# Robust Data Processing in the Presence of Uncertainty and Outliers: Case of Localization Problems

Anthony Welte and Luc Jaulin
Lab STICC
École Nationale Supérieure
de Techniques Avancées Bretagne
(ENSTA Bretagne)
2 rue François Verny
29806 Brest, France
Emails: tony.welte@gmail.com,
lucjaulin@gmail.com

Martine Ceberio and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
Emails: mceberio@utep.edu,
vladik@utep.edu

***Abstract*—To properly process data, we need to take into account both the measurement errors and the fact that some of the observations may be outliers. This is especially important in radar-based localization problems, where some signals may reflect not from the analyzed object, but from some nearby object. There are known methods for dealing with both measurement errors and outliers in situations in which we have full information about the corresponding probability distributions. There are also known statistics-based methods for dealing with measurement errors in situations when we only have partial information about the corresponding probabilities. In this paper, we show how these methods can be extended to situations in which we also have partial information about the outliers (and even to situations when we have no information about the outliers). In some situations in which efficient semi-heuristic methods are known, our methodology leads to a justification of these efficient heuristics – which makes us confident that our new methods will be efficient in other situations as well.**

## I. Formulation of the Problem

**Need for data processing.** In many practical situations, we are interested in the values of the quantities $p_1, \ldots, p_m$ which are difficult to measure directly.

For example, when solving a localization problem – whether it is a problem of locating a robot (see, e.g., [3]) or of locating a satellite (see, e.g., [4]) – we are interested in the coordinates $p_1, \ldots$ of this object. It is possible to directly measure physical quantities such as distance, velocity, density, etc. However, coordinates are an artificial construction that does not directly correspond to any physical quantity. As a result, it is not possible to directly measure coordinates of an object.

The quantities of interest do affect results of some measurements; namely, the value of the corresponding easier-to-measure quantity $y$ depends, in a known way, on the values $p_1, \ldots, p_m$ – and on some auxiliary quantities $x_1, \ldots, x_n$ that describe the measurement's setting:

$$y = f(p_1, \ldots, p_m, x_1, \ldots, x_n).$$

For example, to determine 3-D coordinates $(p_1, p_2, p_3)$ of an object, we can measure the distance

$$y = \sqrt{\sum_{i=1}^{3} (p_i - x_i)^2}$$

between the object of interest and another object with known coordinates $(x_1, x_2, x_3)$.

So, to find the values of $p_i$, we measure the value $y_k$ of the corresponding quantity $y$ under different settings $(x_{k1}, \ldots, x_{kn})$, and then reconstruct the desired values $p_i$ from the condition that

$$y_k = f(p_1, \ldots, p_m, x_{k1}, \ldots, x_{kn}) \tag{1}$$

for all the measurements $k = 1, \ldots, K$.

For example, to locate an object, we measure the distance between this object and several objects with known coordinates. This is how, e.g., radar-based systems determine the coordinates of an airplane.

Such reconstruction is an important case of *data processing*.

**Need to take into account measurement uncertainty and outliers.** Measurement are never absolutely accurate; see, e.g., [18]. As a result, there is always a non-zero difference between the measurement result $y_k$ and the actual (unknown) value $f(p_1, \ldots, p_m, x_{k1}, \ldots, x_{kn})$ of the corresponding quantity:

$$\Delta y_k \stackrel{\text{def}}{=} y_k - f(p_1, \ldots, p_m, x_{k1}, \ldots, x_{kn}) \neq 0. \tag{2}$$

It is important to take into account this measurement uncertainty when processing data.

Measurement errors are usually reasonably small. Hence, the measured value $y_k$ is usually close to the actual value

$f(p_1, \ldots, p_m, x_{k1}, \ldots, x_{kn})$. However, the measuring instrument is not always 100% reliable. Sometimes, the measuring instrument malfunctions, and we get *outliers*, values which are very different from the actual values of the corresponding quantity. In processing data, we also need to take into account the existence of outliers.

This is especially important in localization problems, where the radar-type signal, instead of reflecting from the desired object, reflects from some other objects. In this case, the corresponding measurement result describes the distance to a different object – i.e., from the viewpoint of our problem, is an outlier.

