NEW RESULTS ON THE IDENTIFICATION OF INTERVAL PREDICTOR MODELS

M.C. Campi [*] **G. Calafiore** [**] **S. Garatti** [***]

[*] *Dipartimento di Elettronica per l'Automazione - Università di Brescia, Italy. e-mail: campi@ing.unibs.it*
[**] *Dipartimento di Automatica e Informatica - Politecnico di Torino, Italy. e-mail: giuseppe.calafiore@polito.it*
[***] *Dipartimento di Elettronica ed Informazione - Politecnico di Milano, Italy. e-mail: sgaratti@elet.polimi.it*

Abstract: In this paper, the problem of identifying a predictor model for an unknown system is studied. Instead of standard models returning a prediction value as output, we consider models returning prediction intervals. Identification is performed according to some optimality criteria, and, thanks to this approach, we are able to provide, independently of the data generation mechanism, an exact evaluation of the reliability (i.e. the probability of containing the actual true system output value) of the prediction intervals returned by the identified models. This is in contrast to standard identification where strong assumptions on the system generating data are usually required.

Keywords: Identification, Set-valued maps, Learning theory, Convex optimization

1. INTRODUCTION

In this paper, we are interested in deriving predictor models from data, i.e. models that can be used for prediction purposes. Prediction is not only important per se, but it also plays a significant role in many application endeavors such as predictive control and signal processing.

Along the standard routes in system identification (Ljung, 1999; Söderström and Stoica, 1989), the model is typically obtained by first selecting a parametric model structure, and then by estimating the model parameters using an available batch of observations. The so obtained identified model may then be used to predict the future output of the system.

As is obvious, however, the predicted output value is always an approximation of the actual value the system output will be, so that crediting the predicted value with reliability will depend on the prediction accuracy the application at hand demands. In turn, this entails that a predicted value is of little use if derived without a tag certifying its accuracy.

A practical way to assign the accuracy tag is to provide an interval (or region) of confidence around the predicted value to which the future output is guaranteed to belong with a certain probability. In the standard identification approach, this is typically done a-posteriori by estimating the level of the noise affecting the model and by deriving the confidence interval from such an estimate.

A crucial observation which has been pointed out many times in the literature is that the confidence (or probability) of the prediction interval may be difficult to evaluate if the system generating the observations is structurally different from the parametric model. This entails that reliable results on the interval confidence can be obtained only if strong hypotheses on the structure and order of the mechanism that generates the data are made.

In this paper, we follow a novel approach for the construction of predictor models which returns prediction intervals with guaranteed confidence under general conditions. In contrast to the standard way of proceeding, we consider from the beginning model structures returning an interval as output (these models are called interval predictor models (IPMs) and are strictly connected to set-valued map (Aubin, 1990; Aubin and Cellina, 1984)). In this way, the model structure is directly tailored to the final purpose of obtaining a prediction interval. For the selection of the model within the chosen structure, only the interval models which are compatible with the observed data (in a sense rigorously defined in Section 3) are considered, and, among these, the model returning the smallest possible prediction interval is chosen.

Through this new approach, we gain a fundamental advantage: the reliability of the estimated interval can be quantified independently of the data-generating mechanism. In other words, we are able to evaluate the probability with which a future output is guaranteed to belong to the predicted interval, whatever the system generating the data is.

The results of the present paper build on previous work of the same authors (Calafiore and Campi, 2002), (Calafiore *et al.*, 2002) and (Calafiore and Campi, 2005). Our main contributions here are: i) identification is developed in a more general framework allowing for the presence of outliers; ii) furthermore, we provide a considerably improved bound on the number of samples required to attain the desired reliability.

The paper is structured as follows. In Section 2 interval predictor models are introduced, while the identification of these models is presented in Sections 3 and 4. Section 5 addresses the fundamental problem of evaluating the reliability of the identified model. Finally, some simulation examples are given in Section 6. Due to space limitations, proofs are omitted.

2. INTERVAL PREDICTOR MODELS

In this section, we introduce the key element of our approach: models that return an interval as output (Interval Predictor Models – IPMs).

