Probabilistic set-membership approach for robust regression

Luc Jaulin

ENSIETA, DTN, 2 rue François Verny,

29806 Brest Cédex 09.

Email : jaulinlu@ensieta.fr,

Web: www.ensieta.fr/jaulin/

Abstract

Interval constraint propagation methods have been shown to be efficient and reliable to solve difficult nonlinear bounded-error estimation problems. However they are considered as unsuitable in a probabilistic context, where the approximation of a probability density function by a set cannot be accepted as reliable. This paper shows how probabilistic estimation problems can be transformed into a set estimation problem by assuming that some rare events will never happen. Since the probability of occurrence of those rare events can be computed, we can give some prior lower bounds for the probability associated to solution set of the corresponding set estimation problem. The approach will be illustrated on a parameter estimation problem.

Index Terms

Interval analysis, probability, robust regression, set-membership estimation.

I. INTRODUCTION

Parameter set estimation deals with characterizing a (preferably small) set which encloses the parameter vector $\mathbf{p}$ of a model from data collected on the system [18]. In the context of bounded-error estimation, the measurement error is assumed to be bounded and characterizing the posterior feasible set amounts to solve a set inversion problem for which interval constraint propagation [14], [17] methods have been shown to be particularly efficient, even when the model is nonlinear [11]. In a probabilistic context, the measurement error is not anymore described by membership intervals, but by *probability density functions* (pdf). When some prior pdf for is available, the Bayes rule makes possible to obtain the posterior pdf (see, *e.g.*, [7]). The set to be estimated becomes the *minimal volume credible set* [2] and corresponds to the minimal volume set enclosing the associate random vector with a given probability. Unfortunately, this problem cannot be cast into a set inversion problem and interval methods, which can still be useful to characterize credible sets [10], are limited to small dimensional problems with few data.

Here, the approach that will be considered is different. We do not assume that some prior pdf are available for the vector to be estimated. Instead, we fix a given probability, say α, which corresponds to the risk we accept to take a priori. In practice, α is a tiny positive number. Then we choose a collection of rare events for the error such that the prior probability of occurrence of one of these events if lower than α. Finally we solve the associated set inversion problem using a set membership approach. To my knowledge, this approach has never been proposed before.

Interval methods have already been combined with probabilistic theory [12] [6] in order to solve estimation problems [1] [13]. The main difference between our approach and the above mentioned papers is that here, we solve a traditional probabilistic estimation problem using interval tools and thus our approach is fully consistent with traditional probabilistic estimation.

Section II presents the principle of our new approach for probabilistic estimation that we named *probabilistic-*

set estimation. The approach is then used in Section III to deal with robust nonlinear parameter estimation. The principle of set inversion and interval analysis for set estimation, needed to understand the resolution, is recalled on Section IV. An illustrative test case is treated in Section V. Section VI then concludes the paper.

II. Probabilistic-set approach for estimation

A parameter estimation problem can be represented by an error model equation

$$\mathbf{e} = \mathbf{f}\left(\mathbf{y}, \mathbf{p}\right) = \mathbf{f}_{\mathbf{y}}\left(\mathbf{p}\right), \tag{1}$$

where

- $\mathbf{e} \in \mathbb{R}^m$ is the error vector,
- $\mathbf{y} \in \mathbb{R}^m$ is the collected data vector, with the same dimension than $\mathbf{e}$, which is assumed to be known exactly a posteriori, and
- $\mathbf{p} \in \mathbb{R}^n$ is the parameter vector to be estimated.

A parameter estimation problem amounts to find $\mathbf{p}$ from $\mathbf{y}$ and some assumptions on the error $\mathbf{e}$. The classical approaches for parameter estimation are the Bayesian estimation and the set estimation that will now be recalled. In a probabilistic approach, we generally assume that a prior probability distributions $\Pi_{\mathbf{e}}$,$\Pi_{\mathbf{p}}^{\text{prior}}$ are known for $\mathbf{e}, \mathbf{p}$. The Bayes rule (see *e.g.* [7], [18], [9]) gives us the posterior probability distribution function for $\mathbf{p}$

$$\Pi_{\mathbf{p}}^{\text{post}}(\mathbf{p}) = \frac{\Pi_{\mathbf{e}}(\mathbf{f}_{\mathbf{y}}\left(\mathbf{p}\right)).\Pi_{\mathbf{p}}^{\text{prior}}(\mathbf{p})}{\int_{\mathbf{p}\in\mathbb{R}^n} \Pi_{\mathbf{e}}(\mathbf{f}_{\mathbf{y}}\left(\mathbf{p}\right)).\Pi_{\mathbf{p}}^{\text{prior}}(\mathbf{p}).d\mathbf{p}}. \tag{2}$$

as illustrated by Figure 1.

In a set-membership context (see Figure 2), we generally assume that the error vector $\mathbf{e}$ belongs to a prior set $\mathbb{E}$ which corresponds to all $\mathbf{e}$ such that $\Pi_{\mathbf{e}}(\mathbf{e}) \neq 0$, or equivalently $\mathbb{E}$ is the support of the random vector $\mathbf{e}$. The following theorem provides the link between the Bayesian and the set approaches. Even if its use is classical in the set-membership community, it is given here for the first time, to my knowledge.

