Interval Predictor Models with a Formal Characterization of Uncertainty and Reliability

Luis G. Crespo[1], Daniel P. Giesy[2] and Sean P. Kenny[3]

***Abstract*— This paper develops techniques for constructing empirical predictor models based on observations. By contrast to standard models, which yield a single predicted output at each value of the model's inputs, Interval Predictors Models (IPM) yield an interval into which the unobserved output is predicted to fall. The IPMs proposed prescribe the output as an interval valued function of the model's inputs, render a formal description of both the uncertainty in the model's parameters and of the spread in the predicted output. Uncertainty is prescribed as a hyper-rectangular set in the space of model's parameters. The propagation of this set through the empirical model yields a range of outputs of minimal spread containing all (or, depending on the formulation, most) of the observations. Optimization-based strategies for calculating IPMs and eliminating the effects of outliers are proposed. Outliers are identified by evaluating the extent by which they degrade the tightness of the prediction. This evaluation can be carried out while the IPM is calculated. When the data satisfies mild stochastic assumptions, and the optimization program used for calculating the IPM is convex (or, when its solution coincides with the solution to an auxiliary convex program), the model's reliability (that is, the probability that a future observation would be within the predicted range of outputs) can be bounded rigorously by a non-asymptotic formula.**

I. INTRODUCTION

Model identification refers to the process of estimating the value and the uncertainty of the parameters of a mathematical model based on observed data. Several approaches to model identification are available [1], [2]. The most common technique is Bayesian inference. In this approach the objective is to calculate a probability density function (PDF) for the model's parameters using Bayes' rule [2]. The resulting probabilistic model, called the posterior PDF, depends on a prior PDF for p, and the likelihood function; which in turn depends on the set of observations available. In spite of its high computational demands, its inability to enforce key mathematical attributes to the posterior (e.g., independence), and of the potentially high sensitivity of the posterior to the assumed prior, this method is regarded as the benchmark.

This paper develops techniques for constructing IPMs having various characteristics. As in the Bayesian inference approach, the formulations proposed provide a crisp description of the uncertainty in the value of the model's parameters. In contrast to the Bayesian approach however, this prescription does not require any prior description of the uncertainty, the resulting uncertainty model is not probabilistic, and the spread in the predicted output and the model's reliability are both fully characterized. These properties, which are possible thanks to the assumed structure of the mathematical model, require for the output to depend linearly on the parameters and polynomially on the state, i.e., $y = p^\top \varphi(x)$, where $y \in \mathbb{R}^{n_y}$ is the output, $x \in \mathbb{R}^{n_x}$ is the model's *state* or input, $p \in \mathbb{R}^{n_p}$ is the model's parameter, and $\varphi(x) \in \mathbb{R}^{n_p}$ is a vector of monomials. Optimization-based strategies for calculating IPMs and eliminating the effects of outliers are proposed herein. This paper focuses on hyper-rectangular uncertainty sets, which have the advantage that the range of an individual parameter is not affected by the value taken by the others (in contrast to the ellipsoids studied in [3]). If the data-generating mechanism satisfies mild stochastic assumptions and the optimization program used to generate the model is convex, the model's reliability can be evaluated rigorously [3]. This paper extend this idea to non-convex formulations via the principle of equivalence, and proposes a strategy for the removal of the outliers that degrade the tightness of the prediction the most.

II. PROBLEM STATEMENT

A system is postulated to act on a vector of *state variables* to produce an *outcome*. The outcome depends on the values of the state variables and on some other influences, such as intrinsic variability and random noise, acting on the system to affect the value of the outcome. Let $X \subseteq \mathbb{R}^{n_x}$ be a set of state variables, and $Y \subseteq \mathbb{R}^{n_y}$ be a set of outcomes which might result from the system acting on elements of X.

