Guaranteed characterization of exact non-asymptotic confidence regions in nonlinear parameter estimation

Michel Kieffer, Eric Walter

To cite this version:


HAL Id: hal-00819488
https://hal-supelec.archives-ouvertes.fr/hal-00819488

Submitted on 1 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Guaranteed characterization of exact non-asymptotic confidence regions in nonlinear parameter estimation

Michel Kieffer ∗,*∗,*∗∗, Eric Walter ∗,

∗ L2S, CNRS, Supelec, Univ Paris-Sud, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France
** Partly on leave at LTCI, CNRS, Telecom ParisTech, 46 rue Barrault, 75013 Paris
*** Institut Universitaire de France, 103 bld Saint-Michel, 75005 Paris

Abstract:
Recently, a new family of methods has been proposed for characterizing accuracy in nonlinear parameter estimation by Campi et al. These methods make it possible to obtain exact, non-asymptotic confidence regions for the parameter estimates under relatively mild assumptions on the noise distribution, namely that the noise samples are independently and symmetrically distributed.

A key issue is drawing conclusions that are as little prejudiced as possible, and this paper presents a new method for doing so in a guaranteed way, based on the LSCR approach (Leave out Sign-dominant Correlation Regions) recently developed by Campi et al. The paper is organized as follows. The LSCR approach is recalled in Section 2. Section 3 shows how interval analysis can be used to characterize the exact confidence regions provided by LSCR in a global and guaranteed way. Examples are treated in Section 4, and conclusions drawn in Section 5.

1. INTRODUCTION

When a vector \( \mathbf{p} \) of parameters of some approximate mathematical model is estimated from a noisy data vector \( \mathbf{y} \), this is usually via the minimization of some cost function \( J(\mathbf{p}) \), for instance

\[
J(\mathbf{p}) = \| \mathbf{y} - \mathbf{y}_m(\mathbf{p}) \|_2^2,
\]

where \( \mathbf{y}_m(\mathbf{p}) \) is the vector of model outputs, assumed here to be a deterministic function of \( \mathbf{p} \) and \( \| \cdot \|_2 \) is a (possibly weighted) \( l_2 \) norm. Then

\[
\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} J(\mathbf{p}).
\]

This procedure is fraught with difficulties in the general nonlinear case. The parameters of the model may not be identifiable uniquely (i.e., there may be several values of \( \hat{\mathbf{p}} \) that yield exactly the same vector \( \mathbf{y}_m(\hat{\mathbf{p}}) \), in which case there are several global minimizers of the cost function). The numerical algorithm used to compute \( \hat{\mathbf{p}} \) may also get trapped at a parasitic local minimizer because of inadequate initialization.

Even if a single numerical vector \( \hat{\mathbf{p}} \) is obtained and if \( \mathbf{y} \) and \( \mathbf{y}_m(\hat{\mathbf{p}}) \) are reassuringly similar, it would be naive to consider \( \hat{\mathbf{p}} \) as the final answer to the estimation problem. One should instead attempt to attach some quality tag to \( \hat{\mathbf{p}} \) by assessing the reliability of the numerical values thus obtained. This is especially important if one wants to estimate physically meaningful parameters of some knowledge-based model in physics, chemistry, biology, etc., or if decisions have to be taken on the basis of the numerical values of the model parameters to tune controllers or to detect faults, for instance.

* This work has been partly supported by the ANR CPP.
LSCR defines, without any approximation, regions $\Theta$ to which the parameter vector $p^*$ of the true system belongs with some specified probability.

Let $\varepsilon_t(p)$ be a prediction error, such that $\varepsilon_t(p^*)$ is a realization of the noise corrupting the data at time $t$. The procedure for computing one such confidence region is as follows:

1. Select two integers $r \geq 0$ and $q \geq 0$.
2. For $t = 1 + r, \ldots, k + r = n$, compute
   \[ c^r_{t-r} (p) = \varepsilon_t (p) \varepsilon_t (p), \]
   \[ s^r_{r-r} (p) = \sum_{k=1}^{r} c^r_{k-r} (p), \quad i = 1, \ldots, m, \]
   where $\mathbb{I}_i$ is a subset of a set $\mathbb{I}$ of indexes and the collection $G$ of these subsets $\mathbb{I}_i$, $i = 1, \ldots, m$, forms a group under the symmetric difference operation, i.e., $(\mathbb{I}_i \cup \mathbb{I}_j) - (\mathbb{I}_i \cap \mathbb{I}_j) \in G$.
3. Find the set $\Theta_{r,q}^r$ such that at least $q$ of the functions $s^r_{r-r} (p)$ are larger than 0 and at least $q$ are smaller than 0.

