Interval Predictor Models: Identification and Reliability [⋆]

M.C. Campi [a,*] G. Calafiore [b] S. Garatti [c]

[a]*Dipartimento di Elettronica per l'Automazione - Università di Brescia, via Branze 38, 25123 Brescia, Italia*

[b]*Dipartimento di Automatica e Informatica - Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italia*

[c]*Dipartimento di Elettronica e Informatica - Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italia*

Abstract

This paper addresses the problem of constructing reliable *interval predictors* directly from observed data. Differently from standard predictor models, interval predictors return a *prediction interval* as opposed to a *single prediction value*. We show that, in a stationary and independent observations framework, the reliability of the model (that is, the probability that the future system output falls in the predicted interval) is guaranteed a-priori by an explicit and non-asymptotic formula, with no further assumptions on the structure of the unknown mechanism that generates the data. This fact stems from a key result derived in this paper, which relates at a fundamental level the reliability of the model to its complexity and to the amount of available information (number of observed data).

Key words: Set-valued Models, Interval Prediction, Convex Optimization, Model Identification, Statistical Learning.

1 Introduction

In this paper we present a novel approach for the construction of *predictor models* — that is, models that can be used for prediction purposes — directly from observed data, and for assessing the reliability of the prediction generated by these models.

Along the standard routes in system identification (see, e.g., [19] and [24]), a model is typically obtained by first selecting a parametric model structure, and then by estimating the model parameters, either using an available batch of observations, or by on-line parameter estimation. The so-obtained model may be used to predict the future output of the system. The predicted value is however of little use if derived without a tag certifying its accuracy ([6], Chapters 1 and 5, [13], Section 4.1, [21]). A practical way of assigning the accuracy tag is to provide an *interval of confidence* around the predicted value, to which the future output is guaranteed to belong with a certain probability (reliability of prediction). For the construction of this interval, two sources of information are used in this standard identification process: *a-priori* information on the true system, and *a-posteriori* information (the data). The mutual strength of this two-layers information set-up drives the compromise in the choice of the model class complexity (bias vs. variance trade-off). Moreover, the final *reliability* of the obtained model depends on such a compromise, and attaching a reliability certificate to the model calls for the use of a-priori knowledge to quantify the bias component.

1.1 Objectives of this paper

In this paper we propose an alternative approach that, under suitable hypotheses (stationarity and independence of the system variables) stated formally in the next section, permits one to derive a reliable interval of confidence for the system output, with no further assumptions on the structure of the mechanism that generates the data. In other words, no a-priori information on the system structure is used to assess reliability. This result is achieved by abandoning the traditional perspective that the model is a one-valued function. We instead introduce models directly returning an interval as output (IPM – Interval Predictor Models). The IPM selection is driven by the principle that the model should correctly describe the already seen data. Among models correctly describing the data, the one returning the smallest possible prediction interval is chosen.

[⋆] This work was supported by MIUR National Research Project "Innovative Techniques for Identification and Adaptive Control of Industrial Systems" and by FIRB project "Learning, randomization and guaranteed predictive inference for complex uncertain systems".

[*] Corresponding Author: Marco C. Campi.
Phone: +39.030.3715458. Fax: +39.030.380014.

Email addresses:
campi@ing.unibs.it (M.C. Campi),
giuseppe.calafiore@polito.it (G. Calafiore),
sgaratti@elet.polimi.it (S. Garatti).

URLs:
http://bsing.ing.unibs.it/∼campi/ (M.C. Campi),
http://staff.polito.it/giuseppe.calafiore/
(G. Calafiore),
http://home.dei.polimi.it/sgaratti/ (S. Garatti).

By the use of IPMs, a key conceptual separation is obtained: the reliability tag only depends on the chosen model class and by the number of observed data and it is therefore always guaranteed, independently of what the data-generating mechanism is. On the other hand, a-priori information still has a role in the proposed approach since selecting a suitable model class results in an IPM with smaller prediction intervals. A strength of the proposed approach is that of returning the smallest possible interval predictor within the chosen class.

1.2 Problem statement

Let $\Phi \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}$ be given sets, called respectively the *explanatory variable* set and the *outcome* set. An unknown (but a-priori fixed) stochastic data generation mechanism generates the data in the form of a sequence of explanatory variables and corresponding outcomes: $(\varphi(t), y(t))$, $t = 1, 2, \ldots$, with $\varphi(t) \in \Phi$, $y(t) \in Y$. We shall assume that this data generation process is stationary.

Assumption 1 (Stationarity) *The process $x(t) = (\varphi(t), y(t))$, $t = 1, 2, \ldots$, with $\varphi(t) \in \Phi \subseteq \mathbb{R}^n$ and $y(t) \in Y \subseteq \mathbb{R}$, is a (strict-sense) stationary discrete-time stochastic process. The (unknown) marginal distribution of the process at time t, which is the same for any t, is denoted with $\mathbb{P}$.*

Remark 1 $\mathbb{P}$ can be interpreted as a "probabilistic cloud" in the $\Phi \times Y$ space. The case in which $y(t)$ is obtained as a function of $\varphi(t)$ is just a particular case where the probability $\mathbb{P}$ is concentrated over the function. In general, the fact that $\mathbb{P}$ is a "cloud" accommodates situations where the fluctuation in $y(t)$ is caused by other sources (noise sources) besides the explanatory variable $\varphi(t)$. ⋆

Remark 2 Stationarity simply says that the system is operating in steady-state. No assumption is made on the marginal $\mathbb{P}$ so that the structural or functional form relating $\varphi(t)$ to $y(t)$ can be arbitrary. The system can be e.g. linear corrupted by noise, nonlinear corrupted by noise, or anything else. ⋆

For the reliability results we shall develop in Section 4, the following additional hypothesis of independence is made on $x(t)$.

Assumption 2 (Independence) *The process $x(t) = (\varphi(t), y(t))$, $t = 1, 2, \ldots$, is an independent sequence.*

We underline that independence is just a technical additional assumption that we introduce for two reasons: (i) this basic setting permits one to better understand the ideas behind the theorems by focusing on conceptual aspects; (ii) admittedly, quantifying the IPM reliability is more involved in the dependent case. This latter assumption is relieved in Section 4.2 of this paper, where an extension to M-dependent processes is provided. In the independent case, time ordering is not significant and t can also be seen just as an index to enumerate the data.

The problem addressed in this paper is described as follows.

Problem (Reliable interval prediction) *Suppose a finite number N of data from the unknown process $x(t)$ have been observed, and call $D_N = \{\varphi(t), y(t)\}_{t=1,\ldots,N}$ the collection of these observations. We want to find a* rule *that when fed with $\varphi(N+1)$ returns an informative (i.e., not too large) interval I to which the next (unobserved) output $y(N+1)$ belongs with high probability. Moreover, this probability should be quantified only on the basis of the structure of the rule and of the number of observations, without further assumptions on the mechanism that generates the data.*

To achieve these goals we first introduce interval predictor models IPMs, which are the tools through which prediction intervals are generated. Then, we show that these models can actually "learn" from data, that is once the IPM has been "trained" on a batch D_N, it may fail to correctly predict future outcomes with an a-priori quantifiable probability only.

1.3 Structure of the paper

In Section 2, interval predictor models are introduced and the notion of reliability of such models is defined. The problem of identifying an interval predictor from data is the subject of Section 3. The main results on reliability assessment for interval predictors are given in Section 4, which also contains some remarks and comments on the general philosophy underlying the method. Illustrative numerical examples are presented in Section 5, and conclusions are finally drawn in Section 6. To keep the focus of the discussion on the main concepts and to improve readability, all technical proofs are given in the Appendix.

2 Interval predictor models

In this section, we introduce the key instrumental element of our approach, that is models that return an interval as output: Interval Predictor Models (IPMs). The origin of this kind of models has to be found in the theory of differential inclusions and set-valued dynamical systems (see e.g. [1], [2] and [3]). Interval models have been previously considered in other contributions along routes that are quite different from that of this paper. In [22] and [23], interval predictors identification is performed under certain a-priori Lipschitz conditions on the underlying system function. Utilizing this prior assumption allowed the authors of these papers to establish guaranteed results without resorting to any stationarity assumption. Of course, however, the results are reliant on the a-priori known bounds and this sets serious limitations to the applicability of the method. Set prediction has also been developed in [14], [15], and [17]. By combining prior feasible sets with observations, guaranteed regions for the state vector and the system parameters are obtained.