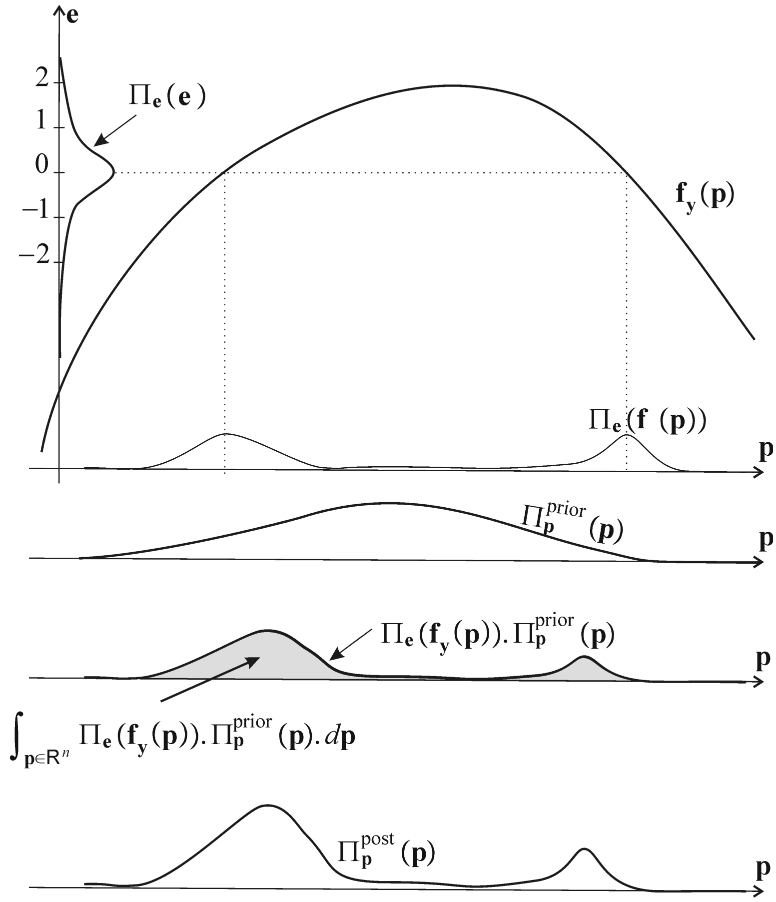

Fig. 1. Principle of the Bayesian estimation

Theorem. If $\mathbb{E}$ is the support of the error $\mathbf{e}$ and $\mathbb{P}$ support of the prior $\mathbf{p}$, then the support of the posterior $\mathbf{p}$ is given by

$$\widehat{\mathbb{P}} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E}) \cap \mathbb{P}. \tag{3}$$

The set $\widehat{\mathbb{P}}$ is called the *posterior feasible set*.

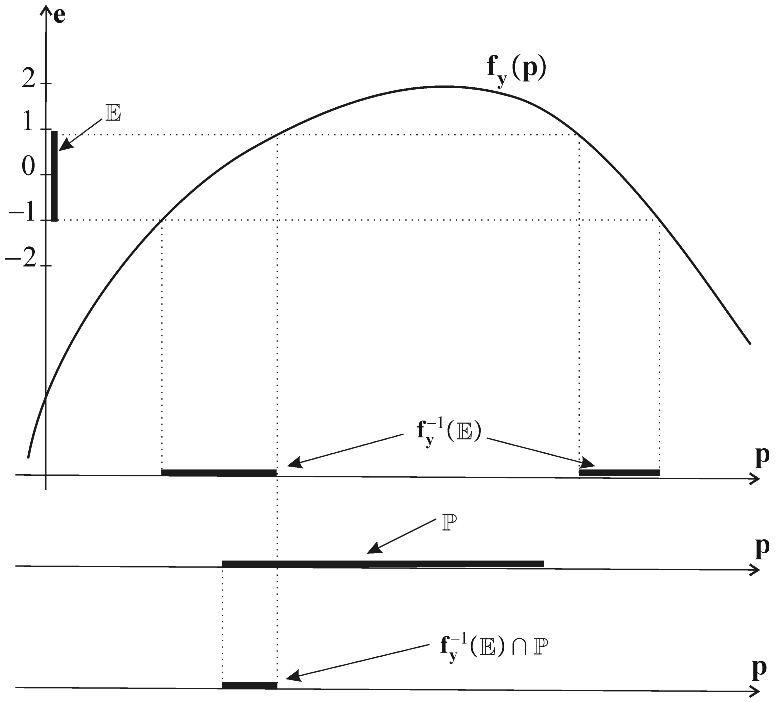

Fig. 2. Illustration of the set membership approach

Proof. Equation (3) can easily be proven from the Bayes rule by taking the supports of the probability distributions that are involved. From (2), we have

$$\begin{aligned} \Pi_{\mathbf{p}}^{\text{post}}(\mathbf{p}) \neq 0 \quad &\Leftrightarrow \quad \Pi_{\mathbf{e}}(\mathbf{f_y}(\mathbf{p})) \neq 0 \text{ and } \Pi_{\mathbf{p}}^{\text{prior}}(\mathbf{p}) \neq 0 \\ &\Leftrightarrow \quad \mathbf{f_y}(\mathbf{p}) \in \Pi_{\mathbf{e}}^{-1}(]0,\infty[) \text{ and } \mathbf{p} \in \mathbb{P} \\ &\Leftrightarrow \quad \mathbf{p} \in \mathbf{f_y}^{-1}(\mathbb{E}) \text{ and } \mathbf{p} \in \mathbb{P} \\ &\Leftrightarrow \quad \mathbf{p} \in \mathbf{f_y}^{-1}(\mathbb{E}) \cap \mathbb{P}. \end{aligned}$$

■

From this theorem, we can conclude that the set-membership approach can be seen as less general than the probabilistic approach. However, characterizing $\widehat{\mathbb{P}}$ is a set inversion problem which can be solved efficiently and in a reliable way for a huge class of nonlinear functions $\mathbf{f}$. This is not the case for probabilistic approaches except for some specific situations, such as the linear-Gaussian case. The most efficient techniques to solve the set inversion problem are based on interval constraint propagation methods.