Let $\Phi \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}$ be given sets, called respectively the *instance* set and the *outcome* set. Then, an interval predictor model is simply a rule that assigns to each instance vector $\varphi \in \Phi$ a corresponding output interval (or region) in Y. That is, an IPM is a set-valued map

$$I : \varphi \to I(\varphi) \subseteq Y. \tag{1}$$

In (1), φ is a regression vector containing explicative variables on which the system output y depends, and $I(\varphi)$ is the prediction interval. For an observed φ, $I(\varphi)$ should contain y with high (guaranteed) probability.

Throughout the paper we will consider IPMs in a parametric from. Precisely, consider a family of functions mapping Φ into Y parameterized by a vector q ranging in some set $Q \subseteq \mathbb{R}^{n_q}$

$$\mathscr{M} = \{y = M(\varphi, q), \quad q \in Q \subseteq \mathbb{R}^{n_q}\},$$

where for a given q, M is a one-valued map $\Phi \to Y$. Then, an IPM is obtained by associating to each $\varphi \in \Phi$ the set of all possible outputs given by $M(\varphi, q)$ as q is let vary over Q, viz.

$$I(\varphi) = \{y : \; y = M(\varphi, q) \text{ for some } q \in Q\}. \tag{2}$$

Remark 1. Note that $y = M(\varphi, q)$ should not be considered as a model family from which a specific model has to be selected. Instead, this parametric model is merely an instrument through which an interval map $I(\varphi)$ is defined.

An example of a parametric IPM is that derived from standard linear regression functions:

$$\mathscr{M} = \{y = \vartheta^T \varphi + e, \; \vartheta \in \Theta \subseteq \mathbb{R}^n, \; |e| \leq \gamma \in \mathbb{R}\}. \tag{3}$$

In this case, $q = [\vartheta^T \; e]^T \in \mathbb{R}^{n+1}$ and $Q = \Theta \times [-\gamma, \gamma]$. Θ can be e.g. a sphere with center c and radius r:

$$\Theta = \mathscr{B}_{c,r} = \{\vartheta \in \mathbb{R}^n : \|\vartheta - c\| \leq r\}, \tag{4}$$

or, more generally, an ellipsoidal region:

$$\Theta = \mathscr{E}_{c,P} = \{\vartheta \in \mathbb{R}^n : (\vartheta - c)^T P^{-1} (\vartheta - c) \leq 1\}, \tag{5}$$

where P is a positive definite matrix.

Note that a parametric IMP as defined in (2) is determined by the set Q where q is let vary. For such a reason, parametric IPMs are usually denoted by I_Q.

A *class* of parametric IPM is simply a collection of I_Q, where Q belongs to a family $\mathscr{Q}$ of feasible sets.

For instance, for the parametric IPM defined by (3),(4), $Q = \mathscr{B}_{c,r} \times [-\gamma, \gamma]$ is univocally determined by c, r and γ and $\mathscr{Q}$ can be obtained considering all possible combinations of such parameters, i.e.

$$\mathscr{Q} = \{Q = \mathscr{B}_{c,r} \times [-\gamma, \gamma] : \; c \in \mathbb{R}^n, r \in \mathbb{R}, \gamma \in \mathbb{R}\}. \tag{6}$$

Similarly, when $Q = \mathscr{E}_{c,P} \times [-\gamma, \gamma]$ we have

$$\mathscr{Q} = \{Q = \mathscr{E}_{c,P} \times [-\gamma, \gamma] : c \in \mathbb{R}^n, P \in \mathbb{S}_+, \gamma \in \mathbb{R}\}, \tag{7}$$

where $\mathbb{S}_+$ is the set of positive definite $n \times n$ matrices.

3. IPMS IDENTIFICATION

Suppose now that the explicative variable φ and the output y are generated according to some data-generating mechanism, and that a bunch of observations $D_N = \{\varphi(t), y(t)\}_{t=1,\ldots,N}$ is available. From these data we want to identify an IPM $\widehat{I}_N$ among a given class of parametric IPMs I_Q, $Q \in \mathscr{Q}$.

Identification is guided by the following two criteria.

On one hand, we require that $\widehat{I}_N$ is not falsified by the observations, i.e. that it is *consistent* with data according to the following definition.