Some basic concepts on IPMs are recalled next. Then, a new type of parametric IPMs of use in the present paper is introduced in Section 2.1; see also [8].

An interval predictor model is simply a rule that assigns to each instance vector $\varphi \in \Phi$ a corresponding output interval in Y. That is, an IPM is a set-valued map

$$I : \varphi \to I(\varphi) \subseteq Y. \qquad (1)$$

In (1), φ is a regression vector containing explanatory variables on which the system output y depends, and $I(\varphi)$ is the prediction interval. We are interested in building IPMs such that, given an observed φ, $I(\varphi)$ is an informative interval containing y with high guaranteed probability. Output intervals are here obtained by considering the span of parametric families of functions, as detailed in the next section.

2.1 Interval models described in parametric form

Consider a family of functions M mapping Φ into Y, parameterized by a vector q ranging in some set $Q \subseteq \mathbb{R}^{n_q}$, i.e.

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

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

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

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

$$\mathscr{M} = \{y = \vartheta^T \varphi + e, \quad \vartheta \in \Theta \subseteq \mathbb{R}^n, \ |e| \le \gamma \in \mathbb{R}\}. \qquad (3)$$

In this case, $q = [\vartheta^T \ e]^T \in \mathbb{R}^{n+1}$ and $Q = \Theta \times [-\gamma, \gamma]$. A possible choice for the set Θ is a ball with center c and radius r:

$$\Theta = \mathscr{B}_{c,r} = \{\vartheta \in \mathbb{R}^n : \|\vartheta - c\| \le r\}. \qquad (4)$$

A more general choice for Θ is an ellipsoidal region:

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

where P is a symmetric positive definite matrix.

For the model structure (3),(4), given an instance φ, the interval output of the IPM obtained through equation (2) can be explicitly computed as

$$I(\varphi) = [c^T \varphi - (r\|\varphi\| + \gamma), c^T \varphi + (r\|\varphi\| + \gamma)]. \qquad (6)$$

To verify this, rewrite (4) as $\{\vartheta \in \mathbb{R}^n : \vartheta = c + \rho, \ \|\rho\| \le r\}$ and write

$$\begin{aligned} y = \vartheta^T \varphi + e = c^T \varphi + \rho^T \varphi + e \le c^T \varphi + r \frac{\varphi^T}{\|\varphi\|} \varphi + \gamma \\ = c^T \varphi + r\|\varphi\| + \gamma. \end{aligned}$$

Similarly, $y \ge c^T \varphi - r\|\varphi\| - \gamma$, leading to (6).

Similar considerations show that for the ellipsoidal model (3),(5) the interval is given by:

$$I(\varphi) = [c^T \varphi - (\sqrt{\varphi^T P \varphi} + \gamma), c^T \varphi + (\sqrt{\varphi^T P \varphi} + \gamma)]. \qquad (7)$$

2.1.1 Classes of IPMs

Note that a parametric IPM as defined in (2) is assigned once a set Q is given. For this reason, parametric IPMs shall be usually denoted by I_Q.

For identification purposes, we shall consider *classes* of parametric IPMs, among which the predictor model is selected. A class of parametric IPMs 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 uniquely determined by c, r and γ, and $\mathscr{Q}$ can e.g. be given by

$$\mathscr{Q} = \{Q = \mathscr{B}_{c,r} \times [-\gamma, \gamma] : \ c \in \mathbb{R}^n, r \ge 0, \gamma \ge 0\}, \qquad (8)$$

that is, $\mathscr{Q}$ is the family of all cylinders obtained by letting the spherical basis and height vary in all possible ways. Similarly, when $Q = \mathscr{E}_{c,P} \times [-\gamma, \gamma]$ we can choose

$$\mathscr{Q} = \{Q = \mathscr{E}_{c,P} \times [-\gamma, \gamma] : c \in \mathbb{R}^n, P \in \mathbb{S}_+, \gamma \ge 0\}, \qquad (9)$$

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

2.2 Reliability of IPMs

Recalling that $\mathbb{P}$ is the probability distribution in the space $\Phi \times Y$, we have the following definition.

Definition 1 (Reliability of an IPM) *Let I be a given IPM. The* reliability *of I is defined as*

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

that is $R(I)$ is the probability that the pair (φ, y) falls in the IPM.

Note that this definition refers to picking a random φ and a y such that y belongs to $I(\varphi)$; in other words, this notion is not conditional to a given φ.

2.3 An example of IPM

Assume that an output $y \in \mathbb{R}$ is generated according to the following data-generating mechanism:

$$y = y(\varphi) = \varphi \cdot (1 + |\varphi|), \quad \text{with } \varphi \in [-1, 1].$$

Suppose this mechanism is actually unknown, and consider a parametric IPM defined according to the following equation:

$$I(\varphi) = \{y : y = \vartheta \varphi, \quad \vartheta \in [1, 2]\}$$

(note that this is a particular instance of a predictor model as in (2),(3),(4)). The prediction interval $I(\varphi)$ can be explicitly computed according to (6), leading to $I(\varphi) = [1.5\varphi - 0.5|\varphi|, 1.5\varphi + 0.5|\varphi|]$. The map $I(\varphi)$ is depicted in Figure 1, where function $y(\varphi)$ is also represented. In this case, for each φ the output $y(\varphi)$ is contained in the predicted interval $I(\varphi)$, so that the reliability of the predicted interval is 100%.

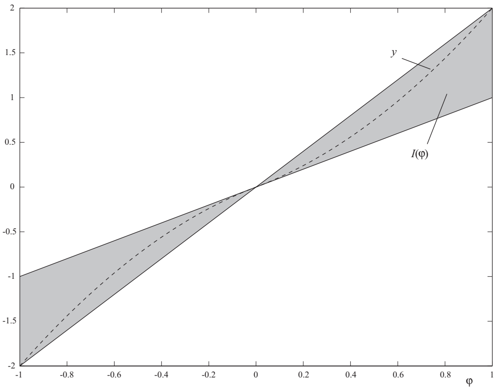

Figure 1. $I(\varphi)$ (textured region) and $y(\varphi)$ (dashed line).

Remark 3 Note that $y = \vartheta\varphi$ should not be considered as a function family from which a specific function has to be selected to represent the data-generating system. Instead, these parametric functions are merely an instrument through which the interval map $I(\varphi)$ is defined. As a matter of fact, $y(\varphi)$ does not belong to the family $y = \vartheta\varphi$, and yet the predicted interval $I(\varphi)$ always contains the output $y(\varphi)$ for suitable values of ϑ that depend on φ. $\star$

In this example, the φ to y map is deterministic. In more general situations the data-generating system is not a deterministic map. Rather it is a "cloud" in the $\Phi \times Y$ space and the interval map is used to describe the vertical dispersion of y.

3 Identification of IPMs

In this section we discuss the problem of selecting an interval predictor from a parametric class on the basis of criteria of consistency with the observed data and of optimality. This section is focused on the numerical techniques for constructing an IPM, whereas the fundamental issue of assessing the reliability of the IPM is deferred to Section 4.

Suppose that the observations

$$D_N = \{\varphi(t), y(t)\}_{t=1,\dots,N}$$

have been collected. Based on D_N we want to identify an interval predictor model $I_{\widehat{Q}_N}$ from a given class of parametric IPMs I_Q, $Q \in \mathcal{Q}$.

In this work, identification is guided by two different criteria as explained in the following.

First, we require $I_{\widehat{Q}_N}$ to be *consistent* with the available observations, according to the following definition.

Definition 2 *An IPM I is* consistent *with the batch of observations D_N if $y(t) \in I(\varphi(t))$, for $t = 1,\dots,N$.*

In other words, consistency means that the interval $I(\varphi(t))$ for the given $\varphi(t)$ is not falsified by the observed $y(t)$, for $t = 1,\dots,N$.

