Propagation and Provenance of Probabilistic and Interval Uncertainty in Cyberinfrastructure-Related Data Processing and Data Fusion

Paulo Pinheiro da Silva[1], Aaron Velasco[2], Martine Ceberio[1], Christian Servin[1],
Matthew G. Averill[2], Nicholas Del Rio[1], Luc Longpré[1], and Vladik Kreinovich[1]
Departments of [1]*Computer Science and* [2]*Geological Sciences*
University of Texas, El Paso, TX 79968, USA, contact vladik@utep.edu

Abstract. In the past, communications were much slower than computations. As a result, researchers and practitioners collected different data into huge databases located at a single location such as NASA and US Geological Survey. At present, communications are so much faster that it is possible to keep different databases at different locations, and automatically select, transform, and collect relevant data when necessary. The corresponding cyberinfrastructure is actively used in many applications. It drastically enhances scientists' ability to discover, reuse and combine a large number of resources, e.g., data and services.

Because of this importance, it is desirable to be able to gauge the the uncertainty of the results obtained by using cyberinfrastructure. This problem is made more urgent by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – and that scientists have much less control over the quality of different resources than in the centralized database. Thus, with the cyberinfrastructure promise comes the need to analyze how data uncertainty *propagates* via this cyberinfrastructure.

When the resulting accuracy is too low, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result. In this paper, we describe algorithms for propagating uncertainty and for finding the provenance for this uncertainty.

Keywords: cyberinfrastructure, uncertainty, interval uncertainty, probabilistic uncertainty, provenance

1. Cyberinfrastructure: A Brief Overview

Practical problem: need to combine geographically separate computational resources. In different knowledge domains in science and engineering, there is a large amount of data stored in different locations, and there are many software tools for processing this data, also implemented at different locations. Users may be interested in different information about this domain.

Sometimes, the information required by the user is already stored in *one of* the *databases.* For example, if we want to know the geological structure of a certain region in Texas, we can get this information from the geological map stored in Austin. In this case, all we need to do to get an appropriate response from the query is to get this data from the corresponding database.

In other cases, different pieces of the information requested by the user are *stored at different locations*. For example, if we are interested in the geological structure of the Rio Grande Region, then we need to combine data from the geological maps of Texas, New Mexico, and the Mexican state of Chihuahua. In such situations, a correct response to the user's query requires that we access these pieces of information from different databases located at different geographic locations.

In many other situations, the appropriate answer to the user's request requires that we not only collect the relevant data $x_1, \ldots, x_n$, but that we also use some *data processing* algorithms $f(x_1, \ldots, x_n)$ to process this data. For example, if we are interested in the large-scale geological structure of a geographical region, we may also use the gravity measurements from the gravity databases. For that, we need special algorithms to transform the values of gravity at different locations into a map that describes how the density changes with location. The corresponding data processing programs often require a lot of computational resources; as a result, many such programs reside on computers located at supercomputer centers, i.e., on computers which are physically separated from the places where the data is stored.

The need to combine computational resources (data and programs) located at different geographic locations seriously complicates research.

Centralization of computational resources – traditional approach to combining computational resources; its advantages and limitations. Traditionally, a widely used way to make these computational resources more accessible was to move all these resources to a *central location*. For example, in the geosciences, the US Geological Survey (USGS) was trying to become a central repository of all relevant geophysical data. However, this centralization requires a large amount of efforts: data is presented in different formats, the existing programs use specific formats, etc. To make the central data repository efficient, it is necessary:

- to reformat all the data,
- to rewrite all the data processing programs – so that they become fully compatible with the selected formats and with each other, etc.

The amount of work that is needed for this reformatting and rewriting is so large that none of these central repositories really succeeded in becoming an easy-to-use centralized database.

Cyberinfrastructure – a more efficient approach to combining computational resources. Cyberinfrastructure technique is a new approach that provides the users with the efficient way to submit requests without worrying about the geographic locations of different computational resources – and at the same time avoid centralization with its excessive workloads. The main idea behind this approach is that *we keep all (or at least most) the computational resources*

- *at their current locations*,
- *in their current formats*.

To expedite the use of these resources:

- we supplement the local computational resources with the “metadata”, i.e., with the information about the formats, algorithms, etc.,
- we “wrap up” the programs and databases with auxiliary programs that provide data compatibility into *web services*,

and, in general, we provide a cyberinfrastructure that uses the metadata to automatically combine different computational resources.

For example, if a user is interested in using the gravity data to uncover the geological structure of the Rio Grande region, then the system should automatically:

- get the gravity data from the UTEP and USGS gravity databases,
- convert them to a single format (if necessary),
- forward this data to the program located at San Diego Supercomputer Center, and
- move the results back to the user.

This example is exactly what we have been designing under the NSF-sponsored Cyberinfrastructure for the Geosciences (GEON) project; see, e.g., (Aguiar et al., 2004; Aldouri et al., 2004; Averill et al., 2005; Ceberio et al., 2006; Ceberio et al., 2005; Keller et al., 2006; Platon et al., 2005; Schiek et al., 2007; Sinha, 2006; Torres et al., 2004; Wen et al., 2001; Xie et al., 2003), and what we are currently doing under the NSF-sponsored Cyber-Share project. This is similar to what other cyberinfrastructure projects are trying to achieve.

Technical advantages of cyberinfrastructure: a brief summary. In different knowledge domains, there is a large amount of data stored in different locations; algorithms for processing this data are also implemented at different locations. Web services – and, more generally, cyberinfrastructure – provide the users with an efficient way to submit requests without worrying about the geographic locations of different computational resources (databases and programs) – and avoid centralization with its excessive workloads (Gates at el., 2006). Web services enable the user to receive the desired data $x_1, \ldots, x_n$ and the results $y = f(x_1, \ldots, x_n)$ of processing this data.

Main advantage of cyberinfrastructure: the official NSF viewpoint. Up to now, we concentrated on the technical advantages of cyberinfrastructure. However, its advantages (real and potential) go beyond technical. According to the final report of the National Science Foundation (NSF) Blue Ribbon Advisory Panel on Cyberinfrastructure, “a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today’s challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive ‘cyberinfrastructure’ on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy.

Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities, such as understanding global climate change, protecting our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters, as well as to address some of our most fundamental intellectual questions such as the formation of the universe and the fundamental character of matter."

Main advantage of cyberinfrastructure: in short. Cyberinfrastructure greatly enhances the ability of scientists to discover, reuse and combine a large number of resources, including data and services.

2. Data Processing vs. Data Fusion

Practically important situation: it is difficult to directly measure the desired quantity with a given accuracy. In practice, we are often interested in a quantity y which is difficult (or even impossible) to directly measure with the desired accuracy.

In this situation, there are two ways to estimate the value of the desired quantity y with the desired accuracy:

- measuring *other* (related) easier-to-measure quantities and then extracting the value y from these measurements; this is called *data processing*; and
- measuring the same quantity y many times and combining the results of these measurements; this is called *data fusion*.

Important terminological comment. To avoid confusion, we would like to emphasize that sometimes, the term "data processing" refers to *all* possible processing of data by computers. In this more general sense, data fusion can be viewed as a particular case of data processing. In this paper, we limit ourselves to the narrow sense of the term "data processing".

First idea: data processing. One possible way of estimating the desired quantity y with a given accuracy is to look for easier-to-measure quantities $x_1, \ldots, x_n$ which are related to the desired y by a known dependence $y = f(x_1, \ldots, x_n)$. Based on the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring these auxiliary quantities, we can then compute an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for the quantity y.

The entire process of measurement followed by estimation is called an *indirect measurement* of y; see, e.g., (Rabinovich, 2005). The actual computation of $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is known as *data processing.*

Comment. Data processing is one of the main reasons why computers were invented in the first place, and it is still one of the major uses of computers.

Data processing: example from the geosciences. In geosciences, we want to know the structure at different depth. To determine this structure, we need to know the density y of the material at different depths. It is very difficult (and very expensive) to *directly* measure this density. Therefore, geoscientists measure this density *indirectly.*

For example, during an earthquake, geoscientists record the seismic waves at sensors located at different points on the Earth surface. As a result, we obtain the travel times $x_1, \ldots, x_n$ of the seismic signal from the earthquake location to the sensor location. Based on these travel times, we determine the structure of the Earth along the paths of the corresponding seismic waves.

The main limitations of this analysis is that earthquakes are unpredictable, they occur only at some locations and as a result, several important areas of the earth are not well covered by the corresponding paths. Thus, in addition to such *passive* (earthquake-related) seismic analysis, geoscientists also perform *active* seismic experiments, in which they start small-scale explosions in specially allocated areas and measure the travel times of the generated seismic waves. Based on these travel times, we can also determine the desired Earth structure, i.e., to be more precise, the values of the density at different depths and different locations; see, e.g., (Averill, 2007; Hole, 1992; Parker, 1994).

Specifics of data processing in cyberinfrastructure. In *traditional* data processing, when we want to know the value of a difficult-to-measure quantity y, and we know the relation between this quantity and easier-to-measure quantities x_i, we then *measure* the values x_i and use the results $\widetilde{x}_i$ of these measurements to compute the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for the desired quantity y.

As we have mentioned earlier, the main idea of a *cyberinfrastructure* is to keep all the existing measurement results readily available. Thus, with cyberinfrastructure in place, first we *look for the results* $\widetilde{x}_i$ of *measuring* x_i in the existing databases. Only when we do not find these results $\widetilde{x}_i$ – or when these results are not accurate enough – only then we actually start measuring.

Specifics of data processing in cyberinfrastructure: example from the geosciences. For example, if we want to know the geophysical structure in a certain area, instead of performing active seismic experiments we first try to combine all the known results of active seismic experiments which are related to this area. If this information is not sufficient, then we will, of course, have to perform new experiments.

Second idea: data fusion. The second idea is also very straightforward: since we cannot achieve the desired accuracy in the desired quantity y by a *single* measurement, we perform *several* independent measurements of this same quantity, and then combine ("fuse") the resulting (less accurate) values $\widetilde{y}_1, \ldots, \widetilde{y}_n$ into a single (more accurate) estimate $\widetilde{y}$ for y.

This combination can be as simple as taking an arithmetic average $\widetilde{y} = \frac{1}{n} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_n)$, or it can be more complicated: e.g., taking a weighted average or applying some non-linear combination technique. Several such techniques will be described and analyzed later in this paper.

Data fusion: examples. Data fusion is, in effect, a standard procedure that is routinely done in engineering and scientific practice (see, e.g., (Rabinovich, 2005)):

- the super-precise time is obtained by using three (or more) independent precise clocks and combining the results of these measurements;
- in medical practice, important quantities such as high blood pressure are often performed at least twice, etc.