Definition 1. An IPM I is *consistent* with the batch of observations D_N if $y(t) \in I(\varphi(t))$, for $t = 1, \ldots, N$.

On the other hand, we want $\widehat{I}_N$ to be tight and for this purpose we suppose that a cost criterion μ_Q is defined, so that, for each feasible Q, μ_Q assesses the magnitude of the intervals returned by I_Q.

For example, consider parametric IPMs defined by

(3),(4). It can be explicitly computed (Calafiore *et al.*, 2002) that $I_Q(\varphi) = [c^T\varphi - r\|\varphi\| - \gamma, c^T\varphi + r\|\varphi\| + \gamma]$, so that, given a φ, the size of the returned interval depends on r and γ. Then, as a cost criterion we may consider

$$\mu_Q = \gamma + \alpha r, \tag{8}$$

where α is a fixed nonnegative number. If e.g. $\alpha = \mathbb{E}[\|\varphi(t)\|]$, then μ_Q measures the average amplitude of I_Q.

Similarly, for parametric IPMs defined by (3),(5) we have that

$$I_Q(\varphi) = [c^T\varphi - \sqrt{\varphi^T P\varphi} - \gamma, c^T\varphi + \sqrt{\varphi^T P\varphi} + \gamma],$$

and as a cost criterion we may consider

$$\mu_Q = \gamma + \mathrm{Tr}[PW], \tag{9}$$

where W is a weighting matrix and $\mathrm{Tr}[\cdot]$ means trace.

Combining the consistency requirement with that on tightness, the identification of $\widehat{I}_N$ can be then performed solving the following constrained optimization problem with respect to Q.

Problem 1. (IPM identification).
Find $\widehat{I}_N := I_{\widehat{Q}_N}$ where

$$\widehat{Q}_N = \arg\min_{Q\in\mathscr{Q}} \mu_Q \text{ s.t. } y(t) \in I_Q(\varphi(t)),\ t = 1,\dots,N.$$

Problem 1 may look hard to solve. However, it is worth noting that for many standard IPM parameterizations and cost criteria (e.g IPMs based on linear regressive models) Problem 1 turns out to be a *convex* optimization problem which can be solved without much computational effort.

In particular, for parametric IPMs defined by (3),(4) and for $\mathscr{Q}$ and μ_Q defined in (6) and (8), respectively, Problem 1 is equivalent to the following linear programming problem (note that $Q = Q(c,r,\gamma)$ in this case):

Problem 1.a. (Linear IPM - spherical parameter set).
$\widehat{I}_N = I_{Q(\widehat{c}_N,\widehat{r}_N,\widehat{\gamma}_N)}$ where

$$\begin{aligned}
&\widehat{c}_N, \widehat{r}_N, \widehat{\gamma}_N = \arg\min_{c,r,\gamma} \gamma + \alpha r \text{ s.t.}\\
&\quad r,\gamma \geq 0\\
&\quad y(t) \geq \varphi^T(t)c - r\|\varphi(t)\| - \gamma, \quad t = 1,\dots,N\\
&\quad y(t) \leq \varphi^T(t)c + r\|\varphi(t)\| + \gamma, \quad t = 1,\dots,N.
\end{aligned}$$

Similarly (see (Calafiore *et al.*, 2002) for details), for the IPMs defined by (3),(5) with $\mathscr{Q}$ and μ_Q as in (7) and (9) Problem 1 becomes a semi-definite (convex) optimization problem which can be solved with standard methods, see e.g. (Vandenberghe and Boyd, 1996).

4. IDENTIFICATION WITH DISCARDED CONSTRAINTS

It is well known that in many cases there are few data (the so called *outliers*) whose value is anomalous as compared to other observations. These data are of no use to understand the data-generating mechanism. As is clear, in presence of outliers, requiring consistency for *all* the available observations as in Problem 1 may be unsuitable. Indeed, even a single anomalous datum may adversely affect the final result, introducing conservatism (untightness) in the identified model. In this case, a wiser procedure would be to discard "bad data" from available observations, before performing the identification.