In this paper, contrary to other set-membership techniques, the support of the pdf of e is not assumed to be bounded. The principle of the approach is to decompose the error space into two subsets. A subset $\mathbb{E}$ on which we bet that the error vector $\mathbf{e}$ will belong and its complementary set $\overline{\mathbb{E}}$. The prior probability of the event $\mathbf{e} \in \mathbb{E}$ is denoted by π. The set $\mathbb{E}$ is chosen such that π is almost equal to 1 and such that the likelihood is high over $\mathbb{E}$. The event $\mathbf{e} \in \overline{\mathbb{E}}$ is considered as rare and we bet that it will not occur. Once the data vector $\mathbf{y}$ is collected, we compute the *posterior* feasible set

$$\widehat{\mathbb{P}} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E}). \tag{4}$$

If now $\widehat{\mathbb{P}}$ is not empty, then we still bet that the rare event did not occur and we conclude that $\mathbf{p} \in \widehat{\mathbb{P}}$ with a probability of π. If $\widehat{\mathbb{P}} = \emptyset$, then we conclude the rare event occurred. In practice, we might also suspect the model is not accurate enough, or that outliers have corrupted the data set. We shall now give academic examples to illustrate the principle of our approach.

Example 1. Assume that the model is described by the relation $y = p^2 + e$, *i.e.*, $e = y - p^2 = f_y(p)$ and that Π_e is a normal distribution of the form $\mathcal{N}(0,1)$. If we set $\mathbb{E} = [-6, 6]$ then, e will be outside $\mathbb{E}$, with a prior probability of

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-6}^{6} \exp\left(-\frac{e^2}{2}\right) de \simeq 1.97 \times 10^{-9}. \tag{5}$$

Assume that we collect the data $y = 10$. With the probabilistic-set approach, we get

$$\begin{aligned} \widehat{\mathbb{P}} &= f_y^{-1}(\mathbb{E}) = f_y^{-1}([-6,6]) \\ &= \sqrt{y - [e]} = \sqrt{10 - [-6,6]} \\ &= \sqrt{[4,16]} = [-4,-2] \cup [2,4]. \end{aligned} \tag{6}$$

with a prior probability of $1 - 1.97 \times 10^{-9}$. To apply Bayesian approach, we need a prior density probability $\Pi_p^{\text{prior}}(p)$ for the parameter p. Assume that it is given by a normal distribution $\mathcal{N}(3, 1)$. The posterior pdf for p is

$$\begin{aligned} \Pi_p^{\text{post}}(p) &= \frac{\Pi_e(f_y(p)).\Pi_p^{\text{prior}}(p)}{\int_{p\in\mathbb{R}} \Pi_e(f_y(p)).\Pi_p^{\text{prior}}(p)dp} \\ &= \frac{e^{-\frac{(10-p^2)^2}{2}}.e^{-\frac{(p-3)^2}{2}}}{\int_{-\infty}^{\infty} e^{-\frac{(10-p^2)^2}{2}}.e^{-\frac{(p-3)^2}{2}}.dp} \\ &\simeq 2.57\ e^{-\frac{p^4-19p^2-6p+109}{2}}. \end{aligned} \tag{7}$$

This illustrates that the probabilistic-set approach is more easy to perform than applying the Bayes rule, in a nonlinear context. ■

Example 2. Consider the same example, but in a situation where the rare event "$e \notin \mathbb{E}$" occurs. For instance, we measure $y = -10$. The probabilistic-set approach gives

$$\widehat{\mathbb{P}} = f_y^{-1}(\mathbb{E}) = \emptyset, \tag{8}$$

whereas

$$\begin{aligned} \Pi_p^{\text{post}}(p) &= \frac{e^{-\frac{(20-p^2)^2}{2}}.e^{-\frac{(p-3)^2}{2}}}{\int_{-\infty}^{\infty} e^{-\frac{(10-p^2)^2}{2}}.e^{-\frac{(p-3)^2}{2}}dp} \\ &\simeq 6.930\,5 \times 10^{23}.e^{-\frac{p^4-39p^2-6p+409}{2}}. \end{aligned} \tag{9}$$

The probabilistic-set approach concludes to an inconsistency, whereas the Bayesian approach yields to precise posterior pdf for p. Now, in practice, the huge factor (here $6.930\,5 \times 10^{23}$) is often interpreted as an inconsistency. ■

Example 3. The probabilistic-set approach can easily be extended to the situation where the error is a vector. Figure 3 illustrates a situation where $\mathbf{e}$ is of dimension 2. Assume that

$$\begin{array}{llll} \Pr(e_1 \leq -1) & = 0.2, & \Pr(e_2 \leq -2) & = 0.2, \\ \Pr(e_1 \in [-1, 1]) & = 0.4, & \Pr(e_2 \in [-2, 3]) & = 0.6, \\ \Pr(e_1 \in [1, 2]) & = 0.2, & \Pr(e_2 \geq 3) & = 0.2, \\ \Pr(e_1 \geq 2) & = 0.2, & & \end{array}$$

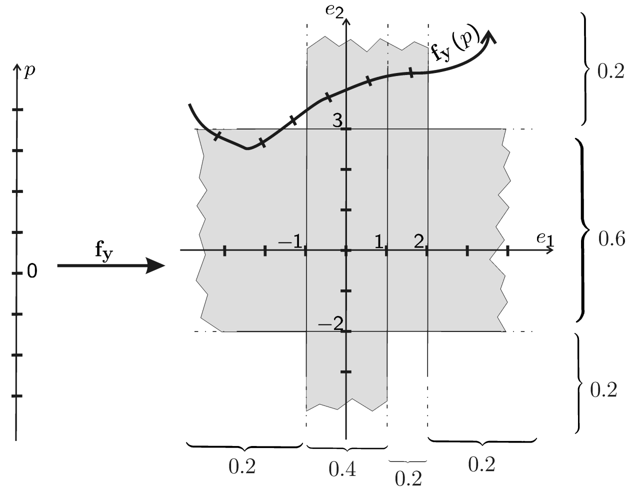

Fig. 3. Illustration of the probabilistic-set approach

and that e_1 and e_2 are independent. Then, the joint pdf for (e_1, e_2) satisfies the probabilities given on the following table.