Specifics of data processing in cyberinfrastructure. In the *traditional* engineering and scientific practice, we actually *measure* the desired quantity y several times. With *cyberinfrastructure* in place, we first *look for the existing results of measuring* the desired quantity, and try to fuse them into a single estimate.

Only if the accuracy of the resulting estimate is not good enough, then we perform additional measurements.

Combination of data processing and data fusion. In real life, to achieve the desired accuracy, it is often necessary both to use multiple measurement *and* to perform indirect measurements. In other words, in many situations, we need to combine data processing and data fusion.

For example, for many geological regions, we already have several density distributions obtained by processing different seismic data. To get a more accurate picture, it is reasonable to combine (fuse) the resulting approximate values of density, i.e., to fuse the existing data processing results.

3. Need for Uncertainty Propagation, and for Provenance of Uncertainty

Need for uncertainty propagation. As we have mentioned, one of the main reasons why we need data processing (i.e., indirect measurements) and data fusion (i.e., multiple measurements) in the first place is that the accuracy of the original direct measurement is not high enough. It is therefore important to make sure that after the proposed data processing and/or data fusion, we get the desired accuracy. In other words, we must find out how the uncertainty (inaccuracy) of the direct measurement results *propagates* via the infrastructure.

The need for uncertainty propagation is enhanced by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – as well as the level of uncertainty of any response derived from such resources. Also, in contrast to the centralized platform, in cyberinfrastructure, scientists have less control about the quality of different resources. Thus, the cyberinfrastructure promise comes along with the need to support the associated uncertainty analysis uncertainty propagation.

Need for the provenance of uncertainty. When the resulting accuracy is sufficient, we get the desired estimate $\widetilde{y}$. However, sometimes, the resulting accuracy is still too low. In this situation, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result.

Comment. In this paper, we mainly deal with the provenance of *uncertainty.* It is worth mentioning that in general, other aspects of provenance are also very important: e.g., to be able to adequately gauge the *reliability* of different measurement results (and thus, to form a decision on how much we trust these results), we must take into account the provenance of these results – i.e., which team performed these measurements, what auxiliary data was used in pre-processing these measurement results, etc.

4. Uncertainty of the Results of Direct Measurements: Probabilistic and Interval Approaches

Measurement uncertainty: general description. To find out how the inaccuracies $\Delta x_i = \widetilde{x}_i - x_i$ of direct measurements (= measurement errors) propagate through the cyberinfrastructure, we need to recall how these inaccuracies Δx_i are usually described.

The manufacturer of the measuring instrument must supply us with an upper bound Δ_i on the measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding "measuring instrument" is practically useless. In this case, once we performed a measurement and got a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

Probabilistic uncertainty. In many practical situations, we not only know the interval $[-\Delta_i, \Delta_i]$ of possible values of the measurement error; we also know the probability of different values Δx_i within this interval. This knowledge underlies the traditional engineering approach to estimating the error of indirect measurement, in which we assume that we know the probability distributions for measurement errors Δx_i.

These probabilities are often described by a normal distribution, so in standard engineering textbook on measurement, it is usually assumed that the distribution of Δx_i is normal, with 0 average and known standard deviation σ_i; see, e.g. (Fuller, 1987; Rabinovich, 2005).

In general, we can determine the desired probabilities of different values of Δx_i by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one use, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error.

Interval uncertainty. There are two cases, however, when in practice, we do not determine the probabilities:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.
- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after we performed a measurement and got a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

What we consider in this paper. For each of the 2 techniques for improving accuracy (data processing and data fusion), we must therefore consider 2 possible situations:

- when we know the probabilities of inaccuracies of direct measurements, and
- when we only know upper bounds (intervals) for these inaccuracies.

So, we have $2 \times 2 = 4$ possible situations. In this paper, we will consider all four situations. We start with data processing under probabilistic and interval uncertainty, and then we cover data fusion under both types of uncertainty.

For three of these four situations, the answer is reasonably straightforward; for the fourth one, we will come up with new formulas.

5. Typical Situation: Measurement Errors are Reasonably Small

Before we start analyzing different situations, let us mention that in this paper, we will only consider a typical situation in which the direct measurements are accurate enough, so that the resulting approximation errors Δx_i are small, and terms which are quadratic (or of higher order) in Δx_i can be safely neglected. In such situations, for data processing, the dependence of the desired value $y = f(x_1, \ldots, x_n) = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n)$ on Δx_i can be safely assumed to be linear.

When approximation errors are small, we can simplify the expression for $\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n)$, if we expand the function f in Taylor series around the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and restrict ourselves only to linear terms in this expansion. As a result, we get the expression

$$\Delta y = c_1 \cdot \Delta x_1 + \ldots + c_n \cdot \Delta x_n,$$

where by c_i we denoted the value of the partial derivative $\dfrac{\partial f}{\partial x_i}$ at the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

In the linear approximation, for small $h > 0$, we have $f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) \approx f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) + c_i \cdot h$, hence we can determine c_i as

$$c_i = \frac{1}{h} \cdot (f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}).$$

Comment. There are practical situations when the accuracy of the direct measurements is not high enough, and hence, quadratic terms cannot be safely neglected (see, e.g., (Jaulin, 2001) and references therein). In this case, the problem of error estimation for indirect measurements becomes computationally difficult (NP-hard) even when the function $f(x_1, \ldots, x_n)$ is quadratic (Kreinovich et al., 1998; Vavasis, 1991). However, in most real-life situations, the possibility to ignore quadratic terms is a reasonable assumption, because, e.g., for an error of 1% its square is a negligible 0.01%.

6. Case of Data Processing

Propagation of uncertainty through data processing: case of probabilistic uncertainty. In the statistical setting, the desired measurement error Δy is a linear combination of independent

Gaussian variables Δx_i. Therefore, Δy is also normally distributed, with 0 average and the standard deviation

$$\sigma = \sqrt{c_1^2 \cdot \sigma_1^2 + \ldots + c_n^2 \cdot \sigma_n^2}.$$

Comment. A similar formula holds if we *do not* assume that Δx_i are normally distributed: it is sufficient to assume that they are independent variables with 0 average and known standard deviations σ_i.

Uncertainty provenance in data processing: case of probabilistic uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, we know which component of the resulting variance σ^2 comes from the inaccuracy σ_i of the i-th measurement. We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\sigma_i^{\text{new}} \ll \sigma_i$) will affect the resulting variance σ^2.

Propagation of uncertainty through data processing: case of interval uncertainty. In the interval setting, we do not know the probability of different errors Δx_i; instead, we only know that $|\Delta x_i| \le \Delta_i$. In this case, the sum $\sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$ attains its largest possible value if each term $c_i \cdot \Delta x_i$ in this sum attains the largest possible value:

- If $c_i \ge 0$, then this term is a monotonically non-decreasing function of Δx_i, so it attains its largest value at the largest possible value $\Delta x_i = \Delta_i$; the corresponding largest value of this term is $c_i \cdot \Delta_i$.
- If $c_i < 0$, then this term is a decreasing function of Δx_i, so it attains its largest value at the smallest possible value $\Delta x_i = -\Delta_i$; the corresponding largest value of this term is $-c_i \cdot \Delta_i = |c_i| \cdot \Delta_i$.

In both cases, the largest possible value of this term is $|c_i| \cdot \Delta_i$, so, the largest possible value of the sum Δy is

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

Similarly, the smallest possible value of Δy is $-\Delta$.

Hence, the interval of possible values of Δy is $[-\Delta, \Delta]$, and the interval of possible values of the actual value y is $[\widetilde{y} - \Delta, \widetilde{y} + \Delta]$.

Uncertainty provenance in data processing: case of interval uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$, we know which component of the resulting approximation error Δ comes from the inaccuracy Δ_i of the i-th measurement. We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\Delta_i^{\text{new}} \ll \Delta_i$) will affect the resulting approximation error Δ.

7. Case of Data Fusion

Propagation of uncertainty through data fusion: case of probabilistic uncertainty. In data fusion, we know several results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y. Under probabilistic uncertainty, we assume that the corresponding n measurements errors are independent normally distributed random variables with 0 mean and known standard deviations σ_i. In this case, for each possible value y, the probability density ρ_i of getting $\widetilde{y}_i$ is equal to

$$\rho_i(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_i} \cdot \exp\left(-\frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right),$$

and thus, the probability density $\rho(y)$ of having the given n measurements is equal to

$$\rho(y) = \rho_1(y) \cdot \ldots \cdot \rho_n(y) = \text{const} \cdot \exp\left(-\sum_{i=1}^{n} \frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right).$$

As a resulting estimate $\widetilde{y}$ for the desired (unknown) quantity y, it is then reasonable to select the most probable value, i.e., the value for which the probability density $\rho(y)$ is the largest.

Maximizing $\rho(y)$ is equivalent to minimizing the quadratic function $-\ln(\rho(y))$; differentiating this quadratic expression with respect to y and equating the derivative to 0, we conclude that

$$\widetilde{y} = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}} \cdot \sum_{i=1}^{n} \frac{\widetilde{y}_i}{\sigma_i^2}.$$

This estimate is a linear combination of normally distributed estimates $\widetilde{y}_i$ with mean y and standard deviation σ_i, with coefficients $c_i = \sigma_i^{-1} / \left(\sum\limits_{j=1}^{n} \sigma_j^{-2}\right)$. Thus, $\widetilde{y}$ is also normally distributed, with the same mean y and the standard deviation $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, i.e., with standard deviation

$$\sigma^2 = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}}.$$

This formula can also be rewritten as

$$\frac{1}{\sigma^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}.$$

Uncertainty provenance in data fusion: case of probabilistic uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\sigma^{-2} = \sum\limits_{i=1}^{n} \sigma_i^{-2}$, we know which component of the resulting variance σ^2 comes from the inaccuracy σ_i of the i-th measurement.

We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\sigma_i^{\mathrm{new}} \ll \sigma_i$) will affect the resulting variance σ^2. Good news is we can predict this accuracy beforehand, without actually performing the measurements – since the resulting accuracy σ depends only on the accuracies σ_i of individual measurements and not on the results of these measurements.

Case of unknown probabilistic uncertainty. Sometimes, we do not know the accuracy of the fused measurements. In this case, we can use the differences between the measurement results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ to estimate the standard deviation $\sigma_1 = \ldots = \sigma_n$ of the corresponding measurements by using the usual statistical formula $\sigma_1^2 = \dfrac{1}{n-1} \cdot \sum\limits_{i=1}^{n} (\Delta y_i - E)^2$, where $E \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} \Delta y_i$.

Propagation of uncertainty through data fusion: case of interval uncertainty. Under interval uncertainty, we know n results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y, and we know the accuracy Δ_i of each measurement. Thus, for each i, we know that the actual (unknown) value y of the desired quantity must belong to the interval $\mathbf{y}_i \stackrel{\text{def}}{=} [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i]$.