This set is such that the probability that $p^*$ belongs to it can be computed exactly
\[ Pr (p^* \in \Theta_{r,q}^r) = 1 - 2q/m. \]

The shape and size of $\Theta_{r,q}^r$ depend not only on the values given to $r$ and $q$ but also on the group $G$ and its number of elements $m$. A procedure for generating a group of appropriate size is suggested in Gordon [1974].

The set $\Theta_{r,q}^r$ may be defined more formally as
\[ \Theta_{r,q}^r = \Theta_{r,q}^- \cap \Theta_{r,q}^+, \]
with
\[ \Theta_{r,q}^- = \left\{ p \in \mathbb{P} \text{ such that } \sum_{i=1}^{m} \tau_i^{\varepsilon_i(p)} \geq q \right\}, \]
\[ \Theta_{r,q}^+ = \left\{ p \in \mathbb{P} \text{ such that } \sum_{i=1}^{m} \tau_i^{\varepsilon_i(p)} \geq q \right\}, \]
where $\mathbb{P}$ is the prior domain for $p$ and where
\[ \tau_i^{\varepsilon_i(p)} = \begin{cases} 1 & \text{if } s^r_{r-r} (p) > 0, \\ 0 & \text{otherwise}, \end{cases} \]
and
\[ \tau_i^{\varepsilon_i(p)} = \begin{cases} 1 & \text{if } s^r_{r-r} (p) \geq 0, \\ 0 & \text{otherwise.} \end{cases} \]

The set $\Theta_{r,q}^r$ contains all values of $p \in \mathbb{P}$ such that at least $q$ of the functions $s^r_{r-r} (p)$ are smaller than 0, whereas $\Theta_{r,q}^r$ contains all values of $p \in \mathbb{P}$ such that at least $q$ of the functions $s^r_{r-r} (p)$ are larger than 0.

When the model studied is driven by an input $u_t$, one may obtain a similar confidence region by substituting $c^r_{t-r,s} (p) = u_{t-r} (p) \varepsilon_t (p)$ for $c^r_{t-r} (p) = \varepsilon_t (p) \varepsilon_t (p)$ in the procedure above, thus replacing autocorrelations by intercorrelations. One then computes a set $\Theta_{r,q}^r$, again such that
\[ Pr (p^* \in \Theta_{r,q}^r) = 1 - 2q/m. \]

The fact that the set $\Theta_{r,q}^r$ (or $\Theta_{r,q}^r$) obtained by this approach is exact does not mean that its volume is minimal, and the resulting confidence region may turn out to be much too large to be useful. One may then intersect several such regions. For a given value of $q$ and $m$, assume that $n_e$ confidence regions $\Theta_{r,q}^r$ and $n_u$ confidence regions $\Theta_{s,q}^r$ have been obtained for $n_e$ values of $r$ and $n_u$ values of $s$. The probability that $p^*$ belongs to the intersection $\Theta$ of these $(n_e + n_u)$ regions then satisfies
\[ Pr (p^* \in \Theta) \geq 1 - (n_e + n_u)2q/m. \]

The price to be paid for taking the intersection of several confidence regions is that the probability that $p^*$ belongs to the resulting confidence region is no longer known exactly, as only a lower bound for this probability is available.

3. GUARANTEED CHARACTERIZATION VIA INTERVAL ANALYSIS

For the LSCR approach (as well as for the other approaches to build confidence regions), a key issue is the guaranteed and global numerical characterization of the confidence region. However, and rather surprisingly, this issue seems seldom addressed outside the community of bounded-error estimation. We concentrate here on LSCR confidence regions, but other types of confidence regions could benefit of a similar treatment.