**What is known, what are the remaining problems, and what we do in this paper.** There are many efficient techniques for taking into account measurement uncertainty. There are also techniques for taking into account outliers, and there are techniques for taking into account *both* measurement uncertainty and outliers.

Such methods work well if we have a complete knowledge about the probabilities of different values of the measurement error and the probabilities of different outliers. In practice, however, we often only have a *partial* information about these probabilities – all the way to the case when we have no information about such probabilities at all; see, e.g., [18]. In such extreme situations, there are methods that take into account either measurement uncertainty or outliers – but not both. In this paper, we briefly overview and analyze the existing techniques of taking into account measurement uncertainty and outliers, and then use this analysis to develop a natural new technique for taking into account both measurement uncertainty and outliers.

The structure of this paper is as follows. In Section 2, we describe the methods of dealing with uncertainty – beware, however that we will describe them in such a way so as to prepare us for the new technique. In Section 3, we use our analysis to show how outliers can also be taken into account.

Most of our results are new. In some cases, as a particular case of our general approach, we get a well-known effective outlier-processing technique; the fact that in some cases, we get well-known well-established efficient techniques makes us confident that our method will be efficient in other situations as well.

## II. How Measurement Uncertainty Is Usually Taken Into Account

**Why this section is needed.** In order to formulate our new results, let us briefly recall how measurement uncertainty is usually taken into account. This recollection is necessary, since our new methods for taking into account both uncertainty and outliers are extensions of the existing methods of taking uncertainty into account.

*Comment.* This section is intended for a general reader, a reader who may not be well familiar with the motivations behind (and details of) all the existing techniques for data processing under uncertainty – such as the Maximum Entropy techniques or interval computations. Readers who are well familiar with all these techniques can simply browse through this section.

**Case when we know the exact probability distribution of the measurement error.** Let us first consider a situation in which we have a complete information about the probability density function $\rho(\Delta y)$ that describes the probability distribution of the measurement error. In this case, once we have the measurement results $y_k$ $(1 \le k \le K)$ corresponding to settings $x_k = (x_{k1}, \ldots, x_{kn})$, then for each parameter tuple $p = (p_1, \ldots, p_m)$ and for each $k$, the probability to observe $y_k$ is proportional to $\rho(\Delta y_k) = \rho(y_k - f(p, x_k))$.

Measurement errors corresponding to different measurements are usually independent. Thus, the probability of observing all the observed values $y_1, \ldots, y_K$ is equal to the product of the probabilities of observing each value $y_k$. Thus, this probability is proportional to the product $\prod_{k=1}^{K} \rho(y_k - f(p, x_k))$.

In this case, we usually have different parameter tuples which are consistent with the given observations. If we need to select a single "best estimate", it is reasonable to select the parameter tuple which is the most probable, i.e., for which the product $L \stackrel{\text{def}}{=} \prod_{k=1}^{K} \rho(y_k - f(p, x_k))$ takes the largest possible value. This idea is known as the *Maximum Likelihood Method*; see, e.g., [14]. Under reasonable conditions, this method indeed leads to estimates which are optimal in some reasonable senses; see, e.g., [14], [19].

**Example.** Let us consider a simple example, in which the measurement error is normally distributed with 0 mean and a known standard deviation $\sigma$. In this case, the probability density function has the form

$$\rho(\Delta y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(\Delta y)^2}{2\sigma^2}\right).$$

Minimizing the corresponding product

$$L = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(\Delta y_k)^2}{2\sigma^2}\right) \tag{2}$$

is equivalent to minimizing minus logarithm of this product

$$\psi \stackrel{\text{def}}{=} -\ln(L) = K \cdot \ln(\sqrt{2\pi} \cdot \sigma) + \frac{1}{2\sigma^2} \cdot \sum_{k=1}^{K} (\Delta y_k)^2. \tag{3}$$

One can easily see that this minimization is equivalent to minimizing the sum

$$\sum_{k=1}^{K} (\Delta y_k)^2 = \sum_{k=1}^{K} (y_k - f(p, x_k))^2.$$

This minimization – known as the *Least Squares Method* – is one of the most widely used data processing techniques.