Fusion here is straightforward: the set of all the values y which belong to all n intervals is equal to the *intersection* $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of these intervals. Here, $\underline{y} = \max(\widetilde{y}_1 - \Delta y_1, \ldots, \widetilde{y}_n - \Delta y_n)$, $\overline{y} = \min(\widetilde{y}_1 + \Delta y_1, \ldots, \widetilde{y}_n + \Delta y_n)$, and the accuracy $\Delta = \dfrac{\overline{y} - \underline{y}}{2}$ of the fused estimate can be computed as

$$\Delta = \frac{1}{2} \cdot (\min(\widetilde{y}_1 + \Delta y_1, \ldots, \widetilde{y}_n + \Delta y_n) - \max(\widetilde{y}_1 - \Delta y_1, \ldots, \widetilde{y}_n - \Delta y_n)).$$

Case of unknown interval uncertainty: a reasonable approach. Sometimes, we do not know the accuracy Δ_i of the fused measurements. In this case, it is reasonable to get a single estimate for $\Delta_1 = \ldots = \Delta_n$ for all these measurements. We know that the intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_1 + \Delta_1]$ must intersect – since they all contain the actual (unknown) value of the desired quantity y.

For the intersection to be non-empty, every lower bound $\widetilde{y}_i - \Delta_1$ must be smaller than or equal to every upper bound $\Delta y_j + \Delta_1$. Thus, we must have $\widetilde{y}_i - \widetilde{y}_j \le 2\Delta_1$. So, we can conclude that $\Delta_1 \ge \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i - \min\limits_i \Delta y_i)$.

Case of unknown interval uncertainty: seemingly reasonable proposal and its limitations. The actual value Δ_1 can be larger than this half-difference, but as a first approximation, it may be reasonable to take $\Delta_1 \approx \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i - \min\limits_i \Delta y_i)$. This particular choice may not be the most adequate, since in this case, the intersection of the corresponding intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_1 + \Delta_1]$ consists of a single point – the midpoint $y_{\mathrm{mid}} \stackrel{\text{def}}{=} \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i + \min\limits_i \Delta y_i)$. This conclusion is somewhat misleading because it erroneously suggests that we know the exact value of the estimated quantity.

In the following text, we will return to this problem and show how to get a somewhat more adequate estimate.

Uncertainty provenance in data fusion: case of interval uncertainty. The above formula describes the *propagation* of uncertainty. From the viewpoint of uncertainty propagation, this formula is even simpler than in the probabilistic case. So, from the computational viewpoint, we can say that as far as propagation of uncertainty is concerned, the situation with interval uncertainty is easier-to-handle than the situation with probabilistic uncertainty.

With provenance, however, the situation is exactly opposite. For probabilistic uncertainty, we can predict the resulting accuracy beforehand, without actually performing the measurements – since the resulting accuracy σ depends only on the accuracies σ_i of individual measurements and not on the results of these measurements. In contrast, for the interval uncertainty, for the same accuracies $\Delta_1, \ldots, \Delta_n$ of the individual measurements, we can get different accuracy Δ of the fusion result – depending on the actual measurement results.

Let us illustrate this problem on the simplest example, when the actual (unknown) value is $y = 0$, and we fuse two measurements with the exact same accuracy $\Delta_1 = \Delta_2 = 1$. All we know about the results of these two measurements is that the resulting intervals contain the actual value y. Since this is the only restriction, we can two radically different extreme situations:

- It is possible that in both measurements, we get the same interval $[-1, 1]$ containing 0. In this case, the intersection is exactly the same interval, so the resulting accuracy is $\Delta = 1$, the same accuracy with which we started.

- It is also possible that in the first measurement, we get the interval $[-1, 0]$ and in the second measurement, we get the interval $[0, 1]$. In this case, as a result of data fusion, we get the exact value of the measured quantity, with $\Delta = 0$.

We can also have all possible values in between. In general, if we have n measurements with accuracies $\Delta_1, \ldots, \Delta_n$, then the half-width Δ of the intersection of the corresponding intervals can take any values from 0 to $\min(\Delta_1, \ldots, \Delta_n)$.

Planning data fusion under interval uncertainty: formulation of the problem. If the accuracy of the result of data fusion is not sufficient, we should then supplement the existing measurements with one or several more accurate ones. How accurate should these new measurement be? how many of these more accurate measurements should we make? It is desirable to have some answers to these questions before we go into the time- and resources-consuming process of actually buying the corresponding sensors and performing the measurements – because if we do not get the desired accuracy again, this time-consuming process will be mostly wasting time.

In other words, it is desirable to produce an estimate for the accuracy Δ of the result of fusing measurements with accuracies $\Delta_1, \ldots, \Delta_n$, an estimate that we can obtain before we start the actual measurements. How can we solve this problem?

Planning data fusion under interval uncertainty: main idea. Our main idea of solving the above problem is as follows. We know that the i-th measurement has accuracy Δ_i. This means that the only information that we have about the possible values of the i-th measurement error Δy_i is that this error belongs to the interval $[-\Delta_i, \Delta_i]$.

We have no information about the probabilities of different values of Δy_i within this interval. According to Laplace's principle of indifference, in this situation, it is reasonable to assume that all possible values have the same probability, i.e., that the distribution of Δy_i on the interval $[-\Delta_i, \Delta_i]$ is uniform.

For each combinations of choices of Δy_i, we get different measurement results $\widetilde{y}_i = y + \Delta y_i$ and thus, different intersections $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of the corresponding intervals $\mathbf{y}_i = [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i] = [y + \Delta y_i - \Delta_i, y + \Delta y_i + \Delta_i]$.

We are interested in the accuracy of the fused results. This accuracy can be gauged by the largest possible absolute value Δ of the different between the actual value y and values from the fused interval $\mathbf{y}$.

As we have mentioned, in principle, this accuracy Δ can be as small as large as $\min(\Delta_1, \ldots, \Delta_n)$. However, the probability of such a large inaccuracy Δ is reasonably small; in our estimates of Δ, we would like to ignore small-probability events. In other words, we would like to select an allowable small probability p_0 of mis-estimation, and estimate Δ as the smallest value for which the probability to have $\overline{y} \le y + \Delta$ is at least $1 - p_0$ and the probability to have $\underline{y} \ge y - \Delta$ is also $\ge 1 - p_0$.

Thus, we arrive at the following precise problem.

Planning data fusion under interval uncertainty: precise formulation of the problem. Let $p_0 > 0$ be a fixed real number. We start with an arbitrary value y. Let $\Delta y_1, \ldots, \Delta y_n$ be n independent random variables such that each variable Δy_i is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$. By Δ, we mean that smallest value for which the probability that for the intersection $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of the intervals $\mathbf{y}_i = [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i]$, where $\widetilde{y}_i = y + \Delta y_i$, the following two properties hold:

- the probability to have $\overline{y} \le y + \Delta$ is at least $1 - p_0$, and
- the probability to have $\underline{y} \ge y - \Delta$ is also $\ge 1 - p_0$.

Towards an estimate for Δ. The condition that $\overline{y} \le y + \Delta$ means that

$$\min(y + \Delta y_1 + \Delta_1, \ldots, y + \Delta y_n + \Delta_n) \le y + \Delta.$$

Which number is smaller and which is larger does not change when we shift all these numbers by the same shift y. Thus, $\min(y + \Delta y_1 + \Delta_1, \ldots, y + \Delta y_n + \Delta_n) = y + \min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n)$ and hence, the above inequality takes the form

$$\min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n) \le \Delta.$$

The probability p_{opp} for the opposite inequality

$$\min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n) > \Delta$$

should be $\le p_0$.

The minimum of several sums is $> \Delta$ if and only if each of these sums is $> \Delta$. Thus, the above opposite inequality holds if all n inequalities $\Delta y_i + \Delta_i > \Delta$ hold. Since the variables Δy_i are independent, we thus conclude that $p_{\text{opp}} = p_1 \cdot \ldots \cdot p_n$, where $p_i \overset{\text{def}}{=} \text{Prob}(\Delta y_i + \Delta_i > \Delta) = \text{Prob}(\Delta y_i > \Delta - \Delta_i)$. Since Δy_i is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$, the probability p_i is equal to the ratio of

- the size of the set $(\Delta - \Delta_i, \Delta_i]$ where the corresponding inequality holds to

– the size of the overall set $[-\Delta_i, \Delta_i]$ on which the distribution is defined,

i.e., to $p_i = \dfrac{\Delta_i - (\Delta - \Delta_i)}{2\Delta_i} = \dfrac{2\Delta_i - \Delta}{2\Delta_i} = 1 - \dfrac{\Delta}{2\Delta_i}$. Thus,

$$p_{\rm opp} = \prod_{i=1}^n \left(1 - \frac{\Delta}{2\Delta_i}\right).$$

This product decreases with Δ; thus, the smallest possible value Δ for which $p_{\rm opp} \le p_0$ can be determined from the condition $p_{\rm opp} = p_0$, i.e., $\prod\limits_{i=1}^n \left(1 - \dfrac{\Delta}{2\Delta_i}\right) = p_0$.

Taking logarithms of both sides, we get

$$\sum_{i=1}^n \ln\left(1 - \frac{\Delta}{2\Delta_i}\right) = \ln(p_0).$$

We are interested in the case when data fusion is efficient, i.e., when $\Delta \ll \Delta_i$. In this case, $\dfrac{\Delta}{2\Delta_i} \ll 1$, and we can use an approximate linearized formula $\ln(1 - x) \approx -x$ which is true for small x. This formula leads to $\sum\limits_{i=1}^n \dfrac{\Delta}{\Delta_i} = 2|\ln(p_0)|$, i.e., to $\Delta \cdot \left(\sum\limits_{i=1}^n \dfrac{1}{\Delta_i}\right) = 2|\ln(p_0)|$ and

$$\Delta = \frac{\rm const}{\sum\limits_{i=1}^n \dfrac{1}{\Delta_i}},$$

or, equivalently, $\dfrac{1}{\Delta} = {\rm const} \cdot \sum\limits_{i=1}^n \dfrac{1}{\Delta_i}$.

The second inequality leads to the exact same formula for Δ.

Data fusion under interval uncertainty: result. When we fuse n measurement results with accuracies Δ_i, the accuracy Δ of the fused estimate can be estimated based on the formula

$$\frac{1}{\Delta} = {\rm const} \cdot \sum_{i=1}^n \frac{1}{\Delta_i},$$

in which the constant ${\rm const} = 2|\ln(p_0)|$ depends on the allowed probability p_0 that the actual inaccuracy of the fused value is higher than this estimate.