From an abstract point of view, the IPM identification with violated constraints can be outlined as follows. Let k, $k \ll N$, be a fixed number and let $\mathscr{A}$ be a decision algorithm through which k observations are discarded from D_N. The output of $\mathscr{A}$ is the set $\mathscr{A}(D_N) = \{i_1,\dots,i_{N-k}\}$ of $N-k$ indexes from $\{1,\dots,N\}$ representing the constraints still in place. By $\widehat{I}^{\mathscr{A}}_{N-k}$ we denote the identified IPM when k constraints are removed as indicated by $\mathscr{A}$. That is:

Problem 1′. (Identification with discarded constraints).
Find $\widehat{I}^{\mathscr{A}}_{N-k} := I_{\widehat{Q}^{\mathscr{A}}_{N-k}}$, where

$$\widehat{Q}^{\mathscr{A}}_{N-k} = \arg\min_{Q\in\mathscr{Q}} \mu_Q \textit{ s.t. } y(t) \in I_Q(\varphi(t)),\ t \in \mathscr{A}(D_N).$$

Remark 2. As is obvious, note that $\widehat{I}^{\mathscr{A}}_{N-0} = \widehat{I}_N$, so that Problem 1 is a particular case of Problem 1′.

Two main issues now arise: (i) How should the algorithm $\mathscr{A}$ be chosen? (ii) Which is the loss in reliability when $\widehat{I}^{\mathscr{A}}_{N-k}$ is used in place of $\widehat{I}_N$? Point (ii) will be addressed in the next Section 5. Point (i) is instead the subject of the following Section 4.1.

4.1 Choice of $\mathscr{A}$

In order to achieve the best possible benefit from constraints removal, algorithm $\mathscr{A}$ should be chosen so as to discard those constraints whose removal leads to the largest drop in the optimal cost value $\mu_{\widehat{Q}^{\mathscr{A}}_{N-k}}$. To this end, one can try to solve Problem 1 for all possible combinations of $N-k$ constraints taken out from the initials N constraints, and then choose that combination resulting in the lowest value of μ_Q. This brute-force way of proceeding, however, is computationally very consuming as it requires to solve $N!/(N-k)!k!$ optimization problems, a truly large number in general.

The main aim of this section is to present a better algorithm for solving the problem of constraints removal. The approach taken here is in the same spirit of (Bai *et al.*, 2002) and (Matoušek, 1994) where the problem of constraints removal has been studied in a slightly different setting.

We first give some relevant definitions. To avoid notational cluttering, these definitions are given with reference to a generic constrained optimization problem:

$$\mathscr{P}: \min_{z\in Z\subseteq\mathbb{R}^d} f(z) \quad \text{s.t. } z \in Z_i, \quad i = 1,\dots,m, \tag{10}$$

where $Z_i \subseteq \mathbb{R}^d$. The following assumption is assumed to hold for all problems considered in this paper.

Assumption 1. The solution of $\mathscr{P}$ exists and is unique.

Remark 3. The uniqueness requirement in Assumption 1 could be removed at the price of introducing extra technicalities. We have preferred to assume uniqueness to ease the reading.

Let $w(\mathscr{P})$ denote the smallest value of $f(z)$ attainable for problem $\mathscr{P}$, viz. $w(\mathscr{P}) = f(z^*)$ where z^* is the solution of $\mathscr{P}$. We have the following definition.

Definition 2. (Support constraints). The $l-th$ constraint Z_l is a *support constraint* for $\mathscr{P}$ if $w(\mathscr{P}) < w(\mathscr{P}_l)$, where $\mathscr{P}_l$ is the optimization problem obtained from $\mathscr{P}$ by removing the $l-th$ constraint, namely:

$$\mathscr{P}_l : \min_{z \in Z \subseteq \mathbb{R}^d} f(z) \; s.t. \; z \in Z_i, \; i = 1, \ldots, l-1, l+1, \ldots, m.$$

In other words, a support constraint is a constraint whose elimination improves the optimal solution. The following Theorem holds (see (Calafiore and Campi, 2005) for a proof).

Theorem 1. If $\mathscr{P}$ is a convex optimization problem (i.e. $f(z)$ is a convex function of z and Z_i is a convex set for each i), then the number of support constraints for $\mathscr{P}$ is at most d.