**What is we only have partial information about the probabilities: case of a finite-parametric family.** In some cases, we do not know the exact probability distribution of

the measurement errors, but we are aware that it belongs to a known finite-parametric family of probability distributions $\rho(\Delta y, \theta)$ depending on the parameter tuple $\theta = (\theta_1, \ldots, \theta_\ell)$.

In this case, the corresponding "likelihood function" $L$ takes the form $L = \prod_{k=1}^{K} \rho(\Delta y_k, \theta)$. Now, instead of selecting only only the parameters $p$ of the model, we also need to select the parameters $\theta$ of the corresponding probability distribution. In this case, it is reasonable to select the most probable pair $(p, \theta)$, i.e., the pair for which the product

$$L = \prod_{k=1}^{K} \rho(y_k - f(p, x_k), \theta)$$

takes the largest possible value.

**Example.** Let us assume that the measurement error is normally distributed with 0 mean, but this time, the standard deviation $\sigma$ is unknown. In this case, we have $\ell = 1$ and $\theta_1 = \sigma$. So, we need to maximize the expression (2) – or, equivalently, minimize the expression (3) – with respect to both $p$ and $\sigma$.

Minimizing the expression (3) with respect to parameters $p$ leads to the same Least Squares estimate as before. Once we find $p$, we can differentiate the expression (3) with respect to $\sigma$, equate the derivative to 0, and get the desired expression

$$\sigma = \sqrt{\frac{1}{K} \cdot \sum_{k=1}^{K} (y_k - f(p, x_k)^2}.$$

**What if we only have partial information about the probabilities: non-parametric case.** In many practical situations, we do not know the finite-parametric family containing the actual distribution. For example, often, all we know is the upper bound $\Delta$ on the measurement error; see, e.g., [18]. In this case, the only information that we have about the actual probability distribution $\rho(\Delta y)$ is that this distribution is located somewhere on the interval $[-\Delta, \Delta]$.

There are many such probability distributions. To apply the Maximum Likelihood principle in this case, we need to select a single "most reasonable" distribution from all these possible distributions. Each of these distributions $\rho(\Delta y)$ can be characterized by its uncertainty (entropy)

$$S = -\int \rho(\Delta y) \cdot \ln(\rho(\Delta y))\, d\Delta y$$

that describes how many binary ("yes"-"no") questions we need to ask to uniquely determine the corresponding value $\Delta y$; see, e.g., [2], [8], [16].

Out of all possible distributions $\rho(\Delta y)$ we have distributions located on a single value $v$. For these distributions, we do not need any questions, we already know the value $v$. However, selecting such a distribution would be cheating – in actuality, we do not know the value $\Delta y$, so we would like to select the distribution that to the largest extent reflects this uncertainty. In other words, it is reasonable to select a distribution for which the entropy is the largest possible.

For all the distributions $\rho(\Delta y)$ located on the interval $[-\Delta, \Delta]$, maximum of entropy under the constraint $\int_{-\Delta}^{\Delta} \rho(\Delta y)\, d\Delta y = 1$ can be obtained by using the Lagrange multiplier method, that reduces the corresponding constraint optimization problem to the unconstrained optimization problem

$$-\int_{-\Delta}^{\Delta} \rho(\Delta y) \cdot \ln(\rho(\Delta y))\, d\Delta y + \lambda \cdot \left( \int_{-\Delta}^{\Delta} \rho(\Delta y)\, d\Delta y - 1 \right) \to \max_{\rho(\Delta y)}$$

for an appropriate Lagrange multiplier $\lambda$. Differentiating this expression with respect to $\rho(\Delta y)$ and equating the derivative to 0, we conclude that $\rho(\Delta y) = \text{const}$, i.e., that we have a uniform distribution on the interval $[-\Delta, \Delta]$, with the probability density

$$\rho(\Delta y) = \frac{1}{2\Delta}.$$

This selection makes perfect sense: since we have no reason to believe that some values from the interval $[-\Delta, \Delta]$ are more probable than others, it is therefore reasonable to conclude that all the values from this interval are equally probable. This argument goes back to Laplace and is thus known as *Laplace Indeterminacy Principle*.