Case of unknown interval uncertainty: revisited. Let us recall that in the case of data fusion under unknown interval uncertainty, a reasonable choice for the accuracy $\Delta_1 = \ldots = \Delta_n$ of the fused measurements is the smallest value Δ_1 for which the corresponding intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_i + \Delta_1]$ intersect. The problem with this approach is that for this smallest value, the intersection consists of a single point $y_{\rm mid}$ – making it sound as if we knew the exact value of the estimated quantity y.

To avoid this erroneous impression, a reasonable idea is to estimate the accuracy Δ of the fused result – for $\Delta_1 = \ldots = \Delta_n$ we get $\Delta = \Delta_1/n$ – and "add" this accuracy Δ to this point, i.e., return the interval $[y_{\rm mid} - \Delta, y_{\rm mid} + \Delta]$ as the interval estimate for the desired quantity y.

Comparison between data fusion under probabilistic and interval uncertainty. The above formula for Δ is similar to the formula $\dfrac{1}{\sigma^2} = \text{const} \cdot \sum\limits_{i=1}^{n} \dfrac{1}{\sigma_i^2}$ which describes the accuracy σ of data fusion under probabilistic uncertainty. The main difference is that instead of the variances σ_i^2 and σ^2 we now have upper bounds Δ_i and Δ.

In practical terms, this formal difference can be described as follows.

- If we apply data fusion to n results known with the same probabilistic uncertainty $\sigma_1 = \ldots = \sigma_n$, then we result of data fusion is known with the uncertainty $\sigma = \dfrac{\sigma_i}{\sqrt{n}}$.
- On the other hand, if we we apply data fusion to n results known with the same interval uncertainty $\Delta_1 = \ldots = \Delta_n$, then we result of data fusion is known with the uncertainty $\Delta = \dfrac{\sigma_i}{n}$.

Thus, with interval uncertainty, we get a much faster ($\sim 1/n$) decrease in approximation error than for the probabilistic uncertainty ($\sim 1/\sqrt{n}$). This fact is in line with similar estimates from (Walster, 1988; Walster and Kreinovich, 1996).

Comment. It is worth mentioning that there is a similar difference for data processing: in the interval case, we have $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$, whereas in the probabilistic case, we have $\sigma^2 = \sum\limits_{i=1}^{n} |c_i|^2 \cdot \sigma_i^2$.

The difference between the formulas for data fusion and data processing is similar to the formulas for the resistance R of of an electric circuit consisting of resistances $R_1, \ldots, R_n$: when the resistances are placed sequentially, we get $R = R_1 + \ldots + R_n$ (as for data processing); when the resistance are placed in parallel to each other, we get $\dfrac{1}{R} = \dfrac{1}{R_1} + \ldots + \dfrac{1}{R_n}$ (as for data fusion).

8. Propagation of Uncertainty When We Have Both Data Processing and Data Fusion

Motivations. As we have mentioned earlier, in many real-life situations, to get the desired accuracy, we must apply both data fusion and data processing.

Example. For example, we can fuse several values $y_1, \ldots, y_n$ each of which is obtained by data processing.

Main idea. In this case, to find the accuracy of the final result, we propagate the uncertainty through all these data fusion/data processing steps.

Example. In the above example,

- we first use the formulas for propagating uncertainty under data processing to come up with accuracy values (Δ_i or σ_i) for y_i, and then
- we use the formulas for uncertainty propagation under data fusion to combine these values Δ_i (correspondingly, σ_i) into a single estimate Δ (corr., σ).

9. Towards Optimal Data Processing and Data Fusion

Motivations. Up to now, we concentrated on the *analysis* of given data processing and data fusion scenarios. However, since we have explicit (and simple) formulas for the propagation of uncertainty under data processing and data fusion, we can actually solve the problem of finding the least expensive way to guarantee the given accuracy.

To perform this *optimization*, we must know how the cost of measuring related quantities with different accuracies.

Towards optimal data fusion: preliminary description. For data fusion, let $c^{\rm prob}(\sigma)$ denote the cost of measuring the desired quantity with standard deviation σ, and let $c^{\rm int}(\Delta)$ denote the cost of measuring the desired quantity with the guaranteed upper bound Δ on the measurement error. Typically, $c^{\rm prob}(\sigma) = \dfrac{C}{\sigma^\alpha}$ and $c^{\rm int}(\Delta) = \dfrac{C}{\Delta^\alpha}$ for some constants C and $\alpha > 0$; see, e.g., (Nguyen et al., 2008; Nguyen and Kreinovich, 2008) and references therein.

Towards optimal data fusion: probabilistic case. In the probabilistic case, we must find the values σ_i for which $\sum\limits_{i=1}^n c^{\rm prob}(\sigma_i) \to \min$ under the constraint that the sum $\sum\limits_{i=1}^n \sigma_i^{-2}$ is equal to the given value σ^{-2}. By applying Lagrange multiplier method to this constraint optimization problem, we get an unconstraint optimization problem $\sum\limits_{i=1}^n c^{\rm prob}(\sigma_i) + \lambda \cdot \sum\limits_{i=1}^n \sigma_i^{-2} \to \min$. Differentiating w.r.t. σ_i and equating the derivative to 0, we conclude that $c'(\sigma_i) \cdot \sigma_i^3 = {\rm const} = 2\lambda$.

For a function $c^{\rm prob}(\sigma) = \dfrac{C}{\sigma^\alpha}$, the expression $c'(\sigma_i) \cdot \sigma_i^3$ is monotonic in σ_i and thus, the equality occurs only for one value σ_i – hence in the optimal plan, $\sigma_1 = \ldots = \sigma_n$. To get the desired value σ, we must have $\sigma_i = \sqrt{n} \cdot \sigma$.

Towards optimal data fusion: interval case. Similarly, in the interval case, the problem of minimizing $\sum\limits_{i=1}^n c^{\rm int}(\Delta_i)$ under the constraint $\sum\limits_{i=1}^n \Delta_i^{-1} = \Delta^{-1}$ leads to the equation $c'(\Delta_i) \cdot \Delta_i^2 = {\rm const} = \lambda$.

For a function $c^{\rm int}(\Delta) = \dfrac{C}{\Delta^\alpha}$, the expression $c'(\Delta_i) \cdot \Delta_i^2$ is monotonic in Δ_i and thus, the equality occurs only for one value Δ_i – hence in the optimal plan, $\Delta_1 = \ldots = \Delta_n$. To get the desired value Δ, we must have $\Delta_i = n \cdot \Delta$.

Towards optimal data processing: preliminary description. For data processing, let $c_i^{\rm prob}(\sigma_i)$ denote the cost of measuring the i-th quantity with standard deviation σ_i, and let $c_i^{\rm int}(\Delta_i)$ denote the cost of measuring the i-th quantity with the guaranteed upper bound Δ_i on the measurement error. Just like in the case data fusion, typically, we have $c_i^{\rm prob}(\sigma_i) = \dfrac{C_i}{\sigma_i^{\alpha_i}}$ and $c_i^{\rm int}(\Delta_i) = \dfrac{C_i}{\Delta^{\alpha_i}}$ for some constants C_i and α_i.

Towards optimal data processing: probabilistic case. In the probabilistic case, the problem of minimizing $\sum\limits_{i=1}^n c_i^{\rm prob}(\sigma_i)$ under the constraint $\sum\limits_{i=1}^n c_i^2 \cdot \sigma_i^2 = \sigma^2$ leads to the equation $\dfrac{c'(\sigma_i)}{c_i^2 \cdot \sigma_i} = {\rm const} = -2\lambda$.

For $c_i^{\rm prob}(\sigma_i)=\dfrac{C_i}{\sigma_i^{\alpha_i}}$, we get $\sigma_i=\left(\dfrac{\alpha_i\cdot C_i}{2\lambda\cdot c_i^2}\right)^{1/(2+\alpha_i)}$, where λ can be determined from the equation

$$\sum_{i=1}^n c_i^2\cdot\left(\frac{\alpha_i\cdot C_i}{2\lambda\cdot c_i^2}\right)^{2/(2+\alpha_i)}=\sigma^2.$$

Towards optimal data processing: interval case. In the interval case, the problem of minimizing $\sum\limits_{i=1}^n c_i^{\rm int}(\Delta_i)$ under the constraint $\sum\limits_{i=1}^n |c_i|\cdot\Delta_i=\Delta$ leads to the equation $\dfrac{c'(\Delta_i)}{|c_i|}={\rm const}=-\lambda$.

For $c_i^{\rm int}(\Delta_i)=\dfrac{C_i}{\Delta_i^{\alpha_i}}$, we get $\Delta_i=\left(\dfrac{\alpha_i\cdot C_i}{\lambda\cdot|c_i|}\right)^{1/(1+\alpha_i)}$, where λ can be determined from the equation

$$\sum_{i=1}^n |c_i|\cdot\left(\frac{\alpha_i\cdot C_i}{\lambda\cdot|c_i|}\right)^{2/(2+\alpha_i)}=\Delta.$$

10. Combining Probabilistic and Interval Uncertainty

Motivations. In the previous sections, we assumed that in data processing an in data fusion, either all measurement results are known with probabilistic uncertainty, or all measurement results are known with interval uncertainty.

In practice, some measurement results are known with probabilistic uncertainty (i.e., we know the probabilities of the corresponding measurement errors), and some are only known with interval uncertainty (i.e., we only know the upper bounds on the corresponding measurement errors). In this case, how can we estimate the accuracy of the result of data processing or data fusion?

Case of data processing. For data processing, it is possible to provide an answer to the above question. Indeed, suppose that we produce an estimate $\widetilde{y}=f(\widetilde{x}_1,\ldots,\widetilde{x}_n)$ for the desired quantity y which is based on the results $\widetilde{x}_i$ of directly measuring n related quantities $x_1,\ldots,x_n$.

We are interested in situations in which some of the measurement errors are known with probabilistic uncertainty, and some with interval uncertainty. Without losing generality, we can assume that the values $x_1,\ldots,x_k$ are known with probabilistic uncertainty and the values $x_{k+1},\ldots,x_n$ are known with interval uncertainty. In other words, we know the standard deviations $\sigma_1,\ldots,\sigma_k$ of the first k measurements, and we know the upper bounds $\Delta_{k+1},\ldots,\Delta_n$ of the others. In this case, the above linearized formula $\Delta y=\sum\limits_{i=1}^n c_i\cdot\Delta x_i$ can be rewritten as $\Delta y=\Delta y^{\rm prob}+\Delta y^{\rm int}$, where $\Delta y^{\rm prob}=\sum\limits_{i=1}^k c_i\cdot\Delta x_i$ and $\Delta y^{\rm int}=\sum\limits_{i=k+1}^n c_i\cdot\Delta x_i$.