In the following, the focus will be on the single-outcome ($n_y = 1$) multi-state/input ($n_x \geq 1$) case. It is desired to build a mathematical model of the data-generating mechanism which will predict the outcome corresponding to an unobserved value of the state. The inability to exactly model the data-generating mechanism, which might be subject to intrinsic variability, makes it unreasonable to build a model which will predict a single outcome value. Instead, an IPM will predict an interval into which unobserved data is expected to fall. Engineering judgment is used to pick a collection of monomials in the state variables, $\varphi(x)$, to use as basis functions for the mathematical model. There will be a model parameter for each of these monomials. Data points $z_i = (x_i, y_i)$ for $i = 1, \ldots, N$ are obtained from observations of the system. Instead of the standard practice of fitting all of the data as closely as possible with a single vector p of parameters, the thrust in this work is to restrict as much as possible a set in $\mathbb{R}^{n_p}$ from which p is chosen while, at the same time, having the property that all data

[1] Luis G. Crespo is with the NIA; MS 308, NASA Langley Research Center, Hampton, VA, 23681, USA (Luis.G.Crespo@nasa.gov)

[2,3]Daniel P. Giesy and Sean P. Kenny are with NASA Langley Research Center, Hampton, VA, 23681, USA

points (except, possibly, for a few outliers) can be fit *exactly* by *at least one* element in such a set.

The restriction considered herein forces p to belong to the hyper-rectangle P. For a fixed value of the state x, the propagation of P through $y = p^\top \varphi(x)$ yields an interval. The thrust here is to choose P to make the corresponding y intervals as small as possible and still allow each chosen (i.e., non-outlier) data point (x_i, y_i) to be modeled as $y_i = p^\top \varphi(x_i)$ for some $p \in P$.

In this setting the two problems of interest can be stated as follows. Suppose N observations $\mathsf{z} = \{z_i,\ i = 1, \ldots, N\}$, are available. First, we want to find an empirical model (or rule) that, when evaluated at a new value x_{N+1} of the state vector, returns an informative prediction of the unobserved output y_{N+1}. These empirical models, which are based on the observations comprising z, must meet a set of design requirements prescribed by the analyst (e.g., the predicted range of outputs must contain a percentage of the observations). Second, under additional stochastic assumptions on the sampling process from which the data is obtained, we want to quantify rigorously the probability that y_{N+1} will be compliant with such requirements (e.g., the probability that a new data point will fall outside the predicted range).

III. INTERVAL PREDICTION MODELS

An IPM is simply a rule that assigns to each instance vector $x \in X$ a corresponding outcome interval in Y. That is, an IPM is a set-valued map

$$I : x \to I(x) \subseteq Y, \tag{1}$$

where x is a state vector on which the system's output depends, and $I(x)$ is the prediction interval. Let M be any functional acting on a vector x of state variables and a vector p of parameters to produce an output y, i.e., $y = M(x, p)$. A parametric IPM is obtained by associating to each $x \in X$ the set of all possible outputs y that result from varying p over P:

$$I_y(x, P) = \{y : y = M(x, p) \text{ for all } p \in P\}. \tag{2}$$

$I_y(x, P)$ will be an interval as long as $M(x, p)$ is a continuous function of x and p, and P is a connected set. All instances of M and P considered in this paper will satisfy these restrictions.

Attention will be limited to the case where (i) the output is a linear function of the parameter p, (ii) the output is a polynomial function of the state x, and (iii) the uncertainty set P is the bounded hyper-rectangle:

$$P = \{p : p_{\min} \le p \le p_{\max}\}. \tag{3}$$

Hence, the corresponding IPM is given by

$$I_y(x, p_{\max}, p_{\min}) = \{y : y = p^\top \varphi(x) \text{ for all } p \in P\}. \tag{4}$$

where $\varphi(x)$ is a monomial. The analyst is free to choose which monomials are relevant to the particular application under analysis. A general representation of a multivariate polynomial basis is

$$\varphi(x) = \left[1, x^{i_2}, x^{i_3}, \ldots, x^{i_n}\right]^\top, \tag{5}$$

where $x = [x_1, \ldots, x_{n_x}]$ is the state, and the vector $i_j = [i_{j,1}, \ldots, i_{j,n_x}]$, with $i_j \neq i_k$ for $j \neq k$ has the exponents of the monomials. The inclusion of 1 in $\varphi(x)$ guarantees that every (x, y) pair will be interpolated using some p even if $x = 0$.