Now that we have selected a probability distribution, we can use the Maximum Likelihood method to find the corresponding parameter values $p$. In this case, each probability density $\rho(\Delta y_k)$ is equal to 0 if $\Delta y_k$ is outside the interval $[-\Delta, \Delta]$ and to a constant (equal to $1/(2\Delta)$) when $\Delta y_k$ inside this interval. Thus, the product $L$ of the corresponding probabilities is equal to 0 if one of the values $\Delta y_k$ is inside the interval, and to the same constant $\frac{1}{(2\Delta)^K}$ when $|\Delta y_k| \le \Delta$ for all $k$. So, instead of a *single* tuple $p$, we know need to describe *all* the tuples $p$ for which $|y_k - f(p, x_k)| \le \Delta$ for all $k = 1, \ldots, k$.

The problem of finding the range of such tuples under *interval uncertainty* ($\Delta y_k \in [-\Delta, \Delta]$) is a particular case of *interval computations*; see, e.g., [5], [15]. In interval computations, there are many efficient techniques for solving this problem [5], [15].

**What if we have no information whatsoever about the probabilities of measurement errors.** In some practical situations, we have no information at all about the probability distribution $\rho(\Delta y)$ of the corresponding measurement error. This situation is somewhat similar to the previous one – with the only difference that now, we do not know the bound $\Delta$.

How can we find a good estimate for this value $\Delta$? A reasonable idea is to use the Maximum Likelihood method and select the value $\Delta$ for which the corresponding likelihood $L = \frac{1}{(2\Delta)^K}$ is the largest possible. Once can easily see that the smaller $\Delta$, the larger this likelihood $L$. Thus, selecting the largest possible $L$ is equivalent to selecting the smallest possible $\Delta$.

The only constraints on $\Delta$ is that we should have $\Delta \ge |\Delta y_k|$ for all $k$. This is equivalent to having $\Delta \ge \max_k |\Delta y_k|$.

The smallest value satisfying this inequality is the value $\Delta = \max\limits_k |\Delta y_k|$. Thus, minimizing $\Delta$ means selecting the parameter $p$ for which the corresponding maximum

$$\max_k |\Delta y_k| = \max_k |y_k - f(p, x_k)|$$

is the smallest possible; see, e.g., [9].

The corresponding minimax approach is indeed frequently used in data processing; see, e.g., [1], [5], [6], [11], [12], [13], [20], [21], [22], [23].

## III. New Results: How to Take Both Uncertainty and Outliers into Account

**Which cases are possible?** In the previous section, we considered possible types of knowledge about the probability distribution. In our analysis, we considered the following four cases, in the decreasing order of the available information about the probabilities:

- we know the exact distribution;
- we know the finite-parametric family of distributions;
- we know the upper bound on the (absolute value) of the corresponding difference; and
- we have no information whatsoever, not even the upper bound.

If we take outliers into account, then, in principle, we may have the same four possible types of information about the corresponding probability density function $\rho_0(\Delta y)$. At first glance, it may therefore seem that we can have $4 \times 4 = 16$ possible combinations. In reality, however, not all such combinations are possible.

Indeed, once we gather enough data, we can determine the corresponding probability distributions. Thus, the fact that we do not yet have detailed information about the probability distribution of the measurement error means that we have not yet collected a sufficient number of measurement results. In this case – since the number of outlier is usually much smaller than the number of actual measurement results – we have even fewer outliers. So, if we cannot determine the probability distribution for the measurement errors, even more so, we cannot determine the probability distribution for the outliers either. In general, for the same reason, the amount of information that we have about the outliers is smaller than the amount of information that we have about the measurement errors.

Hence, instead of 16 options, we only have options in which the amount of information about the outlier-related probability distribution $\rho_0(\Delta y)$ does not exceed the amount of information about the probabilities of measurement errors $\rho_(\Delta y)$. Let us consider all these cases one by one.