Clearly, consistency alone does not make $I_{\widehat{Q}_N}$ a good predictor. The second requirement on $I_{\widehat{Q}_N}$ is that it satisfies a tightness criterion, expressed in terms of the minimization of an index μ_Q measuring how wide the intervals returned by the IPM are. The tightness criterion should reflect the specific needs of the problem at hand and is model class dependent. For example, for parametric IPMs with ϑ in a ball defined by (3),(4), given a φ, the size of the predicted interval increases linearly with r and γ (see (6)), so that we may want to consider

$$\mu_Q = \alpha r + \gamma, \tag{10}$$

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

Combining the consistency requirement with the requirement of tightness, the identification of $I_{\widehat{Q}_N}$ can be formulated as the following constrained optimization problem (to ease the notation we shall write $\widehat{I}_N$ in place of $I_{\widehat{Q}_N}$ in the following).

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

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

Problem 1 might be hard to solve in general. However, for some standard IPM parameterizations and cost criteria, Problem 1 turns out to be a *convex* optimization problem, which can be solved at low computational effort.
For instance, for parametric IPMs with ϑ in a ball defined by (3),(4) with $\mathcal{Q}$ and μ_Q as in (8) and (10), Problem 1 becomes the following linear program (note that $Q = Q(c,r,\gamma)$ in this case).

Problem 1.a (spherical parameter set)
Find $\widehat{I}_N := I_{Q(\widehat{c}_N,\widehat{r}_N,\widehat{\gamma}_N)}$ such that

$$\widehat{c}_N, \widehat{r}_N, \widehat{\gamma}_N = \arg\min\nolimits_{c,r,\gamma} \alpha r + \gamma, \text{ subject to}$$

$$r, \gamma \geq 0$$
$$y(t) \geq c^T\varphi(t) - r\|\varphi(t)\| - \gamma, \quad t = 1,\dots,N$$
$$y(t) \leq c^T\varphi(t) + r\|\varphi(t)\| + \gamma, \quad t = 1,\dots,N.$$

A similar result also holds for the IPMs with ϑ in an ellipsoid defined by (3),(5) with $\mathcal{Q}$ as in (9) and

$$\mu_Q = \mathrm{Tr}[PW] + \gamma^2,$$

where W is a weighting matrix and $\mathrm{Tr}[\cdot]$ means trace. If e.g. $W = \mathbb{E}[\varphi(t)\varphi(t)^T]$ then μ_Q relates to the width of I_Q as follows:

$$\begin{aligned}
&\mathbb{E}[(\text{half width})^2] = [\text{see (7)}] = \mathbb{E}[(\sqrt{\varphi(t)^T P\varphi(t)} + \gamma)^2] \\
&\leq 2\mathbb{E}[\varphi(t)^T P\varphi(t) + \gamma^2] = 2\mathbb{E}[\mathrm{Tr}[\varphi(t)^T P\varphi(t)] + \gamma^2] \\
&= 2\mathbb{E}[\mathrm{Tr}[P\varphi(t)\varphi(t)^T]] + 2\gamma^2 = 2\mathrm{Tr}[P\mathbb{E}[\varphi(t)\varphi(t)^T]] + 2\gamma^2 \\
&= 2\mu_Q.
\end{aligned}$$

Notice that minimizing $\mathbb{E}[\text{half width}]$ is not suitable since this quantity is not convex in the optimization variables. In this case, as shown in [11], Problem 1 can be rewritten as follows ($\varepsilon_1, \dots, \varepsilon_N$ are slack variables).

Problem 1.b (ellipsoidal parameter set)
Find $\widehat{I}_N := I_{Q(\widehat{c}_N, \widehat{P}_N, \widehat{\gamma}_N)}$ *such that*

$$\widehat{c}_N, \widehat{P}_N, \widehat{\gamma}^2_N = \arg\min_{c,P,\gamma^2,\varepsilon_1,\dots,\varepsilon_N} \mathrm{Tr}[PW] + \gamma^2, \text{ subject to}$$

$$P \succ 0,$$

$$\begin{bmatrix} \gamma^2 & \varepsilon_t \\ \varepsilon_t & 1 \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} \varphi(t)^T P \varphi(t) & y(t) - c^T \varphi(t) - \varepsilon_t \\ y(t) - c^T \varphi(t) - \varepsilon_t & 1 \end{bmatrix} \succeq 0,$$

$$t = 1, \dots, N.$$

Problem 1.b is a convex semi-definite optimization problem, for which many efficient numerical solvers have been developed (see e.g. [7], [25]).

Numerical examples of IPM identification can be found in Section 5.

3.1 Identification with discarded constraints

It is well known that in some cases there can be some "exceptional" data points (the so-called *outliers*) whose value is anomalous as compared to other observations. 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, generating a wide identified model. In this case, a wiser procedure would be to discard "bad data", and use the remaining ones to do identification, [18], [4], and [16].

The presence of outliers is not the only reason justifying data discarding, though. Indeed, there are situations where one is willing to accept a decrease in prediction reliability in favor of a narrower interval model and, as we will show in this section, this can be obtained by discarding some data even when these data cannot be regarded as outliers. As an example, we may think of prediction of stock market returns or volatilities. Here, a 60-70% confidence prediction interval of small enough size may be more suitable than a 99% confidence prediction interval which is, however, too loose to reveal the future index trend.

From an optimization point of view, the IPM identification with discarded constraints can be outlined as follows. Let $k < N$ 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 that are used in identification. By $\widehat{I}^{\mathscr{A}}_{N,k}$ we denote the identified IPM when k constraints are removed as indicated by $\mathscr{A}$. Precisely:

Problem 1′ (IPM identification with k discarded constraints)
Find $\widehat{I}^{\mathscr{A}}_{N,k} := I_{\widehat{Q}^{\mathscr{A}}_{N,k}}$ *such that*

$$\widehat{Q}^{\mathscr{A}}_{N,k} = \arg\min_{Q \in \mathscr{Q}} \mu_Q,$$
$$\text{subject to} \quad y(t) \in I_Q(\varphi(t)),\ t \in \mathscr{A}(D_N).$$

Notice that, for $k = 0$, $\widehat{I}^{\mathscr{A}}_{N,0} \equiv \widehat{I}_N$, so that Problem 1 is a particular case of Problem 1′.

Two main issues now arise:

(i) How should $\mathscr{A}$ 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 postponed to Section 4, while point (i) is the subject of the next Section 3.2.

3.2 Choice of the data discarding algorithm $\mathscr{A}$

3.2.1 Greedy constraints removal

A straightforward approach to remove constraints is to select in succession those constraints which – if removed one by one – lead each time to the largest immediate improvement in μ_Q. This approach does not of course give the overall optimal result with k constraints removed. It has however the great advantage of being implementable at a low computational effort. Moreover, the reliability analysis of Section 4 is algorithm-independent, and rigorously applies to this greedy approach as well.

3.2.2 Optimal constraints removal

In order to achieve the best possible benefit from constraints removal, algorithm $\mathscr{A}$ should be chosen so to discard those constraints whose removal leads to the largest overall drop in the cost μ_Q. To this end, one can try to solve Problem 1 for all possible combinations of $N-k$ constraints taken out from the initial 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 demanding, since 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 as in [4] and [20], though in a different setting. We inform the reader that is not interested in computational aspects that he/she can jump from here to Section 4, where the key issue of reliability is discussed, without any loss of continuity.

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

$$\mathscr{P}: \text{ Find } \widehat{z} = \arg \min_{z \in Z \subseteq \mathbb{R}^d} f(z),$$
$$\text{subject to } z \in Z_t, \quad t = 1, \dots, N.$$

Existence and uniqueness of $\widehat{z}$ is taken here for granted. More generality can be achieved by extra technicalities. Let $w(\mathscr{P}) := f(\widehat{z})$ be the optimal value for problem $\mathscr{P}$. We have the following definition.