The limits of the output of the IPM prescribed by (4-5) can be explicitly computed as

$$I_y(x, p_{\max}, p_{\min}) = \left[\, \underline{y}(x, p_{\max}, p_{\min}), \overline{y}(x, p_{\max}, p_{\min}) \,\right], \tag{6}$$

where

$$\underline{y}(x, p_{\max}, p_{\min}) = \varphi(x)^\top \bar{p} - \varphi(|x|)^\top m, \tag{7}$$

$$\overline{y}(x, p_{\max}, p_{\min}) = \varphi(x)^\top \bar{p} + \varphi(|x|)^\top m, \tag{8}$$

$\bar{p} = (p_{\max} + p_{\min})/2$, and $m = (p_{\max} - p_{\min})/2$. Therefore, the envelopes of the interval valued function I_y, are linear functions of $p_{\min}$ and $p_{\max}$, and piecewise polynomial functions of the state. As such, they can possibly have derivative discontinuities on the coordinate hyperplanes $\{x \in X : x_i = 0\}$ for each $i = 1, \ldots, n_x$. The spread of I_y, which is the separation between its limits, is

$$\delta_y(x, p_{\max}, p_{\min}) = \varphi(|x|)^\top (p_{\max} - p_{\min}). \tag{9}$$

Note that the spread depends on the size of P, and not on its geometric center.

For a given IPM, we might want to evaluate the contribution of individual terms in M to the resulting model prediction. The contribution of the $\varphi_i(x)$ term is significant when either $\max_{x \in X}\{\varphi_i(|x|)(p_{\max,i} - p_{\min,i})\} \gg 0$ (the term contributes significantly to the predicted spread of the output) or $p_{\min,i} \approx p_{\max,i} \neq 0$ (the term affects the location of the interval value function). Terms not satisfying any of these conditions can be removed from M without degrading the accuracy of the prediction. These criteria can be used to evaluate the effectiveness of the model structure M assumed.

Commonly, the data-generating mechanism is approximated by the Least Square (LS) prediction, $y = p_{\text{LS}}^\top \varphi(x)$, where p_{LS}, the solution to the LS program $p_{\text{LS}} = \operatorname{argmin}_p \sum_{i=1}^N \left(y_i - p^\top \varphi(x_i)\right)^2$, is

$$p_{\text{LS}} = (A^\top A)^{-1} A^\top [y_1, \ldots y_N]^\top, \tag{10}$$

where $A_{i,j} = \varphi_j(x_i)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_p$. While the LS prediction describes the overall trend of the data, I_y describes its spread. Two types of IPMs are introduced next.

A. Type-1 IPMs

In this formulation we seek an IPM given by (4-5) with P given by the following Optimization Program (OP).

Optimization Program 1: The limits of P are given by

$$\begin{aligned}\langle \hat{p}_{\max}, \hat{p}_{\min} \rangle = \operatorname*{argmin}_{p_{\text{a}},\, p_{\text{b}}} \big\{ & E_x[\delta_y(x, p_{\text{b}}, p_{\text{a}})] : \underline{y}(x_i, p_{\text{b}}, p_{\text{a}}) \le y_i \\ & \le \overline{y}(x_i, p_{\text{b}}, p_{\text{a}}), p_{\text{a}} \le p_{\text{b}} \big\}, \end{aligned} \tag{11}$$

where $E_x[\cdot]$ is the expected value operator with respect to the state variable x, and (x_i, y_i) for $i = 1, \ldots, N$ are the observations.

