{out}}(\Delta y)$ are bounded by $\text{const} \cdot |\Delta y|^{-(\alpha_1+\alpha_2-1)}$, so their sum $\rho(\Delta y)$ is also bounded by a similar expression.

Summarizing: the asymptotic expression for $\rho(\Delta y)$ is the sum of three positive terms of the type $|\Delta y|^{-\alpha}$: a term corresponding to $\alpha = \alpha_1$, a term corresponding to $\alpha = \alpha_2$, and a term bounded by $\alpha = \alpha_1 + \alpha_2 - 1$. Since $\alpha_i > 1$, we have $\alpha_1 + \alpha_2 - 1 > \alpha_i$.

In general, when $\alpha < \alpha'$, then for large z, the ratio $\dfrac{z^{-\alpha'}}{z^{-\alpha}}$ tends to 0. This means in the sum of power-law asymptotic expressions, the term with the smallest value of α dominates, in the sense that the asymptotics of the sum follows the power law with the smallest possible exponent α. In our case, since $\alpha_1 + \alpha_2 - 1 > \alpha_i$, this smallest exponent is $\min(\alpha_1, \alpha_2)$.

Case of a general linear combination. One can check that a similar formula holds for the difference $y = x_1 - x_2$ and, more generally, for an arbitrary linear combination $y = a_0 + \sum\limits_{i=1}^{m} a_i \cdot x_i$. Namely, when the measurement 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 $\alpha = \min(\alpha_1, \ldots, \alpha_m)$.

Case of product $y = f(x_1, x_2) = x_1 \cdot x_2$: analysis of the problem. For the product, from $y = x_1 \cdot x_2$ and $y + \Delta y = \widetilde{y} = \widetilde{x}_1 \cdot \widetilde{x}_2 = (x_1 + \Delta x_1) \cdot (x_2 + \Delta x_2)$, we conclude that $\Delta y = \Delta x_1 \cdot x_2 + x_1 \cdot \Delta x_2 + \Delta x_1 \cdot \Delta x_2$.

We know the asymptotics of the probability distribution for Δx_1 and Δx_2, so $\Delta x_1 \cdot x_2$ and $x_1 \cdot \Delta x_2$ should have asymptotics with the same exponents α_1 and α_2. Let us find the asymptotics for the product $r \stackrel{\text{def}}{=} \Delta x_1 \cdot \Delta x_2$. Similarly to the case of addition, the corresponding terms come from three cases:

- when Δx_1 is large and Δx_2 is not large; this leads to terms $\sim |r|^{-\alpha_1}$;
- when Δx_2 is large and Δx_1 is not large; this leads to term $\sim |r|^{-\alpha_2}$;
- when both Δx_1 and Δx_2 are large; this leads to the term $\sim |r|^{-(\alpha_1+\alpha_2-1)}$, which (similarly to the case of addition) can be asymptotically ignored in comparison with terms $\sim |r|^{-\alpha_i}$.

Thus, similarly to the case of addition, we have terms with exponent α_1, we have terms with exponent α_2, and we have other terms which can be asymptotically ignored. Hence, we arrive at the following conclusion.

Case of product $y = f(x_1, x_2) = x_1 \cdot x_2$: result. When the measurement error Δx_1 of the first input has a tail with asymptotics $\rho_1(\Delta x_1) \sim A_1 \cdot |\Delta x_1|^{-\alpha_1}$ and 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 $\alpha = \min(\alpha_1, \alpha_2)$.

Case of product or ratio of several terms. One can check that a similar formula holds for the ratio $y = x_1/x_2$ and, more generally, for an arbitrary combination $y = a_0 \cdot \prod\limits_{i=1}^{m} x_i^{a_i}$. Namely, when the measurement 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 $\alpha = \min(\alpha_1, \ldots, \alpha_m)$.

Comment. 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 with first-salary distribution $\rho_1(x_1)$ and second-salary distribution $\rho_2(x_2)$,
- we know that these distributions are independent, and
- we want to find the distribution of the total salary $y = x_1 + x_2$.

3 Future Work: From Asymptotics to a Complete Description of the Corresponding Probability Distributions

Need for a complete description. In the above text, we only find the exponent α corresponding to 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 expression provides an accurate description of the probability density, and
- the probability density $\rho(\Delta y)$ on the interval $[-\Delta_0, \Delta_0]$ on which the asymptotic expression is not applicable.

Once we know a similar information for the input measurement errors Δx_1 and Δx_2, we can use the formula (3) (or similar formulas corresponding to other data processing algorithms) to estimate the corresponding characteristics for Δy.

What if we only have partial information about the distribution of errors of direct measurements. In practice, we only have partial information about the probability distributions $\rho_i(\Delta x_i)$ of the errors Δx_i of direct measurements.

Usually, we consider situations in which we know an interval on which the random variable is located with certainty. For example, for normal distribution with mean μ and standard deviation σ, we can safely conclude that all possible values are located within the six-sigma interval $[\mu - 6\sigma, \mu + 6\sigma]$, since the probability to be outside this interval is $\leq 10^{-8}$. For such distributions, uncertainty means, e.g., that instead of the exact values of the corresponding cumulative distribution functions $F(x) \stackrel{\text{def}}{=} \text{Prob}(X \leq x)$, 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; see, e.g., [2, 3].

Several algorithms are known for propagating p-boxes via data processing, i.e., for transforming the p-boxes corresponding to the input uncertainty Δx_i to the p-box for the output uncertainty Δy. It is desirable to extend these algorithms so hat they will be able to also cover a similar interval uncertainty about the values A, α, and Δ_0 describing the heavy-tailed distributions.

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.

References

1. Chakrabarti, B. K., Chakraborti, A., Chatterjee, A.: Econophysics and Sociophysics: Trends and Perspectives, Wiley-VCH, Berlin (2006)
2. Ferson, S.: RAMAS Risk Calc 4.0, CRC Press, Boca Raton, Florida (2002)
3. Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., Ginzburg, L.: Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, Sandia National Laboratories (2007) Publ. 2007-0939.
4. B. Mandelbrot, "The variation of certain speculative prices", *J. Business*, 1963, Vol. 36, pp. 394–419.
5. Mandelbrot, B.: The Fractal Geometry of Nature, Freeman, San Francisco, California (1983)
6. Mandelbrot, B., Hudson, R. L.: The (Mis)behavior of Markets: A Fractal View of Financial Turbulence, Basic Books, New York (2006)
7. Rabinovich, S.: Measurement Errors and Uncertainties: Theory and Practice, Springer Verlag, New York (2005)
8. Resnick, S. I.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling, Springer-Verlag, New York (2007)
9. Sheskin, D. J.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida (2007)