As we already know, $\Delta y^{\rm prob}$ is a normally distributed random variable with 0 mean and standard deviation $\sigma = \sqrt{\sum\limits_{i=1}^{k} c_i^2 \cdot \sigma_i^2}$ and $\Delta y^{\rm int}$ is a variable about which we only know that it belongs to the interval $[-\Delta, \Delta]$, where $\Delta = \sum\limits_{i=k+1}^{n} |c_i| \cdot \Delta_i$.

So, we conclude that the approximation error Δy is the sum of two error components: a random one with a known σ and an interval one with a known Δ.

The resulting two-component description of measurement and approximation error is in line with the measurement practice. The above two-component description of an approximation error is in line with the standard practice in measurement theory (see, e.g., (Rabinovich, 2005)), where a measurement error Δx is often described by its two component:

- a *random* error component $\Delta_r x \stackrel{\rm def}{=} \Delta x - E[\Delta x]$ with 0 mean ($E[\Delta_r x] = 0$), for which we usually know the standard deviation σ, and
- a *systematic* error component $\Delta_s x \stackrel{\rm def}{=} E[\Delta x]$ for which we only know the upper bound Δ on its absolute value.

The situation in which we only know the upper bound Δ on the (absolute value of) the total measurement error can be viewed as a degenerate case of this two-component description, with $\sigma = 0$.

Data processing: case when we have a two-component description of all the measurement errors. In view of the prevalence of the two-component error description in measurement practice, it is reasonable to consider the following situation.

We want to estimate the value of the desired difficult-to-measure quantity y. We know the relation $y = f(x_1, \ldots, x_n)$ between this quantity y and easier-to-measure quantities $x_1, \ldots, x_n$. For each of these auxiliary quantities x_i, we know the measurement result $\widetilde{x}_i$ and we know that the corresponding measurement error $\Delta x_i \stackrel{\rm def}{=} \widetilde{x}_i - x_i$ can be represented as a sum of two components $\Delta x_i = \Delta_s x_i + \Delta_r x_i$, where:

- the component $\Delta_s x_i$ is a random variable with 0 mean and known standard deviation σ_i;
- about the component $\Delta_r x_i$, we only know the upper bound Δ_i on the (absolute value of the) measurement error, i.e., we know that $|\Delta_r x_i| \le \Delta_i$.

Based on the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$, we compute an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for y. What can we say about the approximation error $\Delta y \stackrel{\rm def}{=} \widetilde{y} - y$ of this estimate?

In the linearization case, we can conclude that $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$. Since $\Delta x_i = \Delta_s x_i + \Delta_r x_i$, we can conclude that $\Delta y = \Delta_s y + \Delta_r y$, where $\Delta_r y = \sum\limits_{i=1}^{n} c_i \cdot \Delta_r x_i$ and $\Delta_s y = \sum\limits_{i=1}^{n} c_i \cdot \Delta_s x_i$. We already know how to handle each of these two sums, so we conclude that the approximation error Δy also consists of two components:

- the component $\Delta_s y$ is a random variable with 0 mean and known standard deviation

$$\sigma = \sqrt{\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2};$$

- about the component $\Delta_r y$, we only know the upper bound $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$ on the (absolute value of the) measurement error, i.e., we know that $|\Delta_r y| \le \Delta$.

Case of data fusion. In data fusion, we have n results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y. In the previous sections, we assume that either all of these measurement errors are known with probabilistic uncertainty, or that all of them are know with interval uncertainty.

Let us now consider the case when some of the measurement errors are known with probabilistic uncertainty, and some with interval uncertainty. Without losing generality, we can assume that the values $y_1, \ldots, y_k$ are known with probabilistic uncertainty and the values $y_{k+1}, \ldots, y_n$ are known with interval uncertainty. In other words, we know the standard deviations $\sigma_1, \ldots, \sigma_k$ of the first k measurements, and we know the upper bounds $\Delta_{k+1}, \ldots, \Delta_n$ of the others. In this case,

- we can use the data fusion formula for the probabilistic uncertainty to fuse the measurement $\widetilde{y}_1, \ldots, \widetilde{y}_k$ into a single result $\widetilde{y}$ with a standard deviation σ, and
- we can fuse the interval-valued measurements by taking the intersection

$$[\underline{y}, \overline{y}] \stackrel{\text{def}}{=} [\widetilde{y}_{k+1} - \Delta_{k+1}, \widetilde{y}_{k+1} + \Delta_{k+1}] \cap \ldots \cap [\widetilde{y}_n - \Delta_n, \widetilde{y}_n + \Delta_n]$$

of the corresponding intervals.

It is therefore important to fuse the interval estimate with accuracy Δ and the probabilistic estimate with the accuracy σ The result of the fusion depends on the relation between Δ and σ:

- If $\Delta \gg \sigma$, this means that the interval estimate is much much wider than what we can simply conclude based on the probabilistic information. Thus, in this case, the fused information consists simply of the probabilistic estimate.
- If $\sigma \gg \Delta$, this means that the probabilistic estimate is much worse than the interval one. In this case, the fused information consists simply of the interval estimate.
- If the estimates σ and Δ are approximately of the same order, this means that we can keep either one of them.

Comment. Our recommendation for the case when the estimates σ and Δ are approximately of the same order is somewhat vague. For this case, it would be nice to come up with a better answer to the fusion question.

11. Adding Reliability and Trust: Results and Open Problems

Formulation of the problem. In the previous sections, we concentrated on the measurement uncertainty, i.e., on the difference between the measurement result and the actual value of the corresponding quantity. We also assumed that these differences are relatively small.

This smallness assumption holds in many practical situations. However, sometimes, we have values which are completely off: a measuring instrument can malfunction, a computer may have misread this information, etc. Such values are often called *outliers.* In short, in addition to being not 100% accurate, the measurement results are also not 100% *reliable.* How can we take this possible unreliability into account in data processing and data fusion?

How we can describe the reliability of different measurement results. A natural way to describe the reliability of different measurement results is to provide the probability p_i that the i-th measurement result is an outlier. It is usually assumed that in terms of reliability, different measurement results are independent – so that, e.g., the probability that both the i-th and the j-th results are outliers is equal to the product $p_i \cdot p_j$.

These probabilities can be gauged, e.g., based on our knowledge of what team performed these measurements, what is the track records of this particular team, what auxiliary values have been used in pre-processing these results, etc. In other words, these probabilities can be gauged based on the provenance of the corresponding measurement results.

Based on these probabilities, we need to estimate the reliability p of the results of data processing and data fusion.

Case of data processing. Let us first consider the case of data processing, when we transform n measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of n auxiliary quantities into an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of the desired quantity y.

For this estimate to be valid, all n measurement results must be valid (i.e., none of them is an outlier). The probability that the i-th measurement result is not an outlier is equal to $1 - p_i$. Since we assumed independence, the probability $1 - p$ that all n measurement results are not outliers is equal to the product

$$1 - p = \prod_{i=1}^{n} (1 - p_i).$$

Hence,

$$p = 1 - \prod_{i=1}^{n} (1 - p_i).$$

If all the probability p_i are small, we can ignore quadratic and higher order terms in this formula and conclude that $1 - p \approx 1 - \sum\limits_{i=1}^{n} p_i$, i.e., that

$$p \approx \sum_{i=1}^{n} p_i.$$

Comment. Since $p_i > 0$, we have $1 - p < 1 - p_i$ for all i and hence, $p > p_i$. It is worth mentioning that for data processing, the un-reliability p of the result of data processing is larger than each individual probability p_i. Thus, if one of the input measurement results is highly unreliable, the result of data processing is highly unreliable as well.

In particular, if we process n values with the same un-reliability $p_1 = \ldots = p_n$, then the un-reliability p of the result of data processing is n times larger: $p \approx n \cdot p_1 \gg p_1$.

Case of data fusion. The above-described data fusion techniques assume that we use all n results $\widetilde{y}_1, \ldots \widetilde{y}_n$ of measuring the desired quantity y. Thus, for these techniques, the reliability p of the resulting estimate $\widetilde{y}$ can be estimated by using a similar formula

$$p = 1 - \prod_{i=1}^{n} (1 - p_i).$$

For small p_i, we can use a linearized version of this formula $p \approx \sum\limits_{i=1}^{n} p_i$.

So here, just like for data processing, the un-reliability p of the result of data fusion is (much) higher than the un-reliability of individual measurement results.

Case of interval uncertainty. Let us show that in the case of linear uncertainty, we can get much better reliability values than in the general data fusion situation.

Indeed, in the case of interval uncertainty, we start with n intervals $[\underline{y}_i, \overline{y}_i]$ which contain the desired value y. In the "reliable" data fusion, we simply take the intersection of these n intervals, i.e., we take the interval $[\underline{y}, \overline{y}]$, where $\underline{y} = \max(\underline{y}_1, \ldots, \underline{y}_n)$ and $\overline{y} = \min(\overline{y}_1, \ldots, \overline{y}_n)$. The minimum and the maximum are attained for some specific values i and j; thus, we always have $\underline{y} = \underline{y}_i$ for some i and $\overline{y} = \overline{y}_j$ for some appropriate value j. Hence,

- the reliability for the lower endpoint $\underline{y}$ is simply equal to the reliability p_i of the i-th measurement, and
- the reliability of the upper endpoint $\overline{y}$ is equal to the reliability p_j of the j-th measurement.

The corresponding values $p = p_i$ and $p = p_j$ are much better than in the general case when $p \gg p_i$ for all i.

Data fusion can also improve reliability: towards an algorithm. Interval data fusion can lead to even more reliable results: namely, an appropriate data fusion can drastically improve the reliability of the result.

Indeed, let us assume that we have n interval measurements $[\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]$ with reliabilities $p_1, \ldots, p_n$. If these reliability are too high, how can we combine these values to get an estimate for y for which the corresponding probability p does not exceed a given threshold p_0?

Let us illustrate this possibility on the example of the upper endpoint $\overline{y}$ of the desired reliable bound for y. For that, let us sort the values $\overline{y}_i$ into an increasing sequence

$$\overline{y}_{(1)} \le \overline{y}_{(2)} \le \cdots \le \overline{y}_{(n)}.$$

The corresponding probabilities will now be $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$. In the "reliable" data fusion, we simply take the smallest value $\overline{y}_{(1)}$ as the desired estimate $\overline{y}$. For fusing possibly un-reliable data, we can no longer do that.

Instead, let us choose, as $\overline{y}$, the k-th value $\overline{y}_{(k)}$ for some k. This estimate is not valid only in one case: when all k estimates $\overline{y}_{(1)}, \ldots, \overline{y}_{(k)}$ are un-reliable. Since we assumed independence, the probability for this is equal to the product $p_{(1)} \cdot \ldots \cdot p_{(k)}$. Thus, ro guarantee reliability $p \le p_0$, we can select the first k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$. Thus, we arrive at the following algorithm.