In this formulation we search for the limits of the uncertainty box that minimize the expected interval spread such that all the observed responses are within the limits of the interval valued function I_y. When x is a random vector with a standard joint density function, the cost function in (11) can be calculated analytically. Otherwise, the sample mean of δ_y can be used to approximate it. The resulting IPM, which is calculated by solving the convex optimization problem in (11), admits a rigorous reliability assessment (see Section IV). This assessment formally quantifies the probability that a future observation will fall within $I_y(x)$.

Note that the formulation in (11) does not guarantee that $p_{\text{LS}} \in P$. Whereas the LS parameter estimate p_{LS} and the corresponding prediction $y = p_{\text{LS}}^\top \varphi(x)$ describe the overall trend of the data by weighting *all* data points equally, the set P and the corresponding interval valued function I_y describe their spread. This spread is driven by *extreme* observations. As such, there is no basis to expect that $p_{\text{LS}} \in P$ nor that $\underline{y}(x) \leq p_{\text{LS}}^\top \varphi(x) \leq \overline{y}(x)$ for all $x \in X$. The membership of p_{LS} in P can be ensured by replacing the last constraint in (11) with either $p_{\text{a}} \leq p_{\text{LS}} \leq p_{\text{b}}$ (i.e., P contains the LS solution), or $p_{\text{a}} + p_{\text{b}} = 2p_{\text{LS}}$ (i.e., P is centered about the LS solution). In general, the inclusion of these constraints will lead to IPMs with larger expected spreads, with the equality constraint leading to the larger of the two. The formulation resulting from adding one of these two sets of constraints will be called the *Augmented OP1*.

Techniques for making the model prediction tighter based on the identification and removal of outliers in the data set are presented next.

The solution of (11) is driven by extreme data points that might significantly depart from the vast majority of the observations. Techniques for refining IPMs based on the identification and removal of outliers, applicable to the IPMs presented here, are proposed in [3]. These approaches require the solution of a combinatorial number of optimization problems, and as such, their implementation can be computationally expensive. An alternative approach for refining Type-I IPMs is presented next. The presence of outliers in the data yields unnecessarily large δ_y's and P's. Smaller spreads are obtained if outliers are removed from the data sequence used in (11). Outliers are the observation points (x_i, y_i) for some i's between 1 and N, whose removal from the data sequence yields an IPM with a considerably tighter prediction. Prospective outliers are identified by determining the observations for which both (i) the value(s) of p required to match the data point yields an excessively and comparatively large P, and (ii) the offset between the least-square prediction and the observed outcome is comparatively large. In regard to the first criterion, note that there are infinitely many points in p-space for which $y_i = p^\top \varphi(x_i)$. The separation between the center of P, $\bar{p}$, and the hyperplane $\{p : y_i = p^\top \varphi(x_i)\}$ is

$$\rho_i(p_{\max}, p_{\min}) = \left| \frac{2y_i - \varphi(x_i)^\top (p_{\max} + p_{\min})}{\varphi(|x_i|)^\top (p_{\max} - p_{\min})} \right| \|m\|. \quad (12)$$

The metric ρ_i is the length of the semi-diagonal of the smallest hyper-rectangle oriented as P (i.e., same center and diagonal orientation) for which either $y_i = \underline{y}(x_i, u, v)$ or $y_i = \overline{y}(x_i, u, v)$, where $u = \bar{p} + \rho_i m/\|m\|$ and $v = \bar{p} - \rho_i m/\|m\|$. Hence, the metric ρ_i evaluates the extent by which the data point (x_i, y_i) contributes to the spread of P. In this setting, the observation (x_i, y_i) satisfies the first selection criteria when $\rho_i(\hat{p}_{\max}, \hat{p}_{\min}) \approx \|m\|$. The empirical Cumulative Density Function (CDF) of ρ, F_ρ, is defined by $F_\rho(r) = j/N$, where the inequality $\rho_i \leq r$ holds for j of the N values. Prospective outliers are identified by calculating this CDF and determining the observations for which $F_\rho(\rho_i) > \lambda_\rho$ for $0 \ll \lambda_\rho < 1$. For instance, if $\lambda_\rho = 0.95$ the observations for which ρ is in the highest 5% will satisfy the first selection criterion[1].