3.1 Problem statement

In LSCR, one has to characterize an intersection of sets defined as
\[ \Psi_q = \{ p \in \mathbb{P} \text{ such that } \sum_{i=1}^{m} \tau_i (p) \geq q \}, \]
where $\tau_i (p)$ is some indicator function
\[ \tau_i (p) = \begin{cases} 1 & \text{if } f_i (p) > 0, \\ 0 & \text{otherwise,} \end{cases} \]
and where $f_i (p)$ depends on the model structure, the measurements, and the parameter vector $p$.

Characterizing $\Psi_q$ may be alternatively formulated as a set-inversion problem Jaulin and Walter [1993]
\[ \Psi_q = \mathbb{P} \cap \tau^{-1} ([q, m]), \]
with
\[ \tau (p) = \sum_{i=1}^{m} \tau_i (p), \]
which may be efficiently solved via interval analysis Moore [1966], Jaulin et al. [2001]. The next sections briefly recall the basic notions of interval analysis required.

3.2 Inclusion functions and centered forms

Interval analysis considers closed interval $[x] = [x, x]$ of $\mathbb{R}$ and extends all arithmetic operations and elementary functions on real numbers to intervals. For arithmetic operations,
\[ [x] + [y] = [x + y], \]
where $o \in \{ +, -, \cdot, / \}$. The set (17) is easily evaluated from the bounds of $[x]$ and $[y]$ for the addition, subtraction, multiplication, and division when $0 \notin [y]$. 
The range of a continuous function \( f : \mathcal{D} \subseteq \mathbb{R} \to \mathbb{R} \) over an interval \([x] \subset \mathcal{D}\)

\[
f([x]) = \{ f(x) | x \in [x] \}
\]  
(18)
is again easily obtained when \( f \) is monotonic from evaluations involving the bounds of \([x]\). For elementary non-monotonic functions, such as all trigonometric functions, simple algorithms may be put at work to evaluate (18).

Nevertheless, in the general case, obtaining (18) requires simple algorithms may be put at work to evaluate (18).

Monotonic functions, such as all trigonometric functions, are Cartesian product of intervals, and to vector-valued functions.

\( f \) provides the smallest interval containing \( f([x]) \) with \([f([x])] \)

Interval analysis makes it possible, through the notion of inclusion functions, to obtain guaranteed outer-approximations of \( f([x]) \). An inclusion function \([f([x])] \)

\[
\forall [x] \subset \mathcal{D}, \ f([x]) \subset [f([x])],
\]

with \([f([x])] \)

A minimal inclusion function provides the smallest interval containing \( f([x]) \) for all \([x] \subset \mathcal{D}\).

Various types of inclusion functions have been considered. The simplest one is the natural inclusion function \([f([x])] \), obtained by replacing, in the formal expression of \( f \), all occurrences of the real variable \( x \) by its interval counterpart \([x]\) and by performing all operations and elementary function evaluations on intervals. The centered form \([f([x])] \)

\[
[f([x])] = \text{mid}(\mathcal{D})([f([x])])
\]

takes advantage of the mean-value theorem to get a potentially more accurate approximation. It assumes that \( f \) is differentiable, and for any \( m \in [x] \), is defined as

\[
[f([x])] = f(m) + ([x] - m)[f']([x]),
\]

where \([f']([x])] \)

is an inclusion function of the first derivative of \( f \). Usually, \( m \) is taken as the midpoint of \([x] \), \( m = \text{mid}([x]) \).

Inclusion functions are in general pessimistic, i.e., they provide coarse outer-approximations of the range of function, especially for large intervals and when \( x \) are many occurrences of the same real variables. Pessimism decreases when the width of the interval argument decreases, however.

All these notions are extended to interval vectors or boxes, which are Cartesian product of intervals, and to vector-valued functions. See, e.g., Jaulin et al. [2001] for more details.