Definition 3 (Support constraint) *The l-th constraint Z_l is a* support constraint *for $\mathscr{P}$ if $w(\mathscr{P}_l) < w(\mathscr{P})$, where $\mathscr{P}_l$ is the optimization problem obtained from $\mathscr{P}$ by removing the l-th constraint, namely:*

$$\begin{aligned} \mathscr{P}_l: \text{ Find } \widehat{z}_l = \arg \min_{z \in Z \subseteq \mathbb{R}^d} f(z),& \\ \text{subject to } z \in Z_t,& \\ t = 1,\ldots,l-1,l+1,\ldots,N.& \end{aligned}$$

In other words, a support constraint is a constraint whose elimination improves the optimal solution. The following theorem is taken from [9]; see Theorem 2 and Section 4.3 in that reference.

Theorem 1 *If $\mathscr{P}$ is a convex optimization problem (i.e. if $f(z)$ is a convex function of z, Z is a convex set, and Z_t is a convex set for any t), then the number of support constraints for $\mathscr{P}$ is at most d, the number of optimization variables.*

We also need the following definition.

Definition 4 (Non-degenerate problem) *$\mathscr{P}$ is non-degenerate if $w(\mathscr{P}_{\rm sc}) = w(\mathscr{P})$, where*

$$\begin{aligned} \mathscr{P}_{\rm sc}: \text{ Find } \widehat{z}_{\rm sc} = \arg\min_{z \in Z} f(z),& \\ \text{subject to } z \in Z_t, \text{ for any } Z_t& \\ \text{that is a support constraint of } \mathscr{P}.& \end{aligned}$$

Thus, a non-degenerate problem is one such that the optimal solution with the sole support constraints in place is the same as the optimal solution with all constraints. A degenerate $\mathscr{P}$ is illustrated in the example below.

Example 1 Let

$$\begin{aligned} \mathscr{P}: \text{ Find } (\widehat{z}_1,\widehat{z}_2) = \arg \min_{(z_1,z_2) \in \mathbb{R}^2} z_2,& \\ \text{subject to } \quad (z_1,z_2) \in Z_a \cap Z_b \cap Z_c,& \end{aligned} \tag{11}$$

where Z_a, Z_b and Z_c are as in Figure 2.

In this case, only Z_a is a support constraint since removing Z_b or Z_c does not change the optimal solution. However, considering the optimization problem subject to Z_a only leads to a different solution than the original problem. ⋆

We now go back to the problem of optimally removing k constraints from the initial set of constraints D_N (with a little abuse of terminology, we say "constraints D_N" for "constraints generated by D_N"). We shall consider a sequence of optimization problems obtained from Problem 1 by removing some constraints from the initial set D_N. For each of these problems we assume that the optimal solution exists and is unique, and moreover that the problem is non-degenerate. While these assumptions can be relaxed (e.g. tie-break rules can be introduced to avoid non-uniqueness, see e.g. [9]), we prefer to maintain them to avoid unduly technical complications.

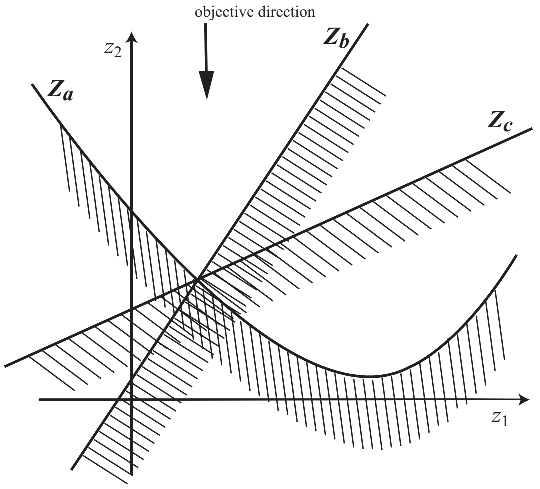

Figure 2. Constraints for the optimization problem (11)

For $F \subseteq D_N$, let $w(F)$ be the optimal value for Problem 1 with F in place of D_N. We will also denote by ${\rm sc}(F)$ and by ${\rm sc}_i(F)$ respectively the set of support constraints and the i-th support constraint of the problem with constraints F. Finally, suppose that Problem 1 is a convex problem (this is true e.g. for Problems 1.a and 1.b) so that $|{\rm sc}(F)| \leq d$, $\forall F \subseteq D_N$, according to Theorem 1 ($|\cdot|$ denotes cardinality).

The following Algorithm $\mathscr{A}^*$ optimally discards k constraints. The strength of this algorithm lies in that, instead of considering all possible combinations of $N-k$ constraints from the N initial 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 has to solve the optimization problems at level k in the tree (that is with k constraints removed).

The pseudo-code of the algorithm is 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).

Algorithm $\mathscr{A}^*$

```
0. $D^1_N := D_N$; $M_0 := 1$; $i := 0$;
1. $M_{i+1} := 0$
   FOR $h = 1$ TO $M_i$
   FOR $l = 1$ TO $|{\rm sc}(D^h_{N-i})|$
      $M_{i+1} := M_{i+1} + 1$
      $D^{M_{i+1}}_{N-i-1} := D^h_{N-i} - {\rm sc}_l(D^h_{N-i})$
   END
   END
2. IF $i+1 < k$ THEN $i := i+1$; GO TO 1.
   ELSE $\mathscr{A}^*(D_N) := D^r_{N-k}$ where $D^r_{N-k}$ is such that
   $w(D^r_{N-k}) \leq w(D^j_{N-k})$, $j = 1,\ldots,M_k$.
```

The optimality of Algorithm $\mathscr{A}^*$ is guaranteed by the following theorem.

Theorem 2 *Under the non-degeneracy assumption (Definition 4), algorithm $\mathscr{A}^*$ is optimal, in the sense that it returns a set of $N-k$ constraints resulting in the largest possible drop of the cost value μ_Q.*

Proof: see Appendix A.

We next provide an evaluation of the computational effort required to implement algorithm $\mathscr{A}^*$. The core of Algorithm $\mathscr{A}^*$ is the inner `FOR` loop where one support constraint at a time has to be removed from D^h_{N-i}. In order to spot the support constraints in D^h_{N-i} one tests all the constraints: each constraint is eliminated in turn and one checks whether the optimal solution improves.

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 (i.e. constraints whose boundary contains the solution of the optimization problem), $\mathrm{sc}(D^h_{N-i})$ can be determined by searching among active constraints of D^h_{N-i} only. This may further reduce the number of optimization problems to test significantly.

4 Reliability of IPMs

This section contains the main results of the paper. Here, we tackle the fundamental issue of quantifying the reliability $R(I)$ of the IPM (recall Definition 1) identified according to Problem $1'$. The reliability result applies to any constraints removal algorithm $\mathscr{A}$ and, in particular, to the greedy algorithm in Section 3.2.1, to the optimal algorithm $\mathscr{A}^*$ of Section 3.2.2, and, of course, to the particular case when no observations are removed.

A quantification of the reliability of the IPM will be given in the next two sections. Section 4.1 concentrates on an independent setting (e.g. data are generated by a static system fed by an independent input), while extensions to dependent settings are discussed in Section 4.2.

4.1 Independent observations

The following main theorem permits one to quantify the reliability of an IPM whenever the optimization Problem $1'$ used for its identification is convex (i.e. μ_Q and the constraints are convex in the optimization variables).

Theorem 3 *Let $x(t)=(\varphi(t),y(t))$, $t=1,2,\ldots$, satisfy Assumptions 1 and 2. Moreover, suppose that Problem $1'$ is a* convex *constrained optimization problem, and that its solution exists and is unique. Then, for any $\varepsilon\in(0,1)$ and $k<N-d$ (k is the number of constraints discarded by $\mathscr{A}$ and d is the number of optimization variables in Problem $1'$) it holds that*

$$\mathrm{Prob}_{\mathbb{P}^N}\{R(\widehat{I}^{\mathscr{A}}_{N,k})\geq 1-\varepsilon\}>1-\beta, \qquad (12)$$

where

$$\beta=\beta_0\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\cdot\frac{\varepsilon^i}{(1-\varepsilon)^i}, \qquad (13)$$

$$\beta_0=\frac{N!}{(N-d)!d!}(1-\varepsilon)^{N-d}, \qquad (14)$$

and $\mathbb{P}^N$ is the probability with which data $x(t)$, $t=1,\ldots,N$, are observed.

Proof: see Appendix B.