Finally, for all problems considered in this paper, we require in the following assumption that the optimal solution with the sole support constraints in place is the same as the optimal solution with all constraints (see Example 1 after the assumption for a degenerate case where this does not apply).

Assumption 2. Given a problem $\mathscr{P}$ as in (10), consider the following optimization problem

$$\mathscr{P}_{sc} : \min_{z \in Z} f(z) \;\; \text{s.t. the support constraints of } \mathscr{P} \text{ are satisfied}$$

Then, $w(\mathscr{P}_{sc}) = w(\mathscr{P})$.

Example 1. Consider the following optimization problem:

$$\min_{(z_1,z_2) \in \mathbb{R}^2} z_2 \quad \text{s.t. } (z_1, z_2) \in Z_a \cap Z_b \cap Z_c, \tag{11}$$

where Z_a, Z_b and Z_c are as in Figure 1. In this case, only Z_a is a support constraint as removing Z_b or Z_c the optimal solution does not change. However, considering the optimization problem subject to Z_a only leads to a different solution than the original problem.

Go back now to the problem of optimally removing k constraints from the initial set of constraints associated to D_N (with a little abuse of notation, we will say "constraints D_N"). Given a subset F of D_N, we will denote by $w(F)$ the smallest value of μ_Q attainable for the optimization Problem 1 obtained by substituting D_N with F. We will also denote by $\text{sc}(F)$ and by $sc_i(F)$, respectively, the set of support constraints and the i-th support constraint of the problem with the constraints F. Finally, suppose that Problem 1 is a convex problem (this is true e.g. for Problem 1.a) so that $|sc(F)| \leq d$, $\forall F \subseteq D_N$, according to Theorem 1 ($|\cdot|$ denotes cardinality).

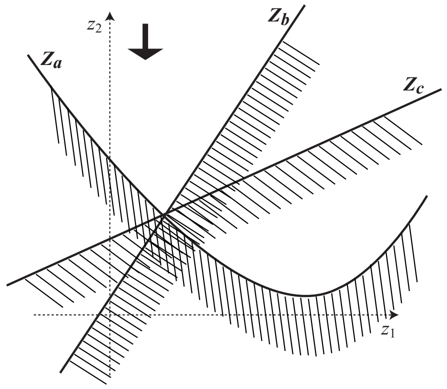

Figure 1. Constraints of the optimization problem (11)

The following Algorithm $\mathscr{A}^*$ optimally discards k observations. Yet, instead of considering all the possible combinations of $N-k$ constraints from the N initials ones, it only considers a subset of situations. Precisely, it constructs a tree of optimization problems as follows: the root is given by Problem 1, with the initial set of constraints D_N; each problem in the tree is obtained from a parent problem simply removing one of the parent problem support constraints. In the end, one simply has to solve the optimization problems at level k in the tree (that is with k constraints removed). Formally, the algorithm goes as follows (here, D^h_{N-i} denotes the constraints of the h-th problem at level i, while M_i is the number of problems at level i and X is a variable that at the end of the algorithm contains the optimal set of $N-k$ constraints).

Algorithm $\mathscr{A}^*$

```
0. $D^1_N := D_N$; $X := D^1_N$; $M_0 = 1$; $i := 0$;
1. $M_{i+1} := 0$
   FOR $h = 1$ TO $M_i$
   FOR $l = 1$ TO $|sc(D^h_{N-i})|$
      $M_{i+1} := M_{i+1} + 1$
      $D^{l+(h-1)\cdot M_i}_{N-i-1} := D^h_{N-i} - sc_l(D^h_{N-i})$
      IF $i+1 = k$ AND $w(D^{l+(h-1)\cdot M_i}_{N-i-1}) \leq w(X)$
      THEN $X := D^{l+(h-1)\cdot M_i}_{N-i-1}$
   END
   END
2. IF $i < k$ THEN $i := i+1$; GO TO 1.
   ELSE $\mathscr{A}^*(D_N) := X$
```

The following theorem holds true.

Theorem 2. Algorithm $\mathscr{A}^*$ is optimal in the sense that it returns a set of $N-k$ constraints resulting in the largest drop of the cost value μ_Q.

Proof: see (Campi *et al.*, 2005).