### 3.3 SIVIA

Consider a set-inversion problem where one has to characterize the set

\[
\mathcal{X} = [x] \cap \mathcal{T}^{-1}(\mathcal{Y}),
\]

where \( f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}^m \), \( \mathcal{Y} \subseteq \mathbb{R}^m \), and \([x] \subset \mathcal{D}\) is some initial search box for \( \mathcal{X} \). The aim of the Set Inverter Via Interval Analysis (SIVIA) Jaulin and Walter [1993] is to provide an inner approximation \( \mathcal{X} \) and an outer approximation \( \overline{\mathcal{X}} \) of \( \mathcal{X} \), represented by subpavings, i.e., unions of non-overlapping boxes. The distance between \( \mathcal{X} \) and \( \overline{\mathcal{X}} \) is indicative of the quality of the approximation of \( \mathcal{X} \). SIVIA requires an inclusion function \([f]\) to be available for \( f \).

SIVIA iteratively partitions the box \([x]\) into subboxes on which the following tests are applied. Consider a given subbox \([\mathbf{x}]\) of \([x]\). If \([f([\mathbf{x}])] \subset \mathcal{Y}\) then (19) implies that

\[
f([\mathbf{x}]) \subset \mathcal{Y}
\]

and thus that \([\mathbf{x}] \subset \mathcal{X}\). In this case, \([\mathbf{x}] \) is stored in \( \mathcal{X} \) and in \( \overline{\mathcal{X}} \). If \([f([\mathbf{x}])] \) \(\cap\) \(\mathcal{Y} = \emptyset\) then (19) implies that \( f([\mathbf{x}]) \) \(\cap\) \(\mathcal{Y} = \emptyset\) and thus that \([\mathbf{x}] \cap \mathcal{X} = \emptyset\).

In this case, \([\mathbf{x}] \) is not considered any further. If none of the two previous tests is true, \([\mathbf{x}] \) is undetermined. In this case, if \([\mathbf{x}] \) is large enough, i.e., if the largest width of its component intervals is larger than some precision parameter \(\varepsilon\), \([\mathbf{x}] \) is bisected into \([\mathbf{x}]_1\) and \([\mathbf{x}]_2\) on which the previous tests are applied again. If \([\mathbf{x}] \) is too small to be bisected, it is stored in \( \overline{\mathcal{X}} \).

The efficiency of SIVIA is conditioned by the accuracy of the inclusion function available for \( f \).

To address the set inversion problem introduced in Section 3.1 with SIVIA, an inclusion function for \( \tau \) is necessary, which will be based on inclusion functions for the \( \tau \)'s, and consequently for the \( f_i \)'s.

### 3.4 Contractors for guaranteed characterization

Large enough undetermined boxes need to be bisected by SIVIA. Indetermination often results from range overestimation by inclusion functions. As a consequence, boxes have to be bisected many times to allow one to conclude on the position of the resulting boxes with respect to \( \mathcal{X} \). This entails intractable computational complexity, even for a moderate dimension of \( \mathbf{p} \).

Contractors Jaulin et al. [2001] partly address this issue. A contractor \( C_{\mathcal{X},\mathcal{Y}} \) associated with the generic set-inversion problem (21) is a function taking a box \([x]\) as input and returning a box

\[
C_{\mathcal{X},\mathcal{Y}}([x]) \subset \mathcal{X}
\]

such that

\[
[x] \cap \mathcal{X} = C_{\mathcal{X},\mathcal{Y}}([x]) \cap \mathcal{X},
\]

so no part of \( \mathcal{X} \) in \([x]\) is lost. It allows parts of the candidate box \([x]\) that do not belong to \( \mathcal{X} \) to be eliminated, without the need to perform any bisection. Various types of contractors have been proposed in the literature, e.g., the contractors by interval constraint propagation, by parallel linearization, the Newton contractor, the Krawczyk contractor, etc.

In the specific problems considered here, the role of \( \mathbf{x} \) is taken by \( \mathbf{p} \). The fact that the function \( \tau \) introduced in (16) is not differentiable forbids the use of most classic contractors, so a specific contractor is needed. The new contractor proposed is implemented in two steps. It assumes that the functions \( f_i \) involved in (14) are differentiable. First, a set of \( m \) possibly overlapping subboxes of \([\mathbf{p}]\) are built, trying to remove all values of \( \mathbf{p} \) from \([\mathbf{p}]\) such that \( f_i(\mathbf{p}) < 0 \), \( i = 1, \ldots, m \). Second, the union of all non-empty intersections of at least \( q \) of these boxes is computed to get a possibly contracted box.