Data fusion which improves reliability of interval estimates: an algorithm. We start with n intervals $[\underline{y}_i, \overline{y}_i]$ with are reliable with probabilities p_i. Our objective is to fuse them into a single interval $[\underline{y}, \overline{y}]$ which is the most accurate under the constraint that each of its endpoints $\underline{y}$ and $\overline{y}$ has an un-reliability $\le p_0$ for some given value p_0.

To get the desired value $\overline{y}$, we sort the upper endpoints $\overline{y}_i$ into an increasing sequence $\overline{y}_{(1)} \le \overline{y}_{(2)} \le \ldots \le \overline{y}_{(n)}$, select the smallest k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$, and take $\overline{y} = \overline{y}_{(k)}$.

Similarly, to get the desired value $\underline{y}$, we sort the lower endpoints $\underline{y}_i$ into a decreasing sequence $\underline{y}_{(1)} \ge \underline{y}_{(2)} \ge \ldots \ge \underline{y}_{(n)}$, select the smallest k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$, and take $\underline{y} = \underline{y}_{(k)}$.

Comment. When all input intervals have the same reliability $p_1 = \ldots = p_n$, the condition

$$p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$$

takes the form $p_1^k \le p_0$. The smallest k for which this inequality holds can be easily computed as $k = \left\lceil \dfrac{|\ln(p_0)|}{|\ln(p_1)|} \right\rceil$.

It is worth mentioning that in this case, we do not need to spend $O(n \cdot \log(n))$ times on sorting the bounds, since the k-th value in the ordered sequence can be computed in linear time; see, e.g., (Cormen et al., 2001).

12. Case Study: Seismic Inverse Problem in the Geosciences

12.1. DESCRIPTION OF THE CASE STUDY

Seismic inverse problem in the geosciences: brief reminder. As a case study, we consider the seismic inverse problem in the geosciences; see, e.g., (Averill, 2007; Hole, 1992; Parker, 1994). In this problem, we measure the travel times $x_1, \ldots, x_n$ of the seismic signals and based on these travel times, we reconstruct the velocity of sound $y = f(x_1, \ldots, x_n)$ at different points inside the Earth. There exist several algorithms for reconstructing this velocity. In our research, we use one of most widely used algorithms proposed by J. Hole (Hole, 1992).

Our objective is to estimate the uncertainty of the resulting velocity estimates.

What we plan to do. The problem of estimating uncertainty has been actively researched in geosciences; see, e.g., (Averill et al., 2005; Averill et al., 2007; Doser et al., 1998; Maceira et al., 2005). In this section, we will apply the above-described techniques to this problem, explain the results and their limitations, and provide a heuristic method of overcoming these limitations, a method which can be applied to other problems as well.

Details of this application are presented in (Averill, 2007).

Case study: brief description. The data used for our analysis was obtained from the Potrillo Volcanic Experiment (PVF), a large-scale active source seismology profile designed to investigate the crustal structure across southern New Mexico and Far West Texas. This field experiment was conducted in 2003.

The PVF experiment was composed of 8 shots of 1000–2000 lbs.; 793 seismic recorders (TEXANS) were deployed at variable spacing of 100 m, 200 m and 600 m over 205 km. The location map for the (PVF) experiment is give below. Stars show shot point locations. Small gray dots represent receiver locations. Black box outlines model space for tomography.

The resulting velocity distribution is given presented in the following picture. Velocity model is gridded at 1 × 1 km spacing. Illuminated coloring shows location of ray coverage within the model. Coverage model showing location and coverage density for rays traced within the model is presented in the next picture.

What is the accuracy with which we know these values of velocity?

Main source of direct measurement errors in the seismic inverse problem. The input to the seismic inverse problem consists of travel times x_i. Each travel time is the time that a seismic wave takes to travel from the location of the explosion to the corresponding sensor. It is determined as the first moment of time at which we detect the incoming seismic wave (on top of the noise). For the low-energy artificial explosions which are used in seismic experiments, the signal-to-noise ratio is rather small, especially for sensors located several dozens kilometers away from the experiment location. As a result, we may miss the first peak of the seismic wave and erroneously identify the second peak as the arrival time of the seismic wave.

This "picking error" is the main source of error in measuring travel time. A typical size of a picking error is the time distance between the two peaks of the seismic wave, i.e., about 150 ms.

12.2. First Try: Probabilistic Approach

First try: probabilistic approach. As we have mentioned, traditionally in science and engineering, the probabilistic approach is used to estimate the uncertainty of the result of data processing. In this approach, we assume that the errors of different direct measurements are independent random variables, with 0 mean and known standard deviations σ_i.

This method have been successfully used in geosciences. In particular, it was used in the analysis of the passive seismic inverse problem, when we use only the travel times of the seismic waves generated by the earthquakes; see, e.g., (Maceira et al., 2005). For this problem, the independence assumption makes sense since different earthquakes at different locations are indeed independent.

In view of these past successes, we decided to apply this technique to our active seismic inverse problems.

In principle, we can use the above formula. In principle, to find the desired value σ, we can use the above formula $\sigma = \sqrt{\sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2}$, where each partial derivative can be determined by numerical differentiation, as $c_i = \dfrac{1}{h} \cdot (f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y})$.

Limitations of directly using the above formula. The direct use of the above formula requires that we call the program f (in our case, the program for solving the seismic inverse problem) $n + 1$ times, where n is the total number of inputs: once to compute $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and n more times to compute the values $c_1, \ldots, c_n$.

The problem with directly using the above formula is that the program f requires several hours time to compute (e.g., in the geological applications, computing f may involve solving an inverse problem), and the number n of inputs x_i is in the thousands. Thus, calling the program f $n + 1$ times requires 1,000 times longer than several hours – i.e., several months.

Monte-Carlo simulations: main idea. In the probabilistic setting, we can use straightforward (Monte-Carlo) simulation, and drastically save the computation time. In this approach, we use a computer-based random number generator to simulate the normally distributed error. A standard normal random number generator usually produces a normal distribution with 0 average and standard deviation 1. So, to simulate a distribution Δx_i with a standard deviation σ_i, we multiply the result α_i of the standard Gaussian random number generator by σ_i. In other words, we take $\Delta_i = \sigma_i \cdot \alpha_i$, and we simulate x_i as $\widetilde{x}_i - \Delta x_i$.

As a result of N Monte-Carlo simulations, we get N values

$$\Delta y^{(1)} = c_1 \cdot \Delta x_1^{(1)} + \ldots + c_n \cdot \Delta x_n^{(1)}, \ldots, \Delta y^{(N)} = c_1 \cdot \Delta x_1^{(N)} + \ldots + c_n \cdot \Delta x_n^{(N)}$$

which are normally distributed with the desired standard deviation σ. So, we can determine σ by using the standard statistical estimate

$$\sigma = \sqrt{\frac{1}{N-1} \cdot \sum_{k=1}^{N} \left(\Delta y^{(k)}\right)^2}. \tag{1}$$

Computation time required for Monte-Carlo simulation. The relative error of the above statistical estimate depends only on N (as $\approx 1/\sqrt{N}$), and not on the number of variables n. Therefore, the number N_f of calls to f that is needed to achieve a given accuracy does not depend on the number of variables at all.

The error of the above algorithm is asymptotically normally distributed, with a standard deviation $\sigma_e \sim \sigma/\sqrt{2N}$. Thus, if we use a "two sigma" bound, we conclude that with probability 95%, this algorithm leads to an estimate for σ which differs from the actual value of σ by $\le 2\sigma_e = 2\sigma/\sqrt{2N}$.

This is an error with which we estimate the error of indirect measurement; we do not need too much accuracy in this estimation, because, e.g., in real life, we say that an error is ±10% or ±20%, but *not* that the error is, say, ±11.8%. Therefore, in estimating the error of indirect measurements, it is sufficient to estimate the characteristics of this error with a relative accuracy of, say, 20%.

For the above "two sigma" estimate, this means that we need to select the smallest N for which $2\sigma_e = 2\sigma/\sqrt{2N} \le 0.2 \cdot \sigma$, i.e., to select $N_f = N = 50$.

In many practical situations, it is sufficient to have a standard deviation of 20% (i.e., to have a "two sigma" guarantee of 40%). In this case, we need only $N = 13$ calls to f.

On the other hand, if we want to guarantee 20% accuracy in 99.9% cases, which correspond to "three sigma", we must use N for which $3\sigma_e = 3 \cdot \sigma/\sqrt{2N} \le 0.2 \cdot \sigma$, i.e., we must select $N_f = N = 113$, etc.

For $n \approx 10^3$, all these values of N_f are much smaller than $N_f = n$ required for numerical differentiation.

Additional advantage: parallelization. In Monte-Carlo algorithm, we need 50 calls to f. If each call requires a minute, the resulting time takes about an hour, which may be too long for on-line results. Fortunately, different calls to the function f are independent on each other, so we can run all the simulations in parallel.

The more processors we have, the less time the resulting computation will take. If we have as many processors as the required number of calls, then the time needed to estimate the error of indirect measurement becomes equal to the time of a single call, i.e., to the time necessary to compute the result $\widetilde{y}$ of this indirect measurement. Thus, if we have enough processors working in parallel, we can compute the result of the indirect measurement *and* estimate its error during the same time that it normally takes just to compute the result.

In particular, if the result $\widetilde{y}$ of indirect measurement can be computed in real time, we can estimate the error of this result in real time as well.

Probabilistic case: results. Following the above algorithm, we randomly perturbed the travel-time data by a Gaussian distribution with a standard deviation equal to the "picking error" of 150 ms. The perturbed data was used to generate a new velocity model. This process was repeated multiple times, and the resulting velocity models are used to calculate the RMS difference σ in velocity.

The majority of the values are less than 0.01 km/s (see below).

From our experience of comparing different results, we know that the actual difference between different estimates $\widetilde{y}$ for the velocity is much higher. Thus, these results are misleadingly low.

Also, the results seem to be qualitatively misleading: the values of σ are the highest near the shots (where the reconstruction is more accurate) and smaller elsewhere.

Comment. When instead of single value $\sigma_i = 150$ ms, we used more realistic different values at different sensor locations (obtained by using a technique from (Zelt and Forsyth, 1994)), we got similar results.

Comment. In addition to the Monte-Carlo approach, we also tried the jack-knife approach (see, e.g., (Lees and Crosson, 1989; Tihelaar and Ruff, 1989)) in which the data is divided into two sets (we divided into even and odd sensors). Each of these two data sets was inverted to generate the velocity distribution. The resulting distributions were compared to the result of processing the combined data set; the differences are taken as an estimate for σ.

The resulting values σ are similar to the probabilistic case: the values σ are too low (generally below 0.06 km/s), and qualitatively wrong: the highest values are located near the ends of the profile, adjacent to the shot points and in regions of lower ray coverage.