In Algorithm $\mathscr{A}^*$, only support constraints are relevant to building a tree level from the previous one. In general, given a set of constraints D^h_{N-i}, in order to spot which among these the support constraints are, one has to solve the optimization problems obtained by removing one by one the constraints in D^h_{N-i}, and test if the optimal solution changes. In the following we provide an evaluation of the total number of optimization problems one has to solve to implement $\mathscr{A}^*$. In $\mathscr{A}^*$ the computation of support constraints has to be repeated for all the problems in the tree, from level 0 to level $k-1$. Since for each problem there are at most d support constraints, the number of problems at level i is at most d^i. Moreover, each of these problems has $N-i$ constraints. Thus, a bound to the number of problems which $\mathscr{A}^*$ requires to solve is $N+(N-1)\cdot d+\ldots+(N-k-1)\cdot d^{k-1} \leq N\cdot\frac{d^k-1}{d-1}$. Note that this number is much smaller than $N!/k!(N-k)!$.

As an additional remark, since support constraints have to be active constraints, $\mathrm{sc}(D^h_{N-i})$ can be determined by searching among active constraints of D^h_{N-i} only. This may significantly reduce the number of optimization problems to test.

5. RELIABILITY OF IPMS

In this section, we tackle the fundamental issue of assessing the *reliability* of the IPM $\widehat{I}^{\mathscr{A}}_{N-k}$, identified according to Problem $1'$ (see Section 4). The reliability result applies to any algorithm $\mathscr{A}$ and, in particular, to $\mathscr{A}^*$ discussed at the end of the previous section.

Assume that the observed data D_N are generated as a realization of a bivariate (strict sense) stationary process $\{x(t)\} = \{\varphi(t), y(t)\}$, $\varphi(t) \in \Phi \subseteq \mathbb{R}^n$ and $y(t) \in Y \subseteq \mathbb{R}$. Stationarity says that the system is operating in steady-state. Apart from this, no assumption is made. The system can be e.g. linear corrupted by noise, nonlinear corrupted by noise, or anything else.

Definition 3. Let I be a given IPM. The *reliability* of I is denoted by $R(I)$ and is the probability that a new unseen datum (φ, y), independent of D_N but generated according to the same mechanism, is consistent with I, i.e.

$$R(I) = \mathrm{Prob}_{\mathbb{P}}\{y \in I(\varphi)\},$$

where $\mathbb{P}$ is the probability of $x(t) \in \mathbb{R}^{n+1}$.

The precise assessment of $R(\widehat{I}^{\mathscr{A}}_{N-k})$ in the i.i.d. case is given by the following theorem.

Theorem 3. Assume that $\{x(t)\} = \{\varphi(t), y(t)\}$ is an independent and identically distributed sequence with unknown probability measure $\mathbb{P}$. Suppose also that Problem $1'$ is a *convex* constrained optimization problem so that the number of its support constraints is no greater than d (see Theorem 1), and that the solution of Problem $1'$ is unique (if not, suitable tie-break rules could be used as explained in (Calafiore and Campi, 2005)).

Then, for any $\varepsilon \in (0,1)$ and δ such that

$$\delta = \frac{N!}{(N-d-k)!d!k!}(1-\varepsilon)^{N-d-k}, \qquad (12)$$

it holds that

$$\mathrm{Prob}_{\mathbb{P}^N}\{R(\widehat{I}^{\mathscr{A}}_{N-k}) \geq 1-\varepsilon\} \geq 1-\delta.$$

Proof: see (Campi *et al.*, 2005).

Remark 4. Theorem 3 states that, if N data points are observed, the reliability of the optimal solution $\widehat{I}^{\mathscr{A}}_{N-k}$ of Problem $1'$ is no worse than $1-\varepsilon$ with high probability greater than $1-\delta$. As a matter of fact, since constraints in Problem $1'$ are random (they depend on a realization $x(1), \ldots, x(N)$ of the data-generating stochastic process $\{x(t)\}$), the resulting optimal interval model $\widehat{I}^{\mathscr{A}}_{N-k}$ is random itself. Therefore, its reliability $R(\widehat{I}^{\mathscr{A}}_{N-k})$ can be equal to $1-\varepsilon$ for a given bunch of random observations and not for another. In the theorem, $1-\delta$ refers to the probability $\mathbb{P}^N = \mathbb{P}\times\ldots\times\mathbb{P}$ of observing a "bad" multi-sample $x(1), \ldots, x(N)$ such that the reliability of $\widehat{I}^{\mathscr{A}}_{N-k}$ is less than $1-\varepsilon$.