**Full information about both distributions.** Let us first consider the ideal case, when we have the complete information about the probabilities. Specifically:

- we know the probability density function $\rho(\Delta y)$ that describes the probability of different values of the measurement error, and
- we know the probability density function $\rho_0(\Delta y)$ that describes the probability of different values of the difference $\Delta y = y - f(p, x)$ corresponding to outliers $y$.

In this case, once we have the measurement results $y_k$ (some of which may come from malfunctioning and are thus outliers), the probability of these observations occurring depends not only on the parameters $p$, but also on which of the values $y_k$ are outliers and which are actual measurement results. Once we know the set $M \subseteq \{1, \ldots, K\}$ of indices $k$ for which $y_k$ is the actual measurement, we can then compute the probability $L$ as

$$L = \left( \prod_{k \in M} \rho(\Delta y_k) \right) \cdot \left( \prod_{k \notin M} \rho_0(\Delta y_k) \right).$$

Now, we can use the Maximum Likelihood approach to determine both the parameter tuple $p$ and the set $M$.

Once $p$ is found, and thus, the values $\Delta y_k = y_k - f(p, x_k)$ are determined, maximizing the product $L$ means:

- selecting $k \in M$ if the value $\rho(\Delta y_k)$ is larger than $\rho_0(\Delta y_k)$, and
- selecting $k \notin M$ if the value $\rho_0(\Delta y_k)$ is larger than $\rho(\Delta y_k)$.

In both cases, the resulting factor in the product $L$ takes the form $\max(\rho(\Delta y_k)), \rho_0(\Delta y_k))$.

The resulting value $L$ takes the following form:

$$L = \prod_{k=1}^{K} \max(\rho(\Delta y_k)), \rho_0(\Delta y_k)) =$$

$$\prod_{k=1}^{K} \max(\rho(y_k - f(p, x_k)), \rho_0(y_k - f(p, x_k))).$$

We thus need to select the parameters $p$ for which this product attains the largest possible value.

*Comment.* From the computational viewpoint, the corresponding problem is similar to the usual maximum likelihood problem, with a new function $g(\Delta y) \stackrel{\text{def}}{=} \max(\rho(\Delta y), \rho_0(\Delta y))$ instead of the original probability density function $\rho(\Delta y)$. It is worth mentioning, however, that, in contrast to the probability density function $\rho(\Delta y)$ for which $\int \rho(\Delta y)\, dy = 1$, for the new function $g(\Delta y)$, we have, in general,

$$\int g(\Delta y)\, dy > \int \rho(\Delta y)\, dy = 1$$

(as long as the probability densities $\rho(\Delta y)$ and $\rho_0(\Delta y)$ are different).

**Full information about $\rho(\Delta y)$, finite-parametric family for $\rho_0(\Delta y)$.** In this case, instead of single distribution $\rho_0(\Delta y)$, we have a finite-parametric family of distributions $\rho_0(\Delta y, \varphi)$ with unknown parameters $\varphi$. In such a situation, we need to determine all the parameters $p$ and $\varphi$ from the requirement that the likelihood

$$L = \prod_{k=1}^{K} \max(\rho(\Delta y_k), \rho_0(\Delta y_k, \varphi)) =$$

$$\prod_{k=1}^{K} \max(\rho(y_k - f(p, x_k)), \rho_0(y_k - f(p, x_k), \varphi))$$

attains the largest possible value.

**Full information about $\rho(\Delta y)$, bound $W$ on the outlier-related differences $\Delta y_k$.** In this case, based on the maximum entropy approach, as a distribution $\rho_0(\Delta y)$, we select a uniform distribution on the interval $[-W, W]$, with the probability density $\rho_0(\Delta y_k) = \dfrac{1}{2W}$.

In such a situation, we determine the parameters $p$ from the requirement that the likelihood

$$L = \prod_{k=1}^{K} \max\left(\rho(\Delta y_k), \frac{1}{2W}\right) =$$

$$\prod_{k=1}^{K} \max\left(\rho(y_k - f(p, x_k)), \frac{1}{2W}\right)$$

attains the largest possible value under the constraint that

$$|\Delta y_k| = |y_k - f(p, x_k)| \le W$$

for all $k = 1, \ldots, K$.