Theorem 3 is a "generalization" theorem, in the sense that the solution obtained by looking at N observations generalizes to unseen data. Precisely, the theorem states that the reliability of $\widehat{I}^{\mathscr{A}}_{N,k}$ is no worse than $1-\varepsilon$, with high probability greater than $1-\beta$. As for the probability $1-\beta$, one should note that $\widehat{I}^{\mathscr{A}}_{N,k}$ is a random element that depends on the observed realization $x(1),\ldots,x(N)$ of the stochastic process $x(t)$. Therefore, its reliability $R(\widehat{I}^{\mathscr{A}}_{N,k})$ can be greater than or equal to $1-\varepsilon$ for some random observations and not for others, and β 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 $1-\varepsilon$. Parameter ε is referred to as the "reliability parameter" while β is the "confidence parameter".

The confidence probability $1-\beta$ is the key to obtain results that are guaranteed independently of the data-generating system. Without this probability, a reliability result would certainly require some a-priori assumption on the data-generating mechanism. It is worth noticing that the confidence parameter can be pushed down to values such that the probability $1-\beta$ is so close to 1 that it loses any practical significance (so that $R(\widehat{I}^{\mathscr{A}}_{N,k})\geq 1-\varepsilon$ is for practical purposes guaranteed) and this is obtained without letting N increase too much. This is due to that β exponentially vanishes with N. See e.g. Table 1 for a numerical example.

We further remark that a confidence probability is common in many different contexts of classical probability theory, starting from the Glivenko-Cantelli theorem, [12], and going down to Vapnik-Chervonenkis uniform law of large numbers, [27], [28], [26].

For $k=0$, equation (13) tells us that $R(\widehat{I}_N)\geq 1-\varepsilon$ holds with confidence at least $1-\beta_0$; see (14). This bound for $k=0$ previously appeared in [10] in a robust control context. In the expression for β in (13), the term $\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\cdot\frac{\varepsilon^i}{(1-\varepsilon)^i}$ represents the confidence degradation due to the discarding of k observations.

A number of remarks on Theorem 3 are now in order.

Remark 4 The requirement that the optimization problem be convex is e.g. satisfied for the model classes used in Problems 1.a and 1.b, and further investigations are expected

N	500	600	700	800	900	1000
β	$4.1\cdot10^{-3}$	$1.3\cdot10^{-6}$	$3.2\cdot10^{-10}$	$5.5\cdot10^{-14}$	$7.5\cdot10^{-18}$	$8.8\cdot10^{-22}$

Table 1
β given by (13) with $\varepsilon = 0.1$, $d = 4$ and $k = 10$.

in the direction of determining other convex classes. We further note that convexity is not only the key property to obtain reliability results; it is also crucial in making the resulting optimization problem computationally tractable. ⋆

Remark 5 The relation (12) holds for given N and k. In certain applications one may want to let k vary with a fixed number of observations to meet a suitable balance between performance and reliability, or one may want to let N increase as the time horizon extends. If so, relation (12) can still be applied and the simultaneous satisfaction of the reliability results for different k and N holds with a confidence $1-\sum_j \beta_j$, where j runs over the different situations. Having a sum of β_j is not a hurdle since β_j is very small in normal situations. ⋆

Remark 6 The reader may wonder how the result in Theorem 3 is possible since, after all, reality has been inspected in correspondence of N points only and, since no assumptions are made on the data-generating system, reality can be anything elsewhere. The reason why this perhaps surprisingly result is possible relies on the role played by the probability $1-\beta$, a role that can be easily appreciated through a simple example.

Suppose that reality is represented by some function, and that the IPM does not correctly predict part of it. Two situations can occur. In the first situation in Figure 3(a), only a small part of reality is outside the IPM (in the terminology of this paper, $R(\widehat{I}_{N,k}^{\mathscr{A}}) \geq 1-\varepsilon$), and the IPM is reliable in the context of Theorem 3.

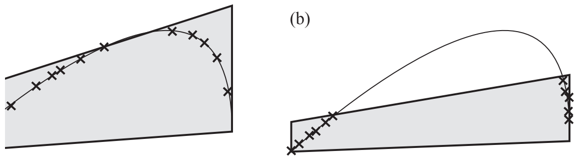

Figure 3. IPM vs. reality, 'x' = data

In the second situation (Figure 3(b)) a large part of reality is outside the IPM ($R(\widehat{I}_{N,k}^{\mathscr{A}}) < 1-\varepsilon$) so that the IPM is not reliable. For this situation to occur, however, data have to be confined where reality and the IPM agree. This happens only with small probability with respect to data extraction, and it is taken into account in Theorem 3 by the confidence parameter β.

To put it differently, it is true that, once the data have been collected, reality can be anything elsewhere so that a data-generating system which is consistent with the observed data and such that the IPM is not reliable can always be constructed. On the other hand, if the data-generating system had been the one we have constructed, we would hardly have seen data leading us to construct such an IPM. Theorem 3 provides a quantitative measure of this unlikelihood, in general situations and uniformly with respect to the probability measure $\mathbb{P}$.

In some more specific and quantitative terms, expression (14) for β_0 has the following intuitive motivation. β_0 represents an upper bound to the probability of obtaining an IPM whose reliability is $< 1-\varepsilon$ (poorly reliable IPM) if no data are discarded. Only few data determine the IPM (actually at most d, inspect the proof). Term $\frac{N!}{(N-d)!d!}$ is the number of possible choices of d data points out of N, that is the total number of potential IPMs. All other $N-d$ data have to be contained in the actually obtained IPM. When the reliability is $< 1-\varepsilon$, the probability of another point lying in the IPM is $< 1-\varepsilon$, so that the probability that all other $N-d$ data are in the IPM is $< (1-\varepsilon)^{N-d}$, and this generates the second term in (14). The degradation term $\sum_{i=0}^{k} \frac{(N-d)!}{(N-d-i)!i!} \cdot \frac{\varepsilon^i}{(1-\varepsilon)^i}$ in (13) to obtain β from β_0 accounts for the possibility of having data points outside the IPM. ⋆

Equation (13) is a fundamental relation linking the level of available information (represented by the number N of observations and the number k of discarded data), the complexity of the model (represented by the number d of decision variables in the IPM identification problem), and the probabilistic levels of reliability ε and confidence β.

In equation (13), the confidence parameter β is explicitly computed from ε, N, k, and d. Such equation, however, should be better thought of as a relation among five different variables $(\varepsilon,\beta,N,k,d)$, and making such a relation explicit with respect to a variable or another is a matter of convenience dictated by the application context. The interpretation of (13) when it is made explicit with respect to other variables than β is briefly discussed next.

- $N = N(\varepsilon,\beta,k,d)$
  This case is related to the design of an identification experiment, where the number of observations to be collected has to be chosen by the user;
- $\varepsilon = \varepsilon(\beta,N,k,d)$
  This is the most typical identification framework where data points are given and one would like to determine the prediction reliability of the identified IPM;
- $k = k(\varepsilon,\beta,N,d)$
  In this case, one wants to establish how many data points can be removed, without going below a chosen reliability level;
- $d = d(\varepsilon,\beta,N,k)$
  In this case, one evaluates the maximal complexity allowed for the explanatory model, for given ε, β, N, and k.

Though analytical expressions may be difficult to obtain in all the above cases, the inversion of equation (13) can be easily performed through numerical methods.

4.2 Dependent observations

Assuming independence in $x(t)$ is a condition that applies to many identification problems in econometrics, in pattern recognition and more generally in learning theory where no dynamics is present. On the other hand, it may be of interest to generalize results to a dependent context. Such generalization is still unavailable in any general form and we limit here to just discuss the case of M-dependent observations, [5].

Definition 5 (**M-dependent sequence**) *A (strict-sense) stationary stochastic sequence $x(t)$ is said to be M-dependent if $x(t)$ and $x(t+s)$ are independent random variables whenever $|s| > M$.*

M-dependence can be also regarded as an approximation of situations where the dependence between data points decays quickly enough so that it is negligible after M time instants.

We have the following theorem, which is here quoted without proof since the proof can be derived along similar lines as that of Theorem 3. It suffices here to say that, in order to test the reliability of the constructed IPM, in the proof of Theorem 4 one simply concentrates on data that are M instants apart from each other.