Remark 5. Note that for $k = 0$, equation (12) reduces to

$$\delta = \frac{N!}{(N-d)!d!}(1-\varepsilon)^{N-d}. \qquad (13)$$

This is the condition guaranteeing $R(\widehat{I}_N) \geq 1-\varepsilon$ with probability no less than $1-\delta$.

Remark 6. It is perhaps worth noticing that, once N and δ have been fixed, the reliability of $\widehat{I}^{\mathscr{A}}_{N-k}$ is not simply the reliability of $\widehat{I}_{N-k}$, even though $\widehat{I}^{\mathscr{A}}_{N-k}$ is obtained through an optimization problem subject to $N-k$ constraints. The reason for this is that the k constraints to be removed from the initial N are *a posteriori* selected (so as to eliminate the constraints which lead to untightness). For this reason, we have $R(\widehat{I}^{\mathscr{A}}_{N-k}) \leq R(\widehat{I}_{N-k})$ as it can be easily verified from equations (12) and (13).

Remark 7. Theorem 3 can be also used to designe an IPM identification experiment. Indeed, suppose to fix ε, δ. Then, equation (12) can be used to determine the number N of observations and the number k of constraints to be removed so as to identify through Problem $1'$ an IPM $\widehat{I}^{\mathscr{A}}_{N-k}$ having reliability $1-\varepsilon$, with probability (confidence) $1-\delta$.

Remark 8. From equation (13), a bound to the number N of samples required to attain a certain reliability $1-\varepsilon$ with confidence $1-\delta$ can be explicitly computed. In fact, after some cumbersome calculations, one can find that $N = \lfloor \frac{2}{\varepsilon}\ln\frac{1}{\beta} + 2d(1+\frac{1}{\varepsilon}\ln\frac{2}{\varepsilon})\rfloor + 1$ ($\lfloor\cdot\rfloor$ = integer part), i.e. that N scales basically as $\frac{1}{\varepsilon}\ln\frac{1}{\delta}$. This greatly improves with respect to the bound given in (Calafiore and Campi, 2002). In particular, the log-dependence on δ allows one to obtain a high confidence without increasing N very much. A similar type of complexity bound has been derived in the context of scenario based optimization in (Calafiore and Campi, 2004).

Remark 9. (Dependent observations). Theorem 3 can be extended to the case of non independent observations. For example, when $D_N = \{x(t)_{t=1,\ldots,N}\}$ is generated by an M-dependent stochastic process (Bosq, 1998), it is quite straightforward to prove that Theorem 3 still holds true if equation (12) is substituted with:

$$\delta = \frac{N!}{(N-d)!d!} \cdot \frac{W!}{(W-k)!k!} \cdot (1-\varepsilon)^{W-k},$$

where $W = \lfloor (N - d(2M+1))/M \rfloor$.

6. NUMERICAL EXAMPLES

Data were generated as $y(t) = u(t)(1+w_1(t)) + w_2(t)$ where $u(t) = \varphi(t)$ was the explicative variable and was a $WGN(0,1)$ (WGN = white gaussian noise), $w_1(t) \approx WGN(0,0.01)$, and $w_2(t)$ was a sequence of independent random variables taking values 0, $+1$, -1 with probability 0.98, 0.01 and 0.01 respectively. $w_2(t)$ merely added outliers to data points.