$[e_2] \backslash [e_1]$	$[-\infty, -1]$	$[-1, 1]$	$[1, 2]$	$[2, \infty]$
$[3, \infty]$	0.04	**0.08**	**0.04**	0.04
$[-2, 3]$	**0.12**	**0.24**	**0.12**	**0.12**
$[-\infty, -2]$	0.04	**0.08**	0.04	0.04

As a consequence, the prior probability for $\mathbf{e}$ to belong to the grey set

$$\mathbb{E} = (\mathbb{R} \times [-2, 3]) \ \cup \ ([-1, 1] \times \mathbb{R}) \ \cup \ ([1, 2] \times [3, \infty[)$$

is given by

$$0.08 + 0.04 + 0.12 + 0.24 + 0.12 + 0.12 + 0.08 = 0.8.$$

If we assume that $\mathbf{e} \notin \overline{\mathbb{E}}$ is a are event (here the probability is 0.2), the posterior feasible set will be $\widehat{\mathbb{P}} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E})$ with a prior probability of 0.8. ■

Remark 1: Representing the pdf Π_e for the error by boxes with an associated probability, as illustrated by Figure 3, can be interpreted as a discretization of Π_e. The resulting object can be represented via *potential clouds* [15],

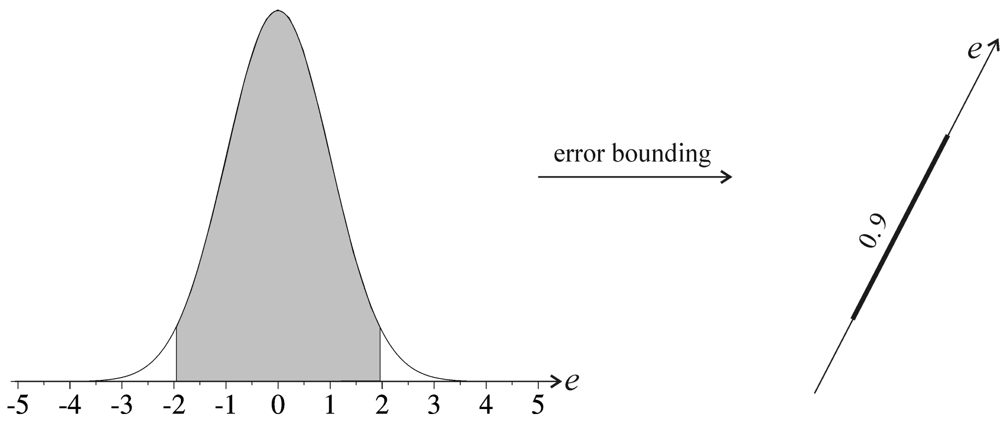

Fig. 4. Error bounding for a gaussian univariate probability function

p-boxes [3] or Dempster-Shafer structures [4], [5]. However, such abstractions will not be needed here and we limit ourselves to the use of classical probabilistic theory. ■

From $\Pi_{\mathbf{e}}$, we can find some feasible set $\mathbb{E}$ for $\mathbf{e}$ associated with a given probability. The procedure which provides $\mathbb{E}$ from $\mathbf{e}$ will be called *error bounding.* For instance, for a normal random variable, it is natural to take the confidence interval (see Figure 4) for $\mathbb{E}$.

For a normal random vector, we have different choices for the error bounding. As illustrated by Figure 5 for a two-dimensional random vector, we can take a disk (which is a confidence region), a square (good for interval methods), or a cross (more robust with respect to outliers and also suited to interval algorithms).

When outliers occur, the errors should not be represented by a Gaussian distribution, but by heavy tailed distributions or by sums of Gaussian distributions. For such distributions, the confidence regions are non convex and should not be represented by boxes or disks. Figure 6 illustrates the error bounding for a random vector $\mathbf{e} = (e_1, e_2)$ where e_1 and e_2 are both independent and satisfy a probability distribution composed by the sum of two Gaussian distributions.