12.3. SECOND TRY: INTERVAL APPROACH

Toward interval estimates. In our probabilistic estimates, we made a simplifying assumption that the measurement errors of different measurements are independent random variable. Since this assumption is false, this means that there is a correlation between these errors.

We do not know the value of this correlation. It is therefore reasonable to now try the more general interval case, which makes no assumption about the correlations.

Interval approach: brief reminder. In this approach, we assume that we know the upper bounds Δ_i on the measurement errors Δx_i, and we compute the upper bounds Δ on the resulting error Δy.

In our example, we take $\Delta_i = 150$ ms.

In principle, we can use the above explicit formula. In principle, to find the desired value Δ, we can use the above formula $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$, where each partial derivative can be determined by numerical differentiation. However, similarly to the probabilistic case, this method requires that we call f $4n+1$ times – which can lead to months of computations.

To avoid these computations, we use the Cauchy-based method described in (Kreinovich et al., 2007; Kreinovich et al., 2004).

Mathematics behind the Cauchy method. In our simulations, we use *Cauchy distribution* – i.e., probability distributions with the probability density $\rho(z) = \dfrac{\Delta}{\pi \cdot (z^2 + \Delta^2)}$; the value Δ is called the *(scale) parameter* of this distribution.

Cauchy distribution has the following property that we will use: if $z_1, \ldots, z_n$ are independent random variables, and each of z_i is distributed according to the Cauchy law with parameter Δ_i, then their linear combination $z = c_1 \cdot z_1 + \ldots + c_n \cdot z_n$ is also distributed according to a Cauchy law, with a scale parameter $\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n$.

Therefore, if we take random variables δ_i which are Cauchy distributed with parameters Δ_i, then the value

$$\delta \overset{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \delta_1, \ldots, \widetilde{x}_n - \delta_n) = c_1 \cdot \delta_1 + \ldots + c_n \cdot \delta_n$$

is Cauchy distributed with the desired parameter $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$. So, repeating this experiment N times, we get N values $\delta^{(1)}, \ldots, \delta^{(N)}$ which are Cauchy distributed with the unknown parameter, and from them we can estimate Δ.

The bigger N, the better estimates we get.

There are two questions to be solved:

- how to simulate the Cauchy distribution;
- how to estimate the parameter Δ of this distribution from a finite sample.

Simulation can be based on the functional transformation of uniformly distributed sample values: $\delta_i = \Delta_i \cdot \tan(\pi \cdot (r_i - 0.5))$, where r_i is uniformly distributed on the interval $[0, 1]$.

In order to estimate Δ, we can apply the Maximum Likelihood Method

$$\rho(\delta^{(1)}) \cdot \rho(\delta^{(2)}) \cdot \ldots \cdot \rho(\delta^{(N)}) \to \max,$$

where $\rho(z)$ is a Cauchy distribution density with the unknown Δ. When we substitute the above-given formula for $\rho(z)$ and equate the derivative of the product with respect to Δ to 0 (since it is a maximum), we get an equation

$$\frac{1}{1 + \left(\dfrac{\delta^{(1)}}{\Delta}\right)^2} + \ldots + \frac{1}{1 + \left(\dfrac{\delta^{(N)}}{\Delta}\right)^2} = \frac{N}{2}. \tag{2}$$

The left-hand side of (2) is an increasing function that is equal to $0 (< N/2)$ for $\Delta = 0$ and $> N/2$ for $\Delta = \max \left|\delta^{(k)}\right|$; therefore the solution to the equation (2) can be found by applying a bisection method to the interval $\left[0, \max \left|\delta^{(k)}\right|\right]$.

It is important to mention that we assumed that the function f is reasonably linear within the box $[\widetilde{x}_1 - \Delta_1, \widetilde{x}_1 + \Delta_1] \times \ldots \times [\widetilde{x}_n - \Delta_n, \widetilde{x}_n + \Delta_n]$. However, the simulated values δ_i may be outside the box. When we get such values, we do not use the function f for them, we use a normalized function that is equal to f within the box, and that is extended linearly for all other values (we will see, in the description of an algorithm, how this is done).

As a result, we arrive at the following algorithm.

Algorithm.

- Apply f to the results of direct measurements: $\widetilde{y} := f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$;
- For $k = 1, 2, \ldots, N$, repeat the following:
  - use the standard random number generator to compute n numbers $r_i^{(k)}$, $i = 1, 2, \ldots, n$, that are uniformly distributed on the interval $[0, 1]$;
  - compute Cauchy distributed values $c_i^{(k)} := \tan(\pi \cdot (r_i^{(k)} - 0.5))$;
  - compute the largest value of $|c_i^{(k)}|$ so that we will be able to normalize the simulated measurement errors and apply f to the values that are within the box of possible values: $K := \max_i |c_i^{(k)}|$;
  - compute the simulated measurement errors $\delta_i^{(k)} := \Delta_i \cdot c_i^{(k)} / K$;
  - compute the simulated "actual values" $x_i^{(k)} := \widetilde{x}_i - \delta_i^{(k)}$;

  - apply the program f to the simulated "actual values" and compute the simulated error of the indirect measurement:

$$\delta^{(k)} := K \cdot \left(\widetilde{y} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)\right);$$

- Compute Δ by applying the bisection method to solve the equation (2).

Comment. To avoid confusion, we should emphasize that, in contrast to the Monte-Carlo solution for the probabilistic case, the use of Cauchy distribution in the interval case is a computational trick and *not* a truthful simulation of the actual measurement error Δx_i: indeed, we know that the actual value of Δx_i is always inside the interval $[-\Delta_i, \Delta_i]$, but a Cauchy distributed random attains values outside this interval as well.

Interval case: results. The results (given below) show some interesting features which we can use to qualitatively interpret the accuracy of the velocity values. In general, the values correspond well to the density and geometry of ray coverage in the model (see an earlier picture). The lowest values are in the upper part of the model and along paths of greatest ray coverage. The highest values or regions of lowest resolution are deeper in the model, near the center of the model with low ray coverage, and beneath the El Paso area (between shotpoints 5 and 6), where urban noise has decreased the number of travel-time picks and their quality. Whereas theses values do provide a good assessment of reliability for different regions of the model, they are clearly not useful in absolute terms.

Comment. When instead of single value $\Delta_i = 150$ ms, we used more realistic different values at different sensor locations (Zelt and Forsyth, 1994), we got similarly over-large results.

Clarifying comment. The above negative result is easy to explain by the following back-of-the-envelope calculations. Let us take two neighboring sensors at a distance $d = 600$ m from each other. Let x_1 be the time by which the seismic wave arrived at the first sensor, and let x_2 be the time by which this wave arrived at the second sensor. This means that this wave took time $t_2 - t_1$ to travel a distance d between the two sensors and thus, its velocity in this area can be estimated as $v = d/(x_2 - x_1)$.

The actual velocity near the surface is about $v \approx 2$ km/s, so the actual time difference is $x_2 - x_1 = d/v \approx 0.3$ sec. The observed value $\widetilde{x}_2 - \widetilde{x}_1$ is thus 0.3 sec, and the upper bound on each measurement error Δx_1 and Δx_2 is 0.15 sec. Thus, the upper bound on the error $\Delta x_2 - \Delta x_1$ is 0.3 sec. So, by using interval uncertainty, we conclude that the actual (unknown) value of $x_2 - x_1$ can take any value from 0 to 0.6 sec. When $x_2 - x_1$ is close to 0, for the corresponding velocity $v = d/(x_2 - x_1)$ we get meaningless thousands of km/s.

12.4. A New Heuristic Approach

Towards the main idea. The *guaranteed* bounds provided by the interval approach are too high. How can we improve these bounds?

One possible solution comes from the following simple observation. For the normally distributed random variable with 0 mean and standard deviation σ, the only guaranteed upper bound is ∞. In practice, however, we can say that with confidence 90%, the actual value of this variable does not exceed 2σ, with confidence 99.9%, it does not exceed 3σ, etc. To get a bound with 90% confidence, we "cut-off" the top 10% of the normal distribution. To get the bound with 99.9% confidence, we "cut-off" the top 0.1% of the normal distribution, etc.

Main idea. Since guaranteed bounds are too high, it is reasonable to restrict ourselves to bounds guaranteed with a given confidence, e.g., bounds which are guaranteed with a confidence of 95% and dismiss the top 5% of uncertainty values. To find such bounds in the Cauchy method, we "cut-off" the top 5% of the corresponding Cauchy distribution. To be more precise, we find the threshold value x_0 for which the probability of exceeding this value is 5% (or any other desired cut-off probability p_0), and then replace values x for which $x > x_0$ with x_0 and for $x < -x_0$ with $-x_0$. For the Cauchy distribution, we have found that a 95% confidence level is obtained for the bounds of $-12.706 \le x_0 \le 12.706$ (see Appendix).

So, to get more realistic estimates for Δ, in the Cauchy approach, we use the "cut-off" Cauchy distribution instead of the original one.

Heuristic approach: results. The results of applying the Cauchy approach with 95% and 90% confidence are presented on the next page. Good news is that, in contrast to the practically useless interval-case values of uncertainty, here, velocity uncertainties Δ are exactly as expected. At the 95% confidence, the values Δ range from 0.01 to 0.3 km/s, and at a 90% confidence level, they range from 0.005 to 0.23 km/s.

On the qualitative level, the values Δ are still as geophysically reasonable as the values computed by the original interval-case method:

- the lowest values of Δ are found near the shotpoints, and along paths of highest ray coverage;
- the highest uncertainties are near the center of the model with lowest ray coverage and beneath the El Paso region between shotpoints 5 and 6.

Conclusions

In the past, communications were much slower than computations. As a result, researchers and practitioners collected different data into huge databases located at a single location such as NASA and US Geological Survey. At present, communications are so much faster that it is possible to keep different databases at different locations, and automatically select, transform, and collect relevant data when necessary. The corresponding cyberinfrastructure is actively used in many applications. It drastically enhances scientists' ability to discover, reuse and combine a large number of resources, e.g., data and services.

Because of this importance, it is desirable to be able to gauge the the uncertainty of the results obtained by using cyberinfrastructure. This problem is made more urgent by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – and that scientists have much less control over the quality of different resources than in the centralized database. Thus, with the cyberinfrastructure promise comes the need to analyze how data uncertainty *propagates* via this cyberinfrastructure.

When the resulting accuracy is too low, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result. In this paper, we describe algorithms for propagating uncertainty and for finding the provenance for this uncertainty.