After collecting 177 observations $u(t), y(t)$, we sought an explanatory interval predictor model of the form (2),(3),(4) with $n = 1$, i.e.

$$I(\varphi(t)) = \{y(t) : y(t) = \vartheta u(t) + e, \\ |e| \leq \gamma, \ \vartheta \in \mathscr{B}_{c,r} \ \}.$$

We set $\mu_Q = \gamma + 0.7r$ (note that $\mathbb{E}[|u(t)|] = 0.7$), and solving Problem 1 we got as optimal IPM parameters $c = 0.708$, $r = 0.537$, $\gamma = 0.024$ and $\mu_Q = 0.4$. The so obtained set-valued map $I(\varphi(t))$ is depicted in Figure 2 along with the collected data points. As it appears, such IPM is untight because of just a single outlier. For this reason, we discarded $k = 1$ observations according to the optimal algorithm $\mathscr{A}^*$ described in Section 4, and solving Problem 1$'$ we found the IPM depicted in Figure 3. We got $c = 1.013$, $r = 0.231$, $\gamma = 0.024$, and $\mu_Q = 0.19$. Precisely, discarding one observation yielded a 50% reduction of the cost μ_Q.

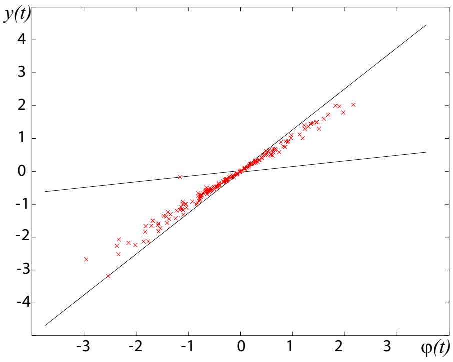

Figure 2. Output interval predictor model identified on the basis of the $N = 177$ available observations.

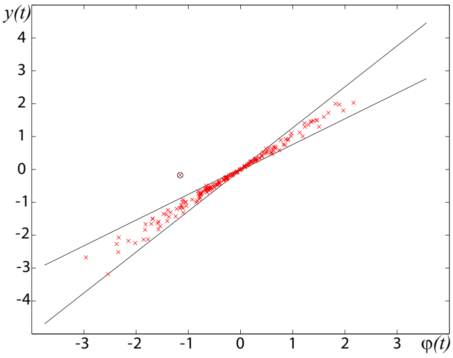

Figure 3. Optimal interval predictor model with $k = 1$ discarded observations.

For what concerns the reliability of the identified IPMs, Theorem (3) *a-priori* states that, with probability at least equal to $1 - \delta = 0.99$, $R(I)$ is no less than 0.9 if no constraints are removed and no less than 0.873 when $k = 1$ constraints are removed (the reliability loss is evaluated as 0.027).

REFERENCES

Aubin, J.P. (1990). *Set-valued analysis*. Birkhäuser. Boston, MA.

Aubin, J.P. and A. Cellina (1984). *Differential inclusions*. Springer-Verlag. Berlin, Germany.

Bai, E., H. Cho and R. Tempo (2002). Optimization with few violated constraints for linear bounded error parameter estimation. *IEEE Transaction on Automatic Control* **47**, 1067–1077.

Bosq, D. (1998). *Non parametric statistics for stochastic processes*. Springer. New York, NY.

Calafiore, G. and M.C. Campi (2002). A learning theory approach to the construction of predictor models. In: *Proceedings of the* 4th *international conference on dynamical systems and differential equations*. pp. 1–9.

Calafiore, G. and M.C. Campi (2004). A new bound on the generalization rate of sampled convex programs. In: *Proceedings of the* 43rd *IEEE conference on decision and control*. Atlantis, Paradise Island, Bahamas, USA.

Calafiore, G. and M.C. Campi (2005). Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming* **102**(1), 25–46.

Calafiore, G., M.C. Campi and L. El Ghaoui (2002). Identification of reliable predictor models for unknown systems: a data-consistency approach based on learning theory. In: *Proceedings of the* 15th *IFAC world congress*. Barcelona, Spain.

Campi, M.C., G. Calafiore and S. Garatti (2005). Identification of interval predictor models. Preprint.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice-Hall. Upper Saddle River, NJ.

Matoušek, J. (1994). On geometric optimization with few violated constraints. *Discrete Computational Geometry* **14**, 365–384.

Söderström, T. and P. Stoica (1989). *System Identification*. Prentice-Hall. Englewood Cliffs, NJ.

Vandenberghe, L. and S. Boyd (1996). Semidefinite programming. *SIAM Review* **38**, 49–95.