The second criterion is based on the prediction error $e_i = \left(y_i - p_{\text{LS}}^\top \varphi(x_i)\right)^2$. As before, prospective outliers are identified by calculating the empirical CDF of e based on the N observations, F_e, and determining the observations for which $F_e(e_i) > \lambda_e$ for $0 \ll \lambda_e < 1$. Observations satisfying both criteria will be removed from the data sequence and a new Type-I IPM will be calculated. The effectiveness of the procedure can be evaluated by monitoring the value of the cost function in (11) before and after the removal of outliers.

The presence of outliers is not the only the reason for discarding data. There may be situations where one is willing to accept a reduction in the model's reliability for the sake of a tighter model prediction. The model's reliability and performance, which are evaluated using the developments of Section IV and the cost function $E_x[\delta_y]$ respectively, should be traded off until the desired balance is attained.

Example 1: Consider the data-generating mechanism:

$$y = x^2 \cos(x) - \sin(3x) e^{-x^2} - x - \cos(x^2) + xg, \quad (13)$$

where $x(t)$ is an independent and identically distributed (IID) sequence of random variables with uniform distribution over $X = [-5.5, 5.5]$, and $g(t)$ is the IID with a standard normal distribution. A data sequence z for $N = 150$ observations was generated using (13). We will calculate IPMs based on this sequence having the structure in (4-5) for $n_p = 7$. Therefore, P will be a hyper-rectangular set in the seven dimensional parameter space. The 150 observations are marked with $\times$'s in Figure 1. The prediction corresponding to the LS solution corresponding to a six-order polynomial, for which $p_{\text{LS}} = [-0.8734, -1.1059, -0.9926, 0.0026, -0.0228, -0.0004, 0.0028]^\top$, is shown as a blue solid line.

Recall that no knowledge of the data-generating mechanism is required to calculate IPMs. A Type-1

[1]Note that F_ρ is a piecewise constant function that can only take on multiples of $1/N$ between 0 and 1. As such, the inequality $F_\rho(r) \leq \lambda$ holds for a range of r values.

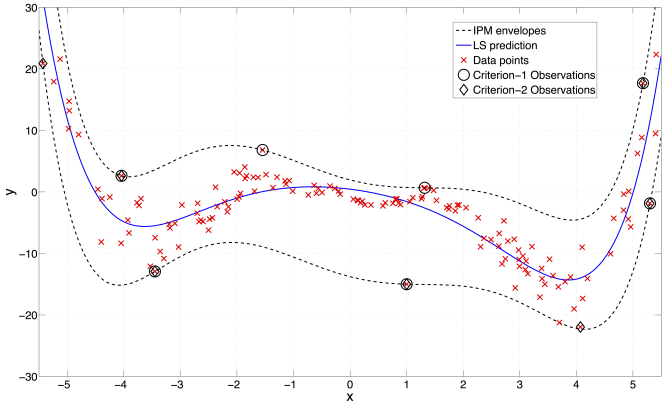

Fig. 1. IPM A: Type-1 IPM for all N observations.

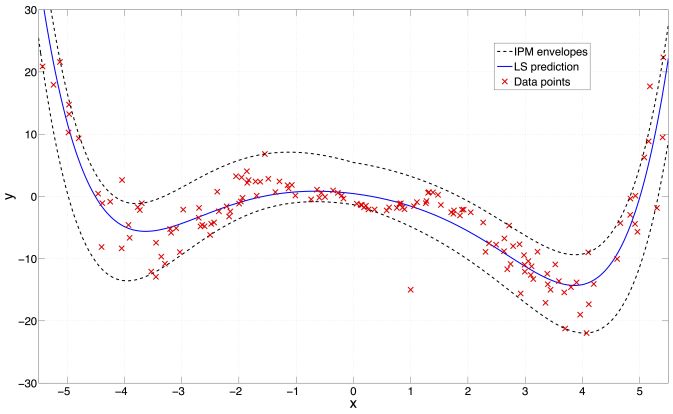

Fig. 2. IPM B: Type-1 IPM after the removal of outliers.