Theorem 4 *Let $x(t) = (\varphi(t), y(t))$, $t = 1,2,\ldots$, satisfy Assumption 1 and assume that it is an M-dependent sequence. Moreover, suppose that Problem 1′ is a* convex *constrained optimization problem, and that its solution exists and is unique. Then, for any $\varepsilon \in (0,1)$ and $k < N-d$ it holds that*

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

where

$$\beta = \frac{N!}{(N-d)!d!}\sum_{i=0}^{k}\frac{W!}{(W-i)!i!}\varepsilon^i(1-\varepsilon)^{W-i},$$

and $W = \lceil (N-d(2M+1))/(M+1) \rceil$.

4.3 *Reviewing the philosophy underlying IPM identification*

The reliability of the identified IPM is guaranteed by Theorem 3 for stationary and independent observations, with no further assumptions on the data-generating mechanism, that is on $\mathbb{P}$. On the other hand, as it is obvious, a-priori knowledge on what is being identified has certainly to play a role in the identification procedure. So, the question is: where does a-priori knowledge enter the picture in the theoretical approach of this paper? The answer is that a selection of an IPM model class which is suitably tailored to the structure of the data-generating mechanism generally leads to a narrower model, that is, one with smaller prediction intervals. Thus, the reliability of the IPM is always guaranteed and the a-priori knowledge impacts the other side of the coin, that is the width of the model. The point is that such a width can be assessed at the end of the identification procedure before the IPM is used.

5 Numerical examples

Two simple examples illustrate the idea of interval predictor models and IPM identification. The second example deals also with the presence of outliers.

5.1 *Example A*

Consider the static data-generating mechanism:

$$y(t) = \sin(2u(t)) + w(t), \tag{15}$$

where $u(t)$ is an i.i.d. (independent and identically distributed) sequence of random variables with uniform distribution in $U = [-1,1]$, and $w(t)$ is i.i.d. with normal distribution $\mathscr{N}(0, 0.01)$.

$N = 300$ observations $u(t), y(t)$ were generated according to (15) and used for identification. We took $\varphi(t) = [u(t)\ u^2(t)]^T$ as explanatory variable, and considered interval models in the form (2),(3),(4) with $n = 2$:

$$\begin{aligned} I(\varphi(t)) \ = \ \{y(t) :\ & y(t) = \vartheta_1 u(t) + \vartheta_2 u(t)^2 + e, \\ & \vartheta = [\vartheta_1\ \vartheta_2]^T \in \mathscr{B}_{c,r},\ |e| \leq \gamma\}, \end{aligned}$$

where $\mathscr{B}_{c,r}$ is the 2-dimensional ball with radius r and center c.

Setting $\mu_Q = 0.6r + \gamma$ (note that $\mathbb{E}[\|\varphi(t)\|] \approx 0.6$) and solving the linear Problem 1.a yielded

$$\widehat{c}_{300} = [1.2870\ \ 0.0220]^T;\ \widehat{r}_{300} = 0.0503;\ \widehat{\gamma}_{300} = 0.3839$$

as optimal solution. The obtained IPM is depicted in Figure 4 directly as a set-valued map from U to $Y = \mathbb{R}$. In the same plot, the available $u(t), y(t)$ data points are also represented.

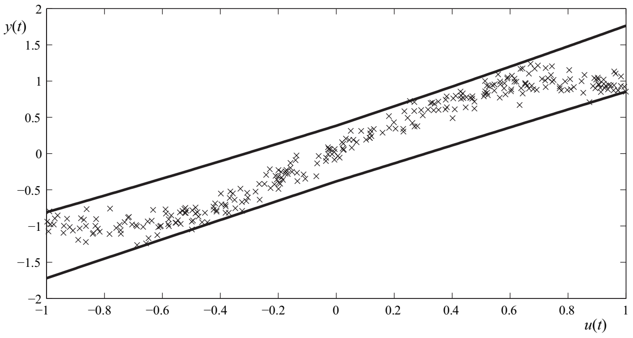

Figure 4. IPM (2-dimensional ϑ) and data points for Example A

Theorem 3 guarantees that the reliability of the identified model is no less than $1-\varepsilon = 0.92$, with high confidence $1-\beta = 0.999$ (equation (13) with $d = 4$).

On the other hand, the obtained IPM is loose. This is apparent from the blank areas between the cloud of data-points and the border of the interval prediction region, and reflects into the optimal cost value $\mu_{\widehat{Q}_{300}} = 0.4140$. A better description of reality can be achieved by suitably modifying the class of IPMs over which identification is performed.

By taking $\varphi(t) = [u(t)\ u^2(t)\ u^3(t)]^T$ and

$$\begin{aligned} & I(\varphi(t)) = \\ & \quad \{y(t) :\ y(t) = \vartheta_1 u(t) + \vartheta_2 u(t)^2 + \vartheta_3 u(t)^3 + e, \\ & \qquad\qquad \vartheta = [\vartheta_1\ \vartheta_2\ \vartheta_3]^T \in \mathscr{B}_{c,r},\ |e| \leq \gamma\}, \end{aligned}$$

with $\mu_Q = 0.67r + \gamma$ as a cost function ($\mathbb{E}[\|\varphi(t)\|] \approx 0.67$ in this case), we obtain

$$\widehat{c}_{300} = [1.9665 \;\; -0.0174 \;\; -1.1606]^T;$$
$$\widehat{r}_{300} = 0.0529; \;\; \widehat{\gamma}_{300} = 0.2320,$$

and the corresponding IPM is as depicted in Figure 5. The optimal cost turns out to be $\mu_{\widehat{Q}_{300}} = 0.2674$, with an almost 40% reduction with respect to the previous situation. Furthermore, Theorem 3 guarantees a reliability no less than $1-\varepsilon = 0.9$ with confidence $1-\beta = 0.999$, where the loss in reliability is due to the increase of the number of optimization variables from 4 to 5.

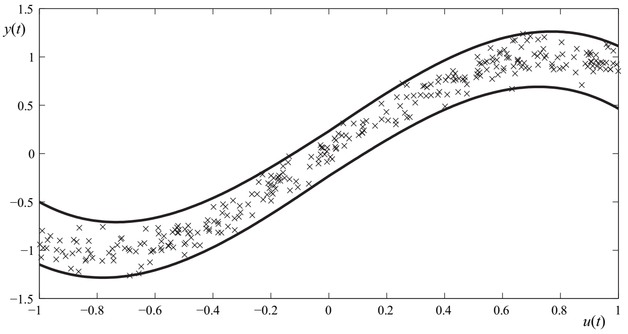

Figure 5. IPM (3-dimensional ϑ) and data points for Example A

5.2 *Example B*

Data were generated according to the following mechanism

$$y(t) = u(t)(1 + w_1(t)) + w_2(t),$$

where the i.i.d. signal $u(t) = \varphi(t)$ is the explanatory variable and has distribution $\mathscr{N}(0,1)$, $w_1(t)$ is i.i.d. with distribution $\mathscr{N}(0, 0.01)$, and $w_2(t)$ is a sequence of independent random variables taking values 0, +1, −1 with probability 0.98, 0.01 and 0.01 respectively. The sequence $w_2(t)$ is regarded as a source of outliers.

After collecting 300 observations $u(t), y(t)$, we sought an interval predictor model of the form (2),(3),(4) with $n = 1$, i.e.

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

Setting $\mu_Q = 0.8r + \gamma$, and solving Problem 1.a returned

$$\widehat{c}_{300} = 1.1204; \;\; \widehat{r}_{300} = 0.0453; \;\; \widehat{\gamma}_{300} = 0.9988;$$

and $\mu_{\widehat{Q}_{300}} = 1.0351$. The resulting set-valued map $I(\varphi(t))$ is depicted in Figure 6, along with the collected data points. As it appears, the identified IPM is loose because of the presence of outliers.