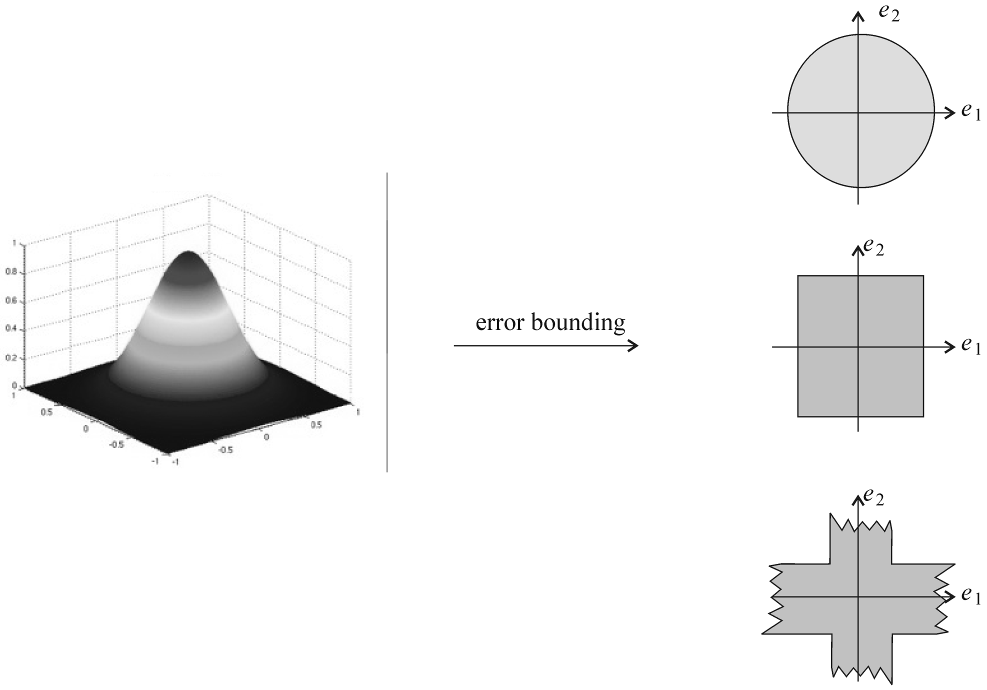

Fig. 5. Error bounding for a normal probability distribution function

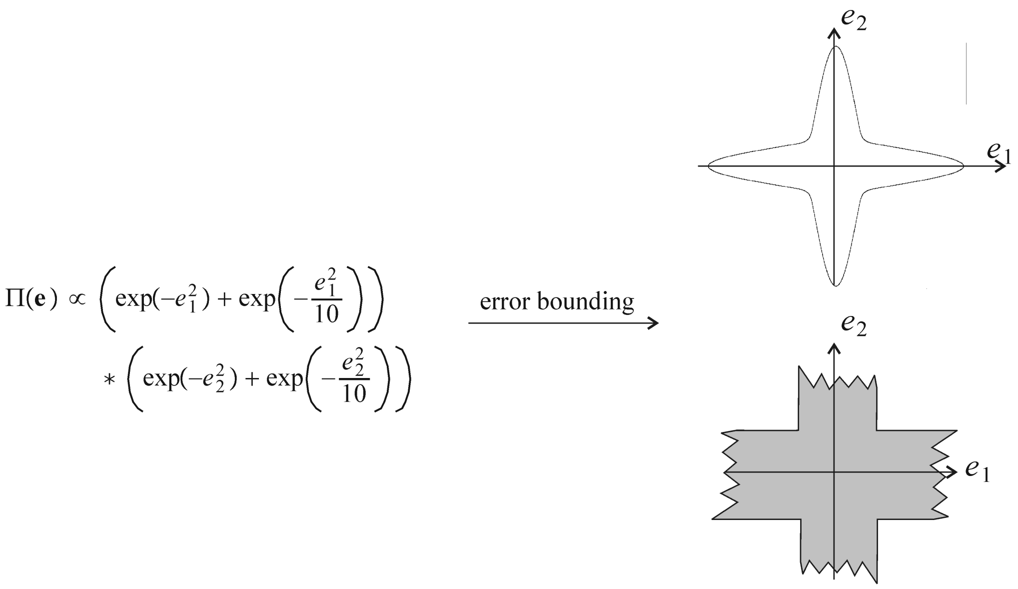

Fig. 6. Error bounding for a white random vector $\mathbf{e}$ the components e_i of which follow a gaussian sum distribution.

III. ROBUST REGRESSION

Robust regression is designed to circumvent some limitations of traditional parametric methods when outliers occur in the set of data [16]. In this section, we will show how the probabilistic-set approach can be used for estimation problems where outliers are involved. Consider again the error model

$$\mathbf{e} = \mathbf{f}_{\mathbf{y}}\left(\mathbf{p}\right). \tag{10}$$

We shall assume that $\mathbf{e}$ is white (i.e., all its components $e_i, i \in \{1, \ldots, m\}$ are independent and identically distributed) Assume that the prior probability for e_i to belong to the interval $[-e_{\max}, e_{\max}]$ is equal to π. The ith component of $\mathbf{e}$ is said to be an *inlier* if $e_i \in [-e_{\max}, e_{\max}]$ and an outlier otherwise. Thus the probability for e_i to be an inlier is π (it does not depend on i). The number k of inliers follows a binomial distribution. Therefore, the probability of having exactly k inliers among m is

$$\beta(k, m, \pi) = \frac{m!}{k!\,(m-k)!}\pi^k.\,(1-\pi)^{m-k}. \tag{11}$$

As a consequence, the probability of having strictly more than q outliers is

$$\gamma\left(q, m, \pi\right) \stackrel{\text{def}}{=} \sum_{k=0}^{m-q-1} \beta(k, m, \pi). \tag{12}$$

Now, the mean for the binomial distribution with m trials is $m.\pi$ and its standard deviation is $\sqrt{m\pi(1-\pi)}$. If m, is large, the binomial distribution is approximately equal to the normal distribution $\mathcal{N}\left(m.\pi, \sqrt{m\pi(1-\pi)}\right)$. The probability that there exists more than q outliers is thus

$$\gamma\left(q, m, \pi\right) \simeq \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{m\left(1-\pi\right) - q - 1}{\sqrt{2m\pi(1-\pi)}}\right)\right). \tag{13}$$

Example. For instance, if $m = 1000$, $q = 900, \pi = 0.2$, we get $\gamma\left(q, m, \pi\right) = 7.04 \times 10^{-16}$. Thus, if 80% of the e_i are outlier, having more than 900 outliers over 1000 data can be seen as a rare event. The graph of the function $\gamma\left(q, 1000, 0.2\right)$ is represented on Figure 7. ■