**Box contraction using the \( f_i \)'s** The first step uses the centered inclusion function of \( f_i \), which, for some \( m \in [\mathbf{p}] \), may be written as

\[
[f_i]([\mathbf{p}]) = f_i(\mathbf{m}) + ([\mathbf{p}] - \mathbf{m})' [g_i]([\mathbf{p}])
\]

\[
= f_i(\mathbf{m}) + \sum_{j=1}^{n_q} ([p_j] - m_j) [g_{i,j}]([\mathbf{p}]),
\]

(24)
where $g_i$ is the gradient of $f_i$. Using (25), we build a contractor $C_{f_i, [0, \infty]}$ for the set of all values of $p \in [p]$ such that $f_i(p) \geq 0$ as follows.

For the $k$-th component $[p_k]$ of $[p]$, when $0 \notin [g_i,k]$ ([p]), $C_{f_i, [0, \infty]}$ associates the contracted interval

$$
[p'_{k,i}] = [p_k] \cap \left( \left( \left( [f_{i,c}] ([p]) \cap [0, \infty] \right) - f_i(m) \right) - \sum_{j=1, j \neq k}^{n_p} (p_j - m_j) [g_{i,j}] ([p]) / [g_j,k] ([p]) + m_k \right) .
$$

(26)

When $0 \notin [g_i,k]$ ([p]), $C_{f_i, [0, \infty]}$ leaves $[p_k]$ unchanged, i.e.,

$$
[p'_{k,i}] = [p_k] .
$$

(27)

**Proposition 1.** For all $[p_i]$, $i = 1, \ldots, m$, built using (26) and (27), one has $[p'_i] \subset [p]$ and

$$
[p'_i] \cap f_i^{-1}([0, \infty]) = [p] \cap f_i^{-1}([0, \infty]) .
$$

(28)

and

$$
[f_i(p)'_0] = f_i(m) + \sum_{j=1}^{n_p} (p_j - m_j) g_{i,j}(\zeta) .
$$

(29)

Since $p_0 \in f_i^{-1}([0, \infty])$.

$$
[f_i(p)'_0] = 0 \in [0, \infty] .
$$

(30)

Provided that $g_{i,k}(\zeta) \neq 0$, (30) implies that the $k$-th component of $p'_i$ satisfies

$$
p_0'_{k} = \left( f_i(p)'_0 - f_i(m) \right)
$$

$$
- \sum_{j=1, j \neq k}^{n_p} (p_j - m_j) g_{i,j}(\zeta) / [g_j,k] ([p]) + m_k .
$$

(31)

Since $[f_{i,c}](\cdot)$, introduced in (25), is an inclusion function for $f_i(\cdot)$, (31) combined with the fact that $p_0 \in [p]$ and $\zeta \in [p]$, implies that

$$
p_0'_{k} \in [p_k] \cap \left( \left( \left( [f_{i,c}] ([p]) \cap [0, \infty] \right) - f_i(m) \right) - \sum_{j=1, j \neq k}^{n_p} (p_j - m_j) [g_{i,j}] ([p]) / [g_j,k] ([p]) + m_k \right) .
$$

(32)

and

$$
[f_i(p)'_0] \subset [p'_i] .
$$

(33)

Provided that $0 \notin [g_i,k] ([p])$. Thus for each $k$, $p_0'_{k} \in [p'_i]$. Since $p_0'$ also belongs to $f_i^{-1}([0, \infty])$, $p_0' \in [p'] \cap f_i^{-1}([0, \infty])$, which completes the proof.