The above results mainly deal either with the *probabilistic* situations, when we either know the probability distributions of different measurement errors (and different errors are independent), or with *interval* situations, when we only know the upper bounds on the measurement errors. Probabilistic estimates tend to *underestimate* the resulting error – since in reality, different measurement errors are correlated (e.g., they have the same systematic error components). Interval estimates tend to *overestimate* because they are based on – often unrealistic – worst-case scenarios. It is thus desirable to combine these estimates to get more realistic error bounds. We describe several such combination methods, their mathematical justifications, and their successful use in processing geospatial data.

Acknowledgements

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Japan Advanced Institute of

Science and Technology (JAIST) International Joint Research Grant 2006-08, and by the Max Planck Institut für Mathematik.

References

Aguiar, M. S., G. P. Dimuro, A. C. R. Costa, R. K. S. Silva, F. A. Costa, and V. Kreinovich. The multi-layered interval categorizer tesselation-based model. In: C. Iochpe and G. Câmara, Editors. *IFIP WG2.6 Proceedings of the 6th Brazilian Symposium on Geoinformatics Geoinfo'2004*, Campos do Jordão, Brazil, November 22–24, 2004, pages 437–454.

Aldouri R., G. R. Keller, A. Q. Gates, J. Rasillo, L. Salayandia, V. Kreinovich, J. Seeley, P. Taylor, and S. Holloway. GEON: Geophysical data add the 3rd dimension in geospatial studies. In: *Proceedings of the ESRI International User Conference 2004*, San Diego, California, August 9–13, 2004, Paper 1898

Averill, M. G. *A Lithospheric Investigation of the Southern Rio Grande Rift*, University of Texas at El Paso, Department of Geological Sciences, PhD Dissertation, 2007.

Averill, M. G., K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks. Using expert knowledge in solving the seismic inverse problem. In: *Proceedings of the 24nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pages 310–314

Averill, M. G., K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks. Using Expert Knowledge in Solving the Seismic Inverse Problem. *International Journal of Approximate Reasoning*, 453):564–578, 2007.

Ceberio, M., S. Ferson, V. Kreinovich, S. Chopra, G. Xiang, A. Murguia, and J. Santillan. How to take into account dependence between the inputs: from interval computations to constraint-related set computations, with potential applications to nuclear safety, bio- and geosciences. In: *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, February 22–24, 2006, pages 127–154.

Ceberio, M., V. Kreinovich, S. Chopra, and B. Ludäscher. Taylor model-type techniques for handling uncertainty in expert systems, with potential applications to geoinformatics. In: *Proceedings of the 17th World Congress of the International Association for Mathematics and Computers in Simulation IMACS'2005*, Paris, France, July 11–15, 2005.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

Doser, D. I., K. D. Crain, M. R. Baker, V. Kreinovich, and M. C. Gerstenberger. Estimating uncertainties for geophysical tomography. *Reliable Computing*, 4(3):241–268, 1998.

Fuller, W. A. *Measurement error models.* J. Wiley & Sons, New York, 1987.

Gates A. Q., V. Kreinovich, L. Longpré, P. Pinheiro da Silva, and G. R. Keller. Towards secure cyberinfrastructure for sharing border information. In: *Proceedings of the Lineae Terrarum: International Border Conference*, El Paso, Las Cruces, and Cd. Juárez, March 27–30, 2006.

Hole, J. A. Nonlinear High-Resolution Three-Dimensional Seismic Travel Time Tomography. *J. Geophysical Research*, 97(B5):6553–6562, 1992.

Jaulin, L., M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*, Springer Verlag, London, 2001.

Keller, G. R., T. G. Hildenbrand, R. Kucks, M. Webring, A. Briesacher, K. Rujawitz, A. M. Hittleman, D. J. Roman, D. Winester, R. Aldouri, J. Seeley, J. Rasillo, T. Torres, W. J. Hinze, A. Gates, V. Kreinovich, and L. Salayandia. A community effort to construct a gravity database for the United States and an associated Web portal. In: A. K. Sinha, Editor. *Geoinformatics: Data to Knowledge*, pages 21–34, Geological Society of America Publ., Boulder, Colorado, 2006.

Kreinovich, V., J. Beck, C. Ferregut, A. Sanchez, G. R. Keller, M. G. Averill, and S. A. Starks. Monte-Carlo-type techniques for processing interval uncertainty, and their potential engineering applications. *Reliable Computing*, 13(1):25–69, 2007.

Kreinovich, V., and S. Ferson. A new Cauchy-Based black-box technique for uncertainty in risk analysis. *Reliability Engineering and Systems Safety*, 85(1–3):267–279, 2004.

Kreinovich, V., A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

Longpré, L., and V. Kreinovich. How to efficiently process uncertainty within a cyberinfrastructure without sacrificing privacy and confidentiality", In N. Nedjah, A. Abraham, and L. de Macedo Mourelle, Editors. *Computational Intelligence in Information Assurance and Security*, pages 155–173, Springer-Verlag, 2007.

Lees, J. M., and R. S. Crosson. Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data. *Journal of Geophysical Research*, 94:5716–5728, 1989.

Maceira, M., S. R. Taylor, C. J. Ammon, X. Yang, and A. A. Velasco, High-resolution Rayleigh wave slowness tomography of Central Asia. *Journal of Geophysical Research*, Vol. 110, paper B06304, 2005.

Nguyen, H. T., O. Kosheleva, V. Kreinovich, and S. Ferson. Trade-Off Between Sample Size and Accuracy: Case of Dynamic Measurements under Interval Uncertainty. *Proceedings of International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics UncLog'08*, JAIST, Japan, March 25–28, 2008 (to appear).

Nguyen, H. T., and V. Kreinovich. Trade-Off Between Sample Size and Accuracy: Case of Static Measurements under Interval Uncertainty. *Proceedings of International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics UncLog'08*, JAIST, Japan, March 25–28, 2008 (to appear).

Parker, R. L. *Geophysical Inverse Theory*, Princeton University Press, Princeton, New Jersey, 1994.

Platon, E., K. Tupelly, V. Kreinovich, S. A. Starks, and K. Villaverde. Exact bounds for interval and fuzzy functions under monotonicity constraints, with potential applications to biostratigraphy. In: *Proceedings of the 2005 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2005*, Reno, Nevada, May 22–25, 2005, pages 891–896.

Rabinovich, S. G. *Measurement Errors and Uncertainty. Theory and Practice*, Springer Verlag, Berlin, 2005.

Schiek, C. G., R. Araiza, J. M. Hurtado, A. A. Velasco, V. Kreinovich, and V. Sinyansky. Images with Uncertainty: Efficient Algorithms for Shift, Rotation, Scaling, and Registration, and Their Applications to Geosciences. In: M. Nachtegael, D. Van der Weken, E. E. Kerre, and Wilfried Philips (eds.), Soft Computing in Image Processing: Recent Advances, Springer Verlag, 2007, pp. 35-64.

Sinha, A. K., Editor. *Geoinformatics: Data to Knowledge*, Geological Society of America Publ., Boulder, Colorado, 2006.

Tichelaar, B. W., and L. R. Ruff. How good are our best models? *EOS*, 70:593–606, 1989.

Torres R., G. R. Keller, V. Kreinovich, L. Longpré, and S. A. Starks. Eliminating duplicates under interval and fuzzy uncertainty: an asymptotically optimal algorithm and its geospatial applications. *Reliable Computing*, 10(5):401–422, 2004.

Vavasis, S. A. *Nonlinear Optimization: Complexity Issues.* Oxford University Press, New York, 1991.

Walster, G. W. Philosophy and practicalities of interval arithmetic. In: *Reliability in Computing*, pages 309–323, Academic Press, N.Y., 1988.

Walster, G. W., and V. Kreinovich. For unknown-but-bounded errors, interval estimates are often better than averaging. *ACM SIGNUM Newsletter*, 31(2):6–19, 1996.

Wen Q., A. Q. Gates, J. Beck, V. Kreinovich, J. R. Keller. Towards automatic detection of erroneous measurement results in a gravity database. In: *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference*, Tucson, Arizona, October 7–10, 2001, pages 2170–2175.

Xie H., N. Hicks, G. R. Keller, H. Huang, and V. Kreinovich. An IDL/ENVI implementation of the FFT based algorithm for automatic image registration. *Computers and Geosciences*, 29(8):1045–1055, 2003.

Zelt, C. A., and P. J. Barton. Three-dimensional seismic refraction tomography: A comparison of two methods applied to data from the Faeroe Basin. *J. Geophysical Research*, 103(B4):7187–7210, 1998.

Zelt, C. A., and D. A. Forsyth. Modeling wide-angle seismic data for crustal structure Grenville province. *J. of Geophys. Res.* 99:11687-11704, 1994.

Appendix

The standard Cauchy distribution is characterized by the probability density function $\rho(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$. We are given a small probability p_0 (e.g., $p_0 = 5\%$), and we want to find the value x_0 such that the probability that $|x| \geq x_0$ is exactly p_0. In other words, we want the probability that $x \geq x_0$ or $x \leq -x_0$ to be equal to p_0. Since the Cauchy distribution is symmetric, the probability that $x \leq -x_0$ is equal to the probability that $x \geq x_0$. Therefore, the probability that $|x| \geq x_0$ is equal to twice the probability that $x \geq x_0$: $p_0 = \text{Prob}(|x| \geq x_0) = 2 \cdot \text{Prob}(x > x_0)$.

For the Cauchy distribution,

$$\frac{p_0}{2} = \text{Prob}(x > x_0) = \frac{1}{\pi} \cdot \int_{x_0}^{\infty} \frac{1}{1+x^2} = \frac{1}{\pi} \cdot (\arctan(\infty) - \arctan(x_0)) =$$

$$\frac{1}{\pi} \cdot \left(\frac{\pi}{2} - \arctan(x_0)\right) = \frac{1}{2} - \frac{1}{\pi} \cdot \arctan(x_0).$$

Thus, we must take $\arctan(x_0) = \frac{\pi}{2} \cdot (1 - p_0)$ and

$$x_0 = \tan\left(\frac{\pi}{2} \cdot (1 - p_0)\right). \tag{3}$$

For small p_0, we can get an even simpler formula. Indeed, in general, $x_0 = \tan\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right) = \frac{\sin\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right)}{\cos\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right)}$. We know that $\sin\left(\frac{\pi}{2} - \alpha\right) = \cos(\alpha)$ and $\cos\left(\frac{\pi}{2} - \alpha\right) = \sin(\alpha)$, so $x_0 = \frac{\cos\left(\frac{\pi}{2} \cdot p_0\right)}{\sin\left(\frac{\pi}{2} \cdot p_0\right)}$. For small α, we have $\sin(\alpha) \approx \alpha$ and $\cos(\alpha) \approx 1$, hence for small p_0, we get $x_0 \approx \frac{1}{\frac{\pi}{2} \cdot p_0}$ and

$$x_0 \approx \frac{2}{\pi \cdot p_0}. \tag{4}$$