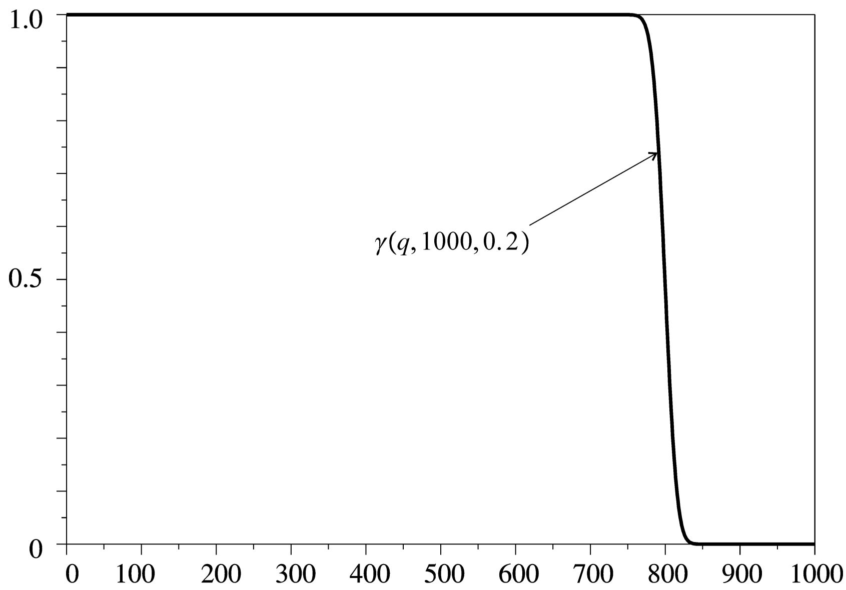

Fig. 7. Graph of the function $\gamma(q, 1000, 0.2)$ which provides the probability of having more than q ouliers

The robust regression approach amounts to choosing the set $\mathbb{E}$ of all $\mathbf{e} \in \mathbb{R}^m$ such that the number of outliers is smaller (or equal) than q, i.e.,

$$\mathbb{E} = \{\mathbf{e} \in \mathbb{R}^m, \text{ card}\{i, e_i \notin [-e_{\max}, e_{\max}]\} \leq q\}. \tag{14}$$

The posterior feasible set $\widehat{\mathbb{P}} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E})$ will thus contain the true parameter vector with a prior probability of $1 - \gamma(q, m, \pi)$.

IV. SET INVERSION WITH INTERVAL ANALYSIS

This section presents the basic notion needed to understand how the set inversion problem

$$\widehat{\mathbb{P}} = \{\mathbf{p} \mid \mathbf{f}_{\mathbf{y}}(\mathbf{p}) \in \mathbb{E}\} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E}) \tag{15}$$

can be solved with interval analysis. Interval arithmetic is a numerical tool originally developed in order to quantify the effect of finite-precision arithmetic on results obtained by a computer [14]. Interval arithmetic extends classical operators on real numbers to intervals in a natural way. Thus, if $[x] = [x^-, x^+]$ and $[y] = [y^-, y^+]$,

$$\begin{aligned} [x] + [y] &= [x^- + y^-, x^+ + y^+] \\ [x] - [y] &= [x^- - y^+, x^+ - y^-] \\ [x] \cdot [y] &= [\min(x^- y^-, x^- y^+, x^+ y^-, x^+ y^+), \\ &\quad \max(x^- y^-, x^- y^+, x^+ y^-, x^+ y^+)]. \end{aligned} \tag{16}$$

For example, we have $([1,2] + [-3,4]) \cdot [-1,5] = [-2,6] \cdot [-1,5] = [-10,30]$. As another example, let us consider the real function $f(x) = x^2 + 2x + 4$. An interval evaluation for f is $[f]([x]) = [x] \cdot [x] + 2[x] + 4$. For $[x] = [-3,4]$, we have:

$$[f]([-3,4]) = [-3,4] \cdot [-3,4] + 2[-3,4] + 4 \tag{17}$$

$$= [-12,16] + [-6,8] + 4 = [-14,28]. \tag{18}$$

Note that the actual image by f of the interval $[x]$, $f([-3,4]) = [3,28]$ is a subset of the interval evaluation $[f]([-3,4]) = [-14,28]$. This illustrates that interval evaluation is usually pessimistic [14]. A *box* or *vector interval* $[\mathbf{p}]$ of $\mathbb{R}^n$ is defined as the Cartesian product of n intervals.

$$[\mathbf{p}] = \left[p_1^-, p_1^+\right] \times \cdots \times \left[p_n^-, p_n^+\right]. \tag{19}$$

It can be proven (see [14]) that the interval evaluation $[\mathbf{f_y}]([\mathbf{p}])$, applied on a given expression of $\mathbf{f_y}$, encloses the set $\mathbf{f_y}([\mathbf{p}])$, i.e.

$$\forall [\mathbf{p}], \mathbf{f_y}([\mathbf{p}]) \subset [\mathbf{f_y}]([\mathbf{p}]). \tag{20}$$

The *width* $w([\mathbf{p}])$ of a box $[\mathbf{p}]$ is the size of its largest side. For instance, the width of the box $[\mathbf{p}] = [1,2] \times [-1,3]$ is equal to 4. A *principal plane* of $[\mathbf{p}]$ is a symmetry plane of $[\mathbf{x}]$ normal to a side of maximum length. To *bisect* a box $[\mathbf{p}]$ means to cut it along one of its principal planes. Bisecting $[\mathbf{p}] = [1,2] \times [-1,3]$ produces two boxes $[\mathbf{p}](1) = [1,2] \times [-1,1]$ and $[\mathbf{p}](2) = [1,2] \times [1,3]$. The algorithm SIVIA (Set Inverter Via Interval Analysis) computes a list of boxes the union of which encloses the solution set $\widehat{\mathbb{P}}$. Note that SIVIA can also compute some boxes that are proved to be inside $\widehat{\mathbb{P}}$, but here, for the sake of simplicity, a simpler version of SIVIA is given.

SIVIA uses the following test to decide if a given box $[\mathbf{p}]$ is outside the solution set $\widehat{\mathbb{P}}$:

$$[\mathbf{f_y}]([\mathbf{p}]) \cap \mathbb{E} = \emptyset \;\Rightarrow\; [\mathbf{p}] \cap \widehat{\mathbb{P}} = \emptyset. \tag{21}$$

SIVIA is given by the following table where the accuracy ε is a small positive real number and $\mathcal{P}^{+}$ is a list of boxes which is initially empty.