IPM, to be referred to as IPM A, based on the 150 observations was calculated first. The limits of the corresponding IPM are shown as dashed lines. Note that all observations fall in between the envelopes. The uncertainty set P corresponding to this IPM is bounded by $\hat{p}_{\min} = [-13.7738, -2.3647, 1.1603, 0.1666, -0.1899, -0.0046, 0.0061]^\top$ and $\hat{p}_{\max} = [1.9543, -2.3647, 1.1603, 0.1666, -0.1899 - 0.0046, 0.0066]^\top$. Note that the spread in the output is mostly caused by parameter variation in the coefficients of the constant and six-order terms (see paragraph preceding Section III-A). Note also that $p_{\text{LS}} \notin P$. The propagation of all the elements in P through (4) yields a family of infinitely many six-order polynomials lying in between the IPM limits. These envelopes are not members of the family, i.e., there is no parameter realization of P whose output is an envelope. This highlights the impossibility of identifying the IPM envelopes by identifying the worst-case parameter realizations within P. The expected spread for IPM A, which is the cost being minimized in the optimization problem, is $E_x[\delta_y] = 8.61$. This metric, which is as an indicator of the model's performance, will be used to assess the tightness of the prediction against other IPMs.

The strategy for identifying outliers based on preprocessing the data sequence is applied next. Recall that an observation will be regarded as an outlier if the values of ρ and e are both in the upper quantiles of the corresponding CDFs. The observations satisfying the selection criteria for $\lambda_\rho = \lambda_e = 0.95$ are indicated in Figure 1. Note that not all the samples near the IPM envelopes are outliers. The five samples satisfying both criteria were removed from the set of 150 and a new Type-1 IPM, to be denoted as IPM B, was calculated. The envelopes of IPM B are shown in Figure 2. The corresponding P is by $\hat{p}_{\min} = [-1.0721, -3.6824, -0.8710, 0.1667, -0.0500, -0.0053, 0.0037]^\top$ and $\hat{p}_{\max} = [3.9264, -1.3838, -0.8362, 0.1667, -0.0500, -0.0053, 0.0037]^\top$. The reduction in the interval spread δ_y is apparent. In this case, parameter variations in the constant, linear and second order terms are the most important. Note that the upper limit of the interval valued function has a derivative discontinuity at $x = 0$ as predicted by (7-8). As before, the LS solution is outside P but the LS prediction is between the IPM limits over the entire X range. This is the case even though the LS prediction is not a member of the family of polynomials associated with the IPM. The performance of IPM B is $E_x[\delta_y] = 5.87$, which is 32% better than that of IPM A. In terms of the size of P, measured by $\|\hat{p}_{\max} - \hat{p}_{\min}\|_2$, IPM B is 65% better than IPM A. Hence, the removal of only five outliers led to a significant improvement in performance.

B. Type-2 IPMs

A formulation leading to an alternative IPM is presented next. In contrast to Type-1 IPMs, this approach searches for P by only using a fixed percentage of the observations. The observations comprising the removed set, whose members can be regarded as outliers, are worst-case in the sense that their removal tightens the model prediction the most. In particular, we seek an IPM given by (4-5), where P is given by the following OP.

Optimization Program 2: The limits of P are given by

$$\langle \hat{p}_{\max}, \hat{p}_{\min} \rangle = \underset{p_{\text{b}},\, p_{\text{a}}}{\text{argmin}} \left\{ E_x \left[\delta_y(x, p_{\text{b}}, p_{\text{a}}) \right] : \right. \tag{14}$$
$$\left. F_{\rho(p_{\text{b}}, p_{\text{a}})} \left(\|p_{\text{b}} - p_{\text{a}}\|/2 \right) \geq \lambda,\ p_{\text{b}} \geq p_{\text{a}} \right\},$$

where $F_{\rho(p_{\text{b}}, p_{\text{a}})}$ is the empirical CDF of $\rho(p_{\text{b}}, p_{\text{a}})$, and $0 < \lambda \leq 1$ is the fraction of observations to be enclosed by I_y.