Algorithm 1.

$$
\begin{align*}
\mathcal{L} &= \{ C_{f_1, [0, \infty]} ([p]), \ldots, C_{f_m, [0, \infty]} ([p]) \} \\
&= \{ [p'_1], \ldots, [p'_m] \} .
\end{align*}
$$

(35)

(36)

Some of them may be empty, in which case, $[p'_i] = \emptyset$ indicates that there is no $p \in [p]$ such that $f_i(p) \geq 0$. Our aim is to evaluate a subbox $[p'_i]$ of $[p]$ such that

$$
\Psi_q \cap [p'_i] = \Psi_q \cap [p] .
$$

(37)

and satisfying

$$
P = \left( \bigcap_{j \in \{1, \ldots, m\}} [p'_j] \right) .
$$

(38)

and

$$
\Psi_q \cap [p'] = \Psi_q \cap [p] ,
$$

(39)

with $\Psi_q$ as defined in (13).

**Proposition 2.** For any box $[p']$, satisfying (39), one has

$$
\Psi_q \cap [p] = \Psi_q \cap [p'] ,
$$

(40)

provided that $0 \notin [g_i,k] ([p])$. Thus for each $k$, $p_0'_{k} \in [p'_i]$. Since $p_0'$ also belongs to $f_i^{-1}([0, \infty])$, $p_0' \in [p'] \cap f_i^{-1}([0, \infty])$, which completes the proof.

Algorithm 1 formalizes a computation carried out on an example in Jaulin [2011]. It aims at building an outer approximating interval of the $q$-relaxed intersection of $m$ scalar intervals. The extension to boxes is obtained by applying Algorithm 1 componentwise.

**Consider a list $\mathcal{L} = \{ [p_1], \ldots, [p_m] \}$ of $m$ scalar intervals. Algorithm 1 builds the smallest interval containing the union of all intersections of $q$ intervals with a complexity $O(\log m)$. This is the smallest interval containing $P$ as defined by (37) in the scalar case.**

**Steps 4 and 8 of Algorithm 1, $p \in [p_j]$ if $p \in [p_j]$ and $p \in [p_j]$ = 0 otherwise.**
When $\mathcal{L}$ is a list of boxes of the same dimension, one may simply apply Algorithm 1 component by component.

4. EXAMPLES

Two examples are considered in what follows to illustrate the guaranteed characterization of exact confidence regions provided by LSCR.

All computations were carried out with Intlab Rump [2001], the interval-analysis toolbox for Matlab.

4.1 Exponential model

Consider first a system such that

$$\begin{align*}
y_t &= a^* \exp(-b^* t) + w_t \\
    y_t^m &= a \exp(-b t),
\end{align*}$$

where, for $t = 0, T, \ldots, (n-1) T$, the $w_t$’s are assumed to be realizations of iid $\mathcal{N}(0, \sigma^2)$ variables.

Take as a model

$$\begin{align*}
y_t^m(p) &= a \exp(-b t),
\end{align*}$$

which is nonlinear in $p = (a, b)^T$. For the true value of the parameter vector, the prediction errors defined as

$$\varepsilon_t(p) = y_t - y_t^m(p),$$

with $y_t^m(p)$ taken from (41), are iid $\mathcal{N}(0, \sigma^2)$. Their distribution is thus symmetric, and LSCR applies.

The characterization of $\Theta_{r=1,q=3}^\tau$ (or $\Theta_{r=1,q=3}^u$) is addressed using SIVIA. For that purpose, inclusion functions for the indi-
The availability of efficient contractors is mandatory to address the problems of characterizing confidence regions for models with a large number of parameters.

REFERENCES


The availability of efficient contractors is mandatory to address the problems of characterizing confidence regions for models with a large number of parameters.

**5. CONCLUSIONS AND PERSPECTIVES**

Interval analysis provides tools to evaluate guaranteed inner and outer-approximations of non-asymptotic confidence regions defined by LSCR. Illustrations have been provided for two models with outputs nonlinear in their parameters. Applications to the characterization of confidence regions defined by SPS may be found in Kieffer and Walter [2012].

Accurate inclusion functions are particularly difficult to obtain for the functions involved in LSCR, due to the many occurrences of the parameters involved in the evaluation of (4). Symbolic manipulations of the expressions involved to reduce the number of occurrences of the parameters may be particularly useful to improve the efficiency of SIVIA and to design better contractors than those considered in Section 3.4.1.