**Full information about $\rho(\Delta y)$, no information whatsoever about the outlier-related differences $\Delta y_k$.** In this case, we select the value $W$ for which the likelihood $L$ as described in the previous example if the largest possible – under the constraint that $|\Delta y_k| \le W$ for all $k$.

One can easily see that the smaller the bound $W$, the larger the density $\dfrac{1}{2W}$ and thus, the larger the likelihood function. Thus, to determine the largest possible value of the likelihood function $L$, we must select the smallest possible value $W$. The constraints on $W$ have the form that $W \ge |\Delta y_k|$ for all $k$. The smallest possible value $W$ that satisfies all these constraints is the value

$$W = \max_{\ell} |\Delta y_\ell| = \max_{\ell} |y_\ell - f(p, x_\ell)|.$$

Substituting this expression into the above formula, we conclude that we need to select the parameters $p$ for which the likelihood

$$L = \prod_{k=1}^{K} \max\left(\rho(y_k - f(p, x_k)), \frac{1}{2 \cdot \max\limits_{\ell} |y_\ell - f(p, x_\ell)|}\right)$$

attains the largest possible value.

**Finite-parametric information about $\rho(\Delta y)$ and about $\rho_0(\Delta)$.** In this case, instead of single distributions $\rho(\Delta y)$ and $\rho_0(\Delta y)$, we have finite-parametric families of distributions $\rho(\Delta y, \theta)$ and $\rho_0(\Delta y, \varphi)$ with unknown parameters $\theta$ and $\varphi$. In such a situation, we need to determine all the parameters $p$, $\theta$, and $\varphi$ from the requirement that the likelihood

$$L = \prod_{k=1}^{K} \max(\rho(\Delta y_k, \theta)), \rho_0(\Delta y_k, \varphi)) =$$

$$\prod_{k=1}^{K} \max(\rho(y_k - f(p, x_k), \theta), \rho_0(y_k - f(p, x_k), \varphi))$$

attains the largest possible value.

**Finite-parametric information about $\rho(\Delta y)$, bound $W$ on the outlier-related differences $\Delta y_k$.** In such a situation, we determine the parameters $p$ and $\theta$ from the requirement that the likelihood

$$L = \prod_{k=1}^{K} \max\left(\rho(\Delta y_k, \theta), \frac{1}{2W}\right) =$$

$$\prod_{k=1}^{K} \max\left(\rho(y_k - f(p, x_k), \theta), \frac{1}{2W}\right)$$

attains the largest possible value under the constraint that

$$|\Delta y_k| = |y_k - f(p, x_k)| \le W$$

for all $k = 1, \ldots, K$.

**Finite-parametric information about $\rho(\Delta y)$, no information about the outlier-related differences $\Delta y_k$.** In this case, similarly to the above case when we had no information about the outlier-related differences $\Delta y_k$, we should select the smallest possible $W$, i.e., $W = \max\limits_{\ell} |\Delta y_\ell|$. Thus, we need to select the parameters $p$ and $\theta$ for which the likelihood

$$L = \prod_{k=1}^{K} \max\left(\rho(y_k - f(p, x_k), \theta), \frac{1}{2 \cdot \max\limits_{\ell} |y_\ell - f(p, x_\ell)|}\right)$$

attains the largest possible value.

**Bound $\Delta$ on the measurement errors, bound $W$ on the outlier-related differences $\Delta y_k$.** In this case, by using the maximum entropy approach, we select the following distributions:

- the measurement errors are uniformly distributed on the interval $[-\Delta, \Delta]$, with the probability density
$$\rho(\Delta y) = \frac{1}{2\Delta};$$
- the outlier-related differences $\Delta y_k$ are uniformly distributed on the interval $[-W, W]$, with the probability density $\rho_0(\Delta y) = \dfrac{1}{2W}$.