Suppose now that one had selected to remove $k = 10$ observations according to the optimal algorithm $\mathscr{A}^*$ of Section 3.2.2. Solving Problem 1′, the IPM depicted in Figure 7 was found, corresponding to

$$\widehat{c}^{\mathscr{A}^*}_{300,10} = 0.9724; \;\; \widehat{r}^{\mathscr{A}^*}_{300,10} = 0.1942; \;\; \widehat{\gamma}^{\mathscr{A}^*}_{300,10} = 0.0197;$$

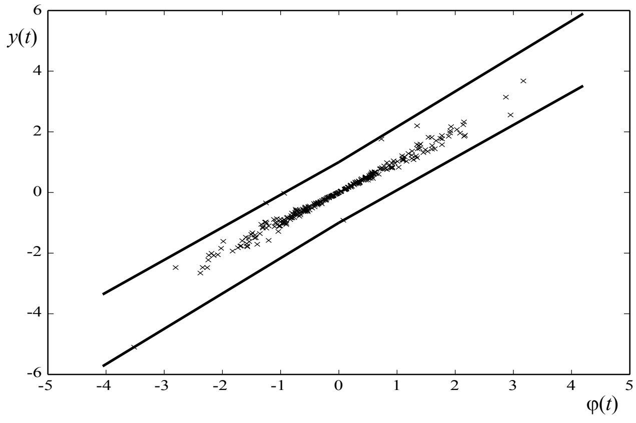

Figure 6. IPM and data points for Example B.

and $\mu_{\widehat{Q}^{\mathscr{A}^*}_{300,10}} = 0.1750$. Thus, discarding 10 observations yielded a 83% reduction of the cost μ_Q. As for the IPM reliability, Theorem 3 says that, with confidence at least equal to $1-\beta = 0.99$, $R(I)$ is no less than 0.935 if no constraints are removed and no less than 0.864 when $k = 10$ constraints are removed. Thus, the reliability loss is only 0.071.

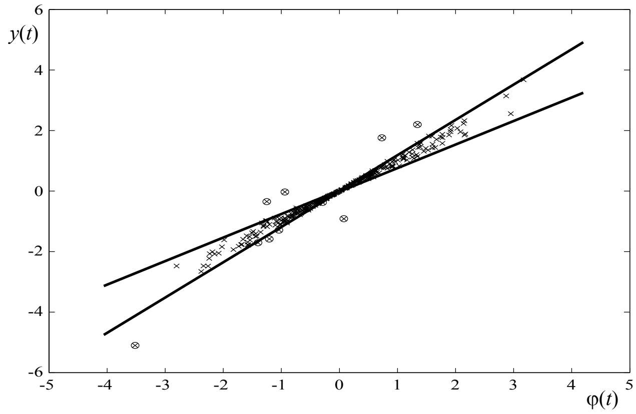

Figure 7. IPM and data points for Example B – $k = 10$ outliers removed.

6 Conclusions

In this paper, we discussed the identification and reliability analysis of interval models for prediction. From the computational side, we showed how to construct IPMs for some specific parametric classes, with the objective of minimizing a measure of the prediction interval while maintaining consistency either (a) with all of the observed data, or (b) with all except k of them. The first case is the easiest and can be efficiently handled via linear programming (in the spherical parameter case), or via convex semi-definite programming (in the ellipsoidal parameter case). In case (b), we proposed an optimal algorithm that alleviates the inherent combinatorial complexity of the partial consistency problem. Alternatively a greedy approach can be used.

From the theoretical side, we provided reliability guarantees given by way of an explicit, non-asymptotic, formula that relates the reliability to the degrees of freedom of the explanatory model class and to the number of available observations.

We believe that these results are bound to launch a new philosophical foundation in system identification. This new approach is here put on solid mathematical grounds and developed algorithmically for spherical and ellipsoidal models. On the other hand, the results are currently given in the stationary and independent or M-dependent setup, whereas the dynamic case requires further study.

Appendix

A Proof of Theorem 2

Let X be an optimal set of $N-k$ constraints which gives the largest drop of the cost criterion. Starting from the root of the tree of problems constructed by $\mathscr{A}^*$, generate a descending path as follows: at the root, eliminate a support constraint which is also an element in $D_N - X$ (if more than one support constraint exist in $D_N - X$, choose any one at will in this set) and move one level down to a problem with $N-1$ constraints. Then, again eliminate a support constraint which is also an element in $D_N - X$. Proceed this way until you get stuck, that is, no support constraints in $D_N - X$ to eliminate can be found, and let D^r_{N-l} be the constraints of the problem that has been reached. We then have:

$$\begin{aligned} w(X) &\geq w(\mathrm{sc}(D^r_{N-l})) && \text{(since } \mathrm{sc}(D^r_{N-l}) \subseteq X) \\ &= w(D^r_{N-l}) && \text{(thanks to non-degeneracy)} \\ &\geq w(X) && \text{(since } X \text{ is optimal),} \end{aligned}$$

so that equality holds throughout and $w(\mathrm{sc}(D^r_{N-l})) = w(X)$. Thus, an optimal problem is reached at some level of the tree, and this trivially entails that the leaves generated at level k from this problem are optimal too.

B Proof of Theorem 3

Let $\Delta = Y \times \Phi$ be the set where observations $x(t) = (\varphi(t), y(t))$ live and let $\Delta^N = \Delta \times \cdots \times \Delta$ the N-fold product of Δ's. Moreover, let $(x(1), \ldots, x(N))$ denote a generic element of Δ^N.

Select a subset $H = \{i_1, \ldots, i_d\}$ of d indexes from $\{1, \ldots, N\}$ and let $\widehat{I}_H$ be the optimal solution of the following optimization problem with d constraints only:

$$\min_{Q \in \mathscr{Q}} \mu_Q, \quad \text{subject to} \quad y(t) \in I_Q(\varphi(t)), \quad t \in H.$$

Based on $\widehat{I}_H$, introduce a subset $\Delta^{N,k}_H$ of Δ^N defined as follows:

$$\Delta^{N,k}_H = \{(x(1), \ldots, x(N)) \in \Delta^N : \widehat{I}_H = \widehat{I}^{\mathscr{A}}_{N,k}\}. \tag{B.1}$$

In other words, $\Delta^{N,k}_H$ contains those observations such that, if we apply algorithm $\mathscr{A}$ to them, we obtain the same IPM as if we optimized with constraints $x(i_1), \ldots, x(i_d)$ only.

Let now H range over the collection $\mathscr{H}$ of all possible choices of d indexes from $\{1, \ldots, N\}$ ($\mathscr{H}$ contains $N!/(N-d)!d!$ sets). We prove that:

$$\Delta^N = \bigcup_{H \in \mathscr{H}} \Delta^{N,k}_H. \tag{B.2}$$

Take any $(x(1), \ldots, x(N))$ in Δ^N and let $x(j_1), \ldots, x(j_{N-k})$, $\{j_1, \ldots, j_{N-k}\} \subset \{1, \ldots, N\}$, be the remaining constraints after that k constraints have been discarded according to algorithm $\mathscr{A}$. These $N-k$ constraints determine the optimal solution $\widehat{I}^{\mathscr{A}}_{N,k}$. From $x(j_1), \ldots, x(j_{N-k})$, eliminate a constraint which is not a support constraint (see Definition 3 for the notion of support constraint). This is possible since in view of Theorem 1 there are at most d support constraints and $N-k > d$. The resulting optimization problem with $N-k-1$ constraints has still $\widehat{I}^{\mathscr{A}}_{N,k}$ as optimal solution. Consider now the set of the remaining $N-k-1$ constraints, and among these, remove a constraint which is not a support constraint. Again, the optimal solution does not change. If we keep going this way, we are eventually left with d constraints and $\widehat{I}^{\mathscr{A}}_{N,k}$ is still the optimal solution. Thus, $(x(1), \ldots, x(N)) \in \Delta^{N,k}_H$ where H are the indexes we are left with at the end of the elimination procedure. Since this is true for any $(x(1), \ldots, x(N)) \in \Delta^N$, (B.2) is proven.