Algorithm SIVIA($[\mathbf{p}]$)
if $[\mathbf{f_y}]([\mathbf{p}]) \cap \mathbb{E} = \emptyset$, then return;
If $w([\mathbf{p}]) < \varepsilon$, {store $[\mathbf{p}]$ into $\mathcal{P}^{+}$; return};
Bisect $[\mathbf{p}]$ into $[\mathbf{p}](1)$ and $[\mathbf{p}](2)$;
SIVIA$([\mathbf{p}](1))$; SIVIA$([\mathbf{p}](2))$;

SIVIA is first called for $[\mathbf{p}] = [\mathbf{p}](0)$, where $[\mathbf{p}](0)$ is a box which contains the solution set $\widehat{\mathbb{P}}$. After completion of the algorithm, we have $\widehat{\mathbb{P}} \subset \mathcal{P}^{+}$.

Remark 2: When the box $\mathbb{E}$ to be inverted is a singleton $\{\mathbf{e}\}$ (for example when dealing with error-free data), the solution set $\widehat{\mathbb{P}}$ is often reduced to a singleton $\{\mathbf{p}\}$ which is easily found by SIVIA or by other punctual approaches. When two or more solutions exist, SIVIA detects all of them in a guaranteed way, contrary to other punctual approaches.

Remark 3: The algorithm can be made much more efficient by using interval constraint propagation, which is not described here.

V. TEST CASE

Consider a set of $m = 500$ data generated as follows

$$\begin{aligned} y_i &= p_1 \sin(p_2 t_i) + e_i, && \text{with a probability } 0.2 \\ &= r_1 \exp(r_2 t_i) + e_i, && \text{with a probability } 0.2 \\ &= n_i && \text{with a probability } 0.6 \end{aligned}$$

where $t_i = 0.02 * (i+1)$, $i \in \{1, 500\}$, e_i is a white signal uniformly distributed inside $[-0.1, 0.1]$ and n_i is a white signal normally distributed with a mean of 2 and a standard deviation of 3. For the generation of the simulated data, we took $\mathbf{p} = (2,2)^{\mathrm{T}}$ and $\mathbf{r} = (4, -0.4)^{\mathrm{T}}$. The set of data is depicted on Figure 8(a). Now, assume that we only know that $y_i = p_1 \sin(p_2 t_i) + e_i$, with a probability 0.2 and that we have no idea of what happen otherwise. Let us now try to estimate p_1 and p_2 from the data. A priori, we know that the probability of getting an outlier is lower than 0.8. Assume that we want to compute a feasible set for $\mathbf{p}$ which encloses the true value for $\mathbf{p}$ with a probability greater than 0.95. We should take

$$q = \min \{q_1 \mid \gamma(q_1, 500, 0.2) < 1 - 0.95\}.$$

Recall that $\gamma(q_1, 500, 0.2)$ represents the probability of having more than q_1 outliers. Since γ is decreasing with respect to q, we easily get $q = 414$, via a dichotomy method. We will thus assume a minimum of $q = 414$ outliers. The resulting feasible set $\widehat{\mathbb{P}}$ (see Figure 8(c) in black inside the two circles) encloses $\mathbf{p}$ with a prior probability greater than $1 - \gamma(414, 500, 0.2) = 0.953$. The two connected components of $\widehat{\mathbb{P}}$ are due to the fact that the model is not globally identifiable. The grey segments of Figure 8(c) represent the boxes generated by the interval set inversion algorithm. The associated filtered data are represented on 8(b), in black. The resulting computing time is less than 5 seconds.

Comparison with a Monte-Carlo approach. The interval approach has been compared with a basic Monte-Carlo method which can also be used to solve set inversion problems. For the same problem, with the same data and the same number of assumed outliers (i.e. $q = 414$), the Monte-Carlo method was not able to find a single feasible

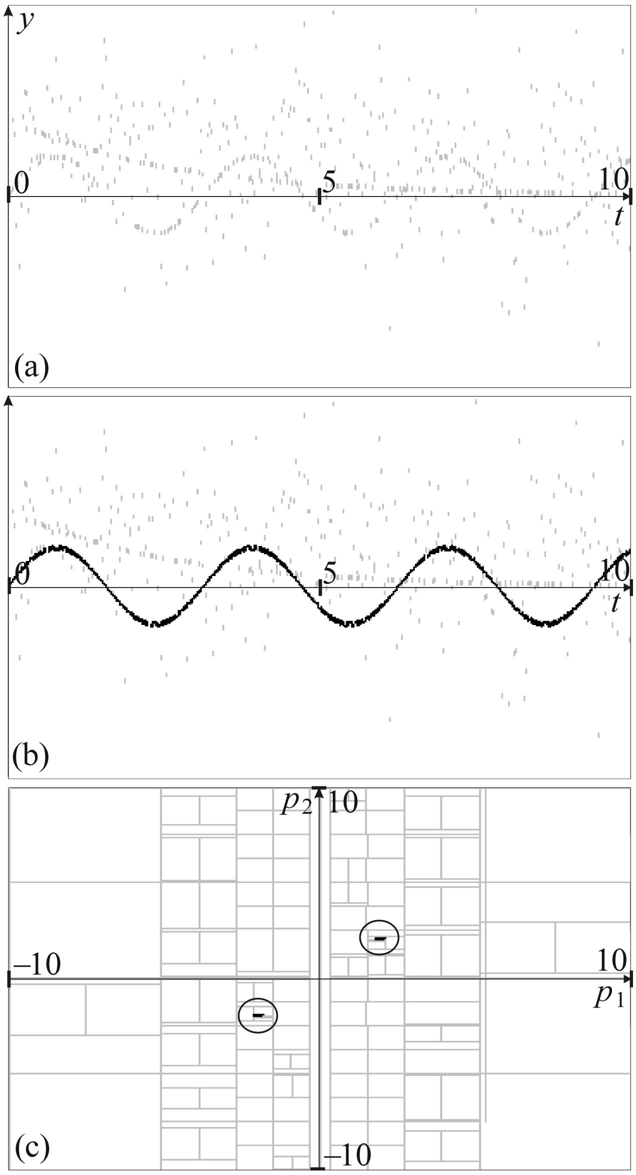

Fig. 8. (a) available data, (b) filtered data, (c) the feasible set in black inside the two circles has two connected components .

point, after ten minutes. To obtain feasible points, the number of allowed outliers was increased up to $q = 460$ for 500 data. The results obtained for 100000 test points are shown on Figure 9. The number of feasible points that have been found around the true value for $\mathbf{p}$ is small and the estimation is poor.