In this case, we need to select the parameters $p$ that maximize the likelihood $L = \prod\limits_{k=1}^{K} g(\Delta y)$, where

$$g(\Delta y) = \max(\rho(\Delta y), \rho_0(\Delta y)).$$

For the above uniform distributions, the auxiliary function $g(\Delta y)$ takes the following form:

- for the values $\Delta y$ for which $|\Delta y| \le \Delta$, we have
$$g(\Delta y) = \frac{1}{2\Delta};$$
- for the values $\Delta y$ for which $\Delta < |\Delta y| \le W$, we have $g(\Delta y) = \dfrac{1}{2W}$; and

- for the values $\Delta y$ for which $|\Delta y| > W$, we have $g(\Delta y) = 0$.

Thus, maximizing the product $L = \prod_{k=1} g(\Delta y_k)$ means minimizing the number of outliers under the constraint that $|\Delta y_k| = |y_k - f(p, x_k)| \le W$ for all $k$. In other words, we select $p$ for which, under the above constraints, the number of observations for which $|y_k - f(p, x_k)| > \Delta$ is the smallest possible.

**Bound $\Delta$ on the measurement errors, no information about the outlier-related differences $\Delta y_k$.** In this case, since we take $W = \max_\ell |y_\ell - f(p, x_\ell)|$, there are no longer any limitations on $p$.

Thus, in this case, the maximum likelihood method simply means selecting the values of the parameters $p$ for which the number of outliers (i.e., values for which $|y_k - f(p, x_k)| > \Delta$) is the smallest possible.

*Comment.* This idea has been actively used, as a heuristic idea, to deal with data processing under outliers, see, e.g., [3], [7], [10]. Several practical applications of this heuristic idea are described, e.g., in [3].

Our probability-based justification for this heuristics was first announces in [17] (see also [4]).

**Final case, when we have no information about the probabilities.** Finally, let us consider the case when we have no information about the probabilities, neither about the probabilities of different values of the measurement errors, nor about the probabilities of different outlier-related differences

$$\Delta y = y - f(p, x).$$

In this case, we need to select the corresponding bounds $\Delta$ and $W$ for which the corresponding likelihood function attains its largest possible value. Similar to the previous cases, for each parameter tuple $p$, the maximum of the likelihood $L$ is attained if we take $W(p) = \max_\ell |\Delta y_\ell|$, so it only remains to select $p$ and $\Delta$.

For each $p$ and $\Delta$, let us denote by $n(p, \Delta)$ the number of values $k$ for which $|y_k - f(p, x_k)| \le \Delta$. In terms of this notation, the desired likelihood value

$$L(p, \Delta) = \prod_{k=1}^{K} g(y_k - f(p, x_k))$$

has the form

$$L(p, \Delta) = \frac{1}{(2\Delta)^{n(p,\Delta)}} \cdot \frac{1}{(2W(p))^{K - n(p,\Delta)}},$$

i.e., equivalently, the form

$$L(p, \Delta) = \frac{1}{(2W(p))^K} \cdot \left(\frac{W(p)}{\Delta}\right)^{n(p,\Delta)}.$$

Maximizing this expression is equivalent to minimizing its minus logarithm

$$\psi(p, \Delta) = -\ln(L(p, \Delta)) =$$

$$K \cdot \ln(2W(p)) + n(p, \Delta) \cdot (\ln(\Delta) - \ln(W(p))).$$

Thus, to get the maximum likelihood, for each $p$, we need to select $\Delta$ for which the expression $\psi(p, \Delta)$ is the smallest possible. We then select the parameters for which the resulting minimum is the smallest possible, i.e., for which the following expression is the smallest possible:

$$\psi(p) = \min_\Delta (K \cdot \ln(2W(p)) + n(p, \Delta) \cdot (\ln(\Delta) - \ln(W(p)))),$$

where $W(p) = \max_\ell |y_\ell - f(p, x_\ell)|$ and

$$n(p, \Delta) = \#\{k : |y_k - f(p, x_k)| \le \Delta\}.$$

*Comment.* To check how well our method works, we have applied this idea to the situations when the values $\Delta y_k$ are distributed according to several reasonable distributions: normal, heavy-tailed power law, etc.