In this formulation we identify the uncertainty set P leading to the interval valued function I_y of minimal spread that contains $100\lambda\%$ of the observations. This makes the IPM insensitive to the $100(1-\lambda)\%$ of the observations for which ρ is the largest. Observe that any box with corners at p_a and p_b satisfying the inequality constraints contains parameters that interpolate at least $100\lambda\%$ of the observations. At the optimum, this fraction is as close to $100\lambda\%$ as possible (i.e., the first inequality becomes an equality), and the largest value of ρ among those corresponding to the $100\lambda\%$ observations retained is as small as possible. The tightening of the prediction for $100\lambda\%$ of the observations yields an empirical model that does not enclose the remaining $100(1-\lambda)\%$ of them. This shows that (14) is a chance-constraint formulation [4], in which one is willing to accept

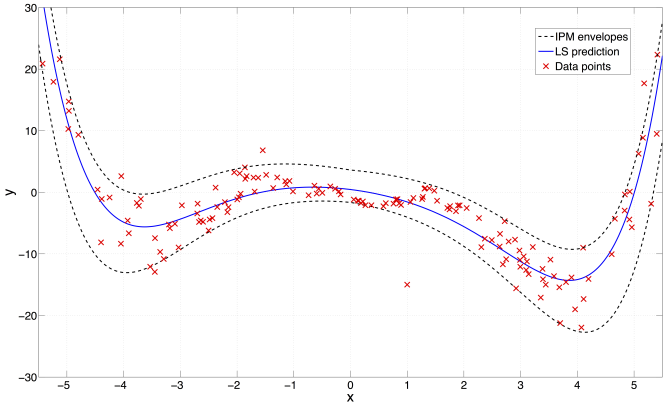

Fig. 3. IPM C: Type-2 IPM for $\lambda = 29/30$.

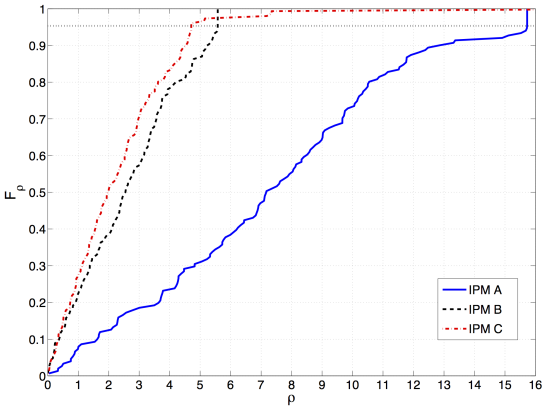

Fig. 4. Empirical CDFs $F_{\rho(\hat{p}_{\max},\,\hat{p}_{\min})}$ for IPM A, IPM B and IPM C.

the occurrence of unfavorable low-probability events for the sake of an improved performance for high-probability events.

The limits $\hat{p}_{\max}$ and $\hat{p}_{\min}$, thus P, depend on the value of λ chosen. To make this dependency explicit, the left hand side of (6) is written as $I_y(x, p_{\max}(\lambda), p_{\min}(\lambda))$.

Denote by $\mathsf{w} = \{w_j\}$ with $j = 1, \ldots, \text{floor}[\lambda N]$ the elements of the sequence z that are within $I_y(x, \hat{p}_{\max}(\lambda), \hat{p}_{\min}(\lambda))$, (i.e., the data points that were not considered outliers). When $\lambda = 1$, $\mathsf{w} = \mathsf{z}$, OP1 and OP2 are equivalent, and the resulting Type-1 and Type-2 IPMs are equal. This is a consequence of the first inequality constraint in (14) being equivalent to $\underline{y}(x_i, p_{\max}, p_{\min}) \leq y_i \leq \overline{y}(x_i, p_{\max}, p_{\min})$ for all $i = 1, \ldots, N$. When $\lambda < 1$, w will be a subset of z. Notice that the formulation in (14) sizes an IPM without prescribing in advance which particular points in z will be outliers. Outliers can be identified by determining the data points (x_i, y_i) for which $F_{\rho(\hat{p}_{\max},\hat{p}_{\min})}(\rho_i) > \lambda$.