Consider now the following subsets of Δ^N:

$$B = \{(x(1), \ldots, x(N)) \in \Delta^N : R(\widehat{I}^{\mathscr{A}}_{N,k}) < 1 - \varepsilon\}$$

(i.e. B is the set of 'bad' observations which lead to an IPM which is not reliable as we would like it to be), and

$$B_H = \{(x(1), \ldots, x(N)) \in \Delta^N : R(\widehat{I}_H) < 1 - \varepsilon\}.$$

We have that:

$$\begin{aligned} B &= B \cap \Delta^N = [\text{using (B.2)}] = B \cap \left(\textstyle\bigcup_{H \in \mathscr{H}} \Delta^{N,k}_H\right) \\ &= \textstyle\bigcup_{H \in \mathscr{H}} (B \cap \Delta^{N,k}_H) = [\text{using (B.1)}] \\ &= \textstyle\bigcup_{H \in \mathscr{H}} (B_H \cap \Delta^{N,k}_H). \end{aligned}$$

A bound for $\mathrm{Prob}_{\mathbb{P}^N}\{B\}$ is now obtained by bounding $\mathrm{Prob}_{\mathbb{P}^N}\{B_H \cap \Delta^{N,k}_H\}$ first, and then summing over $H \in \mathscr{H}$.

Fix any H, e.g. $H = \{1, \ldots, d\}$ to be more explicit. Since the condition $R(\widehat{I}_H) < 1 - \varepsilon$ involves only the first d constraints, the set B_H is a cylinder with base in the cartesian product of the domains of the first d constraints. Fix now $(\bar{x}(1), \ldots, \bar{x}(d))$ in the base of this cylinder. For a point $(\bar{x}(1), \ldots, \bar{x}(d), x(d+1), \ldots, x(N))$ to belong to $B_H \cap \Delta^{N,k}_H$, at least $N-d-k$ constraints among $(x(d+1), \ldots, x(N))$ must be satisfied by $\widehat{I}_H$, for, otherwise, $\widehat{I}_H$ would satisfy less than $N-k$ constraints among $(\bar{x}(1), \ldots, \bar{x}(d), x(d+1), \ldots, x(N))$ and we would not have $\widehat{I}_H = \widehat{I}^{\mathscr{A}}_{N,k}$ as required by definition (B.1) of $\Delta_H^{N,k}$. Therefore, we have that:

$$\begin{aligned}
&\{(x(d+1),\ldots,x(N)):\\
&\quad(\bar{x}(1),\ldots,\bar{x}(d),x(d+1),\ldots,x(N))\in B_H\cap\Delta_H^{N,k}\}\\
&\subseteq\{(x(d+1),\ldots,x(N)):\\
&\qquad\text{at least } N-d-k \text{ constraints are satisfied by } \widehat{I}_H\}\\
&=\Omega_0\cup\Omega_1\cup\ldots\cup\Omega_k,
\end{aligned}$$

where Ω_i is the set where $N-d-i$ constraints among $(x(d+1),\ldots,x(N))$ are satisfied by $\widehat{I}_H$ and i are not.

Let $\zeta=\text{Prob}_{\mathbb{P}}\{y\notin\widehat{I}_H(\varphi)\}$. Then, thanks to the fact that observations are independent, we have that:

$$\begin{aligned}
&\text{Prob}_{\mathbb{P}^{N-d}}\{(x(d+1),\ldots,x(N)):\\
&\quad(\bar{x}(1),\ldots,\bar{x}(d),x(d+1),\ldots,x(N))\in B_H\cap\Delta_H^{N,k}\}\\
&\le\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\zeta^i(1-\zeta)^{N-d-i}\\
&<\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\varepsilon^i(1-\varepsilon)^{N-d-i}, \qquad \text{(B.3)}
\end{aligned}$$

where the latter inequality follows since $\text{Prob}_{\mathbb{P}}\{y\notin\widehat{I}_H(\varphi)\}>\varepsilon$ in B_H, and since $\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\zeta^i(1-\zeta)^{N-d-i}$ is a strictly decreasing function of ζ, as it can be checked by differentiation.

The probability on the left hand side of (B.3) is nothing but the conditional probability that $(x(1),\ldots,x(N))\in B_H\cap\Delta_H^{N,k}$ given $x(1)=\bar{x}(1),\ldots,x(d)=\bar{x}(d)$. Integrating over the base of the cylinder B_H we obtain:

$$\begin{aligned}
&\text{Prob}_{\mathbb{P}^N}\{B_H\cap\Delta_H^{N,k}\}\\
&<\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\varepsilon^i(1-\varepsilon)^{N-d-i}\cdot\text{Prob}_{\mathbb{P}^d}\{\text{base of } B_H\}\\
&\le\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\varepsilon^i(1-\varepsilon)^{N-d-i}.
\end{aligned}$$

Recalling that $B=\bigcup_{H\in\mathscr{H}}(B_H\cap\Delta_H^{N,k})$, the bound for $\text{Prob}_{\mathbb{P}^N}\{B\}$ sought after is finally obtained:

$$\begin{aligned}
&\text{Prob}_{\mathbb{P}^N}\{B\}\le\sum_{H\in\mathscr{H}}\mathbb{P}^N\{B_H\cap\Delta_H^{N,k}\}\\
&<\frac{N!}{(N-d)!d!}\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\varepsilon^i(1-\varepsilon)^{N-d-i}.
\end{aligned}$$

References

[1] J.P. Aubin. *Set-valued analysis*. Birkhäuser, Boston, MA, 1990.

[2] J.P. Aubin and A. Cellina. *Differential inclusions*. Springer-Verlag, Berlin, Germany, 1984.

[3] J.P. Aubin, J. Lygeros, M. Quincampoix, S. Sastry, and N. Seube. Impulse differential inclusions: a viability approach to hybrid systems. *IEEE Transactions on Automatic Control*, 47(1):2–20, 2002.

[4] E. Bai, H. Cho, and R. Tempo. Optimization with few violated constraints for linear bounded error parameter estimation. *IEEE Transactions on Automatic Control*, 47:1067–1077, 2002.

[5] D. Bosq. *Non parametric statistics for stochastic processes*. Springer, New York, NY, 1998.

[6] G. Box and G.M. Jenkins. *Times series analysis: forecasting and control*. Holden-Day, San Francisco, CA, 1970.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University press, Cambridge, UK, 2004.

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

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

[10] G. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51:742–753, 2006.

[11] G. Calafiore, M.C. Campi, and L. El Ghaoui. 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, 2002.

[12] K.L. Chung. *A course in probability theory*. Academic Press, San Diego, CA, USA, 1974.

[13] C.W.J. Granger and P. Newbold. *Forecasting economic time series*. Academic press, New York, NY, 1977.

[14] L. Jaulin, M. Kieffer, I. Braems, and E. Walter. Guaranteed nonlinear estimation using constraint propagation sets. *International Journal of Control*, 74(18):1772–1782, 2001.

[15] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*. Springer, London, UK, 2001.

[16] L. Jaulin and E. Walter. Guaranteed robust nonlinear minmax estimation. *IEEE Transactions on Automatic Control*, 47(11):1857–1864, 2002.

[17] M. Kieffer, L. Jaulin, and E. Walter. Guaranteed recursive nonlinear state bounding using interval analysis. *International Journal of Adaptive Control and Signal Processing*, 6(3):193–218, 2002.

[18] H. Lahanier, E. Walter, and R. Gomeni. OMNE: a new robust membership-set estimator for the parameters of non-linear models. *Journal of Pharmacokinetics and Pharmacodynamics*, 15(2):203–219, 1987.

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

[20] J. Matoušek. On geometric optimization with few violated constraints. *Discrete Computational Geometry*, 14:365–384, 1994.

[21] M. McAleer and M. Deistler. Some recent developments in econometrics. In S. Bittanti, editor, *Time series and linear systems*, volume 86 of *Lecture notes in control and information sciences*. Springer-Verlag, Berlin, Germany, 1986.

[22] M. Milanese and C. Novara. Set-membership identification of nonlinear systems. *Automatica*, 40(6):957–975, 2004.

[23] M. Milanese and C. Novara. Set-membership prediction of nonlinear time series. *IEEE Transactions on Automatic Control*, 50(11):1655–1669, 2005.

[24] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[25] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

[26] V.N. Vapnik. *Statistical learning theory*. John Wiley and sons, New York, NY, USA, 1998.

[27] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of probability and applications*, 16:264–280, 1971.

[28] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of the means to their expectations. *Theory of probability and applications*, 26:532–553, 1981.