In all these cases, we get 5-20% values classified as outliers. This is in line with the usual case of normal distribution, where 5% of the values lie outside the $2\sigma$ interval and are, thus, usually dismissed as outliers,

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grants CAREER 0953339, HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation. This research was performed during Anthony Welte's visit to the University of Texas at El Paso.

The authors are thankful to all the participants of the Summer Workshop on Interval Methods SWIM'2016 (Lyon, France, June 19–22, 2016) for valuable discussions, and to the anonymous referees for useful suggestions.

## REFERENCES

[1] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Verlag, New York, 1985.

[2] B. Chokr and V. Kreinovich, "Ho far are we from complete knowledge: complexity of knowledge acquisition in Demster-Shafer approach", In: R. R. Yager, J. Kacprzyk, and M. Pedrizzi (eds.), *Advances in the Dempster-Shafer Theory of Evidence*, Wiley, New York, 1994, pp. 555–576.

[3] B. Desrochers, S. Lacroix, and L. Jailin, "Set-membership approach to the kidnapped robot problem", *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems IROS'2015*, Hamburg, Germany, September 28 – October 2, 2015, pp. 3715–3720.

[4] V. Drevelle and P. Bonnifait, "A set-membership approach for high integrity height-added satellite positioning", *GPS Solutions*, 2011, Vol. 15, No. 4, pp. 357–368.

[5] L. Jaulin, "Reliable minimax parameter estimation", *Reliable Computing*, 2001, Vol. 7, No. 3, pp. 231–246.

[6] L. Jaulin, M. Kiefer, O. Dicrit, and E. Walter, *Applied Interval Analysis*, Springer, London, 2001.

[7] L. Jaulin and E. Walter, "Guaranteed robust nonlinear mimimax estimation", *IEEE Transactions on Automatic Control*, 2002, Vol. 47, No. 11, pp. 1857–1864.

[8] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[9] V. Kreinovich and S. Shary, "Interval Methods for Data Fitting under Uncertainty: A Probabilistic Treatment", *Reliable Computing*, 2016, Vol. 23, pp. 105–141.

[10] H. Lahanier, E. Walter, and R. Gomeni, "OMNE: a new robust membership-set estimator for the parameters of nonlinear modles", *Journal of Pharmacokinetics and Biopharmacutics*, 1987, Vol. 15, No. 2, pp. 203–219.

[11] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, New York, 2003.

[12] R. McKendall, *Minimax Estimation of a Discrete Location Parameter for a Continuous Distribution*, PhD Dissertation, Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, 1990; available as Teachnical Report MS-CIS-90-28, Computer and Information Science Department, University of Pennsylvania, 1990.

[13] R. McKendall and M. Mintz, "Robust sensor fusion with statistical decision theory", In: M. A. Abidi ad R. C. Gonzalez (eds.), *Data Fusion in Robotics and Machine Intelligence*, Academic Press, Boston, Massachusetts, 1992, pp. 211–244.

[14] R. B. Millar, *Maximum Likelihood Estimation and Inference*, Wiley, Chichester, UK. 2011.

[15] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[16] H. T. Nguyen, V. Kerinovich, B. Wu, and G. Xiang, *Computring Statistics under Intervala nd Fuzzy Uncertainty*, Springer Verlag, Berlin, 2012.

[17] J. Nicola and L. Jaulin, "OMNE is a Maximum Likelihood estimator", *Abstracts of the Summer Workshop on Interval Methods SWIM'2016*, Lyon, France, June 19–22, 2016.

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

[19] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[20] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, New York, 2009.

[21] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*, Springer, London, 1997.

[22] G. A. Watson, "The minimax solution of an overdetermined system of nonlinear equations", *IMA Journal of Applied Mathematics*, 1979, Vol. 23, No. 2, pp. 167–180.

[23] M. A. Wolfe, "On discrete mimimax problems in R using intervak arithmetic", *Reliable Computing*, 1999, Vol. 5, No. 4, pp. 371–383.