Example 2: In this example we calculate a Type-2 IPM for the very same $N = 150$ observations used in Example 1 and $\lambda = 29/30$. Hence, we seek an IPM of minimal spread whose envelopes contain 145 observations, as it was the case for IPM B. Figure 3 shows the envelopes of the resulting IPM, to be referred to as IPM C. The uncertainty set of IPM C is bounded by $\hat{p}_{\min} = [-1.8967, -2.4580, -0.3339, 0.0576, -0.1112, -0.0015, 0.0051]^\top$ and $\hat{p}_{\max} = [2.5118, -0.8003, -0.3104, 0.0850, -0.1093, -0.0015, 0.0052]^\top$. In this case all but the fifth-order term contribute to the spread of the output. Note that IPM B and IPM C, which are both considerably tighter than IPM A, exclude a different set of outliers. The outliers are the observations shown outside the envelopes. The comparison of the three IPMs indicates that the envelopes of IPM C enclose the bulk of the observations most tightly. In addition, recall that Type-2 IPMs identify and eliminate the effect of outliers while the IPM is being calculated, and as such, they don't have to be chosen and removed in advance. Figure 4 shows the empirical CDF of ρ corresponding to IPM A, IPM B and IPM C. Note that $F_{\rho(\hat{p}_{\max},\,\hat{p}_{\min})}$ for IPM A, IPM B and IPM C have a steep vertical jump at $\rho = 15.72$, $\rho = 5.50$ and $\rho = 4.71$ respectively. These values correspond to $F_\rho = 1$ for both IPM A and IPM B, and to $F_\rho = \lambda$ for IPM C. These probability jumps, whose values are 6, 5, and 2% respectively, indicate the percentage of observations for which there is only one parameter point p for which $y_i = p^\top \varphi(x_i)$ and that point lies on the surface of P. The concentration of p values on the surface of P is a consequence of obtaining IPMs for which the interval spread, thereby the size of the corresponding uncertainty set, is minimal. Note that for IPM C, 96.66% of the observations attain a value of ρ which is less than 4.71. This is 30% of the largest value for ρ attained by IPM A and 85% of that by IPM B. This shows that IPM C concentrates $100\lambda\%$ of the observations much closer to the geometric center of P than either IPM A or IPM B. Note however that the largest value of ρ attained by IPM C, which is out of the range shown in the figure, exceeds that of the other two models. While $100\lambda\%$ of the observations are bounded more tightly by IPM C, the value of ρ for the remaining $100(1-\lambda)\%$ has increased considerably. That is often the price of enforcing chance constraints. This can be restated as follows. Pick K between 1 and N so that $K = \text{argmax}_{1 \leq i \leq N}\{\rho_i\}$, so (x_K, y_K) is the outlier with the largest ρ. Also, pick L between 1 and N so that $L = \text{argmax}_{1 \leq i \leq N}\{\rho_i : F_\rho(\rho_i) \leq \lambda\}$, so that (x_L, y_L) has the largest ρ among the retained data points. The smaller ρ_L, the tighter the prediction for $100\lambda\%$ of the observations. The larger $\rho_K - \rho_L$, the better the performance improvement resulting from removing outliers.

IV. MODEL'S RELIABILITY

The developments that follow are based on the scenario approach of Calafiore and Campi [5]. Denote by $\mathbb{P}$ the *unknown* distribution of the process from where the data pairs (x_i, y_i) are obtained. $\mathbb{P}$ can be interpreted as a probabilistic cloud in the $X \times Y$-space. The case in which y is a function of x only is a particular case where $\mathbb{P}$ is concentrated over the function. A general $\mathbb{P}$