The C++ Builder 5 source codes including the interval and the Monte-Carlo algorithm corresponding the test case as well as the data are available at the following link

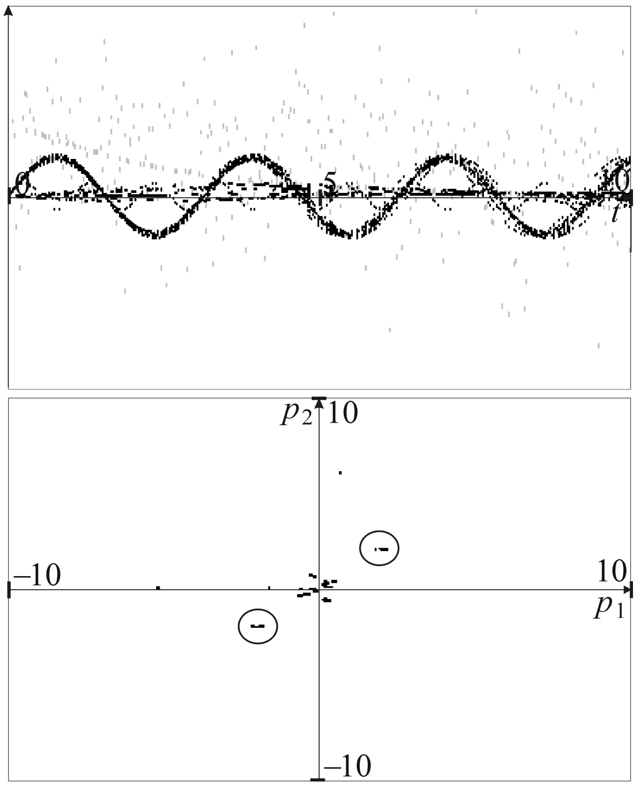

Fig. 9. Results obtained by the Monte-Carlo method

`www.ensieta.fr/jaulin/probint0.html`

VI. Conclusion

In this paper, we have presented a new approach for identification which combines interval propagation methods with a probabilistic representation of uncertainty. The main idea is to transform a probabilistic problem into a set inversion problem. It made possible to solve nonlinear probabilistic estimation problems in a robust and reliable way. The main contribution with respect to other classical interval-based methods is that our approach is able to provide a probability associated to computed set. Contrary to other robust Monte-Carlo based methods (such as the Ransac algorithm [8] widely used in computer vision) our algorithm is deterministic and provides guaranteed results if the assumptions are satisfied (in the context of the paper, the probability of having the assumptions satisfied was assumed to be known).

A possible extension of the approach is to make the prior error feasible set $\mathbb{E}$ depend on a parameter $\alpha \in [0,1]$, where α represents the probability for the error vector $\mathbf{e}$ to belong to $\mathbb{E}$. The resulting parametrized set $\mathbb{E}(\alpha)$ can then be interpreted as a *cloud* [15] and can easily be represented through *potential clouds*. The proposed approach can then be extended to compute a posterior feasible cloud $\widehat{\mathbb{P}}(\alpha)$ for $\mathbf{p}$.

REFERENCES

[1] F. Abdallah, A. Gning, and P. Bonnifait. Box particle filtering for nonlinear state estimation using interval analysis. *Automatica*, 44(3):807–815, 2008.

[2] J. Berger. *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer-Verlag, New York, NY, 1985.

[3] D. Berleant, L. Xie, and J. Zhang. Statool: a tool for distribution envelope determination (denv), an interval-based algorithm for arithmetic on random variables. *Reliable Computing*, 9(2):91–108, 2003.

[4] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[5] T. Denoeux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.

[6] D. Dubois and H. Prade. Random sets and fuzzy interval analysis. *Fuzzy Sets and Systems*, 42(1):87–101, 1976.

[7] P. Eykhoff. *System Identification, Parameter and State Estimation*. John Wiley, London, 1974.

[8] D. Forsyth and J. Ponce. *Computer Vision, a modern approach*. Prentice Hall, 2003.

[9] J. Idier, editor. *Bayesian Approach to Inverse Problems*. ISTE Ltd and John Wiley & Sons Inc, avr. 2008.

[10] L. Jaulin. Computing minimal-volume credible sets using interval analysis; application to bayesian estimation. *IEEE Trans. on Signal Processing*, 54(9):3632–3636, 2006.

[11] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer-Verlag, London, 2001.

[12] V. Kreinovich, G. Dimuro, and A. C. da Rocha Costa. Probabilities, intervals, what next ? extension of interval computations to situations with partial information about probabilities. In *10th IMEKO TC7 International symposium*, 2004.

[13] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg. Outlier detection under interval uncertainty: Algorithmic solvability and computational complexity. In I. Lirkov, S. Margenov, J. Wasniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing*, Proceedings of the 4th International Conference LSSC'2003, 2003.

[14] R. E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, PA, 1979.

[15] A. Neumaier. Clouds, fuzzy sets and probability intervals. *Reliable Computing*, 10(4):249Ű–272, 2004.

[16] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[17] M. van Emden. Algorithmic power from declarative use of redundant constraints. *Constraints*, 4(4):363–381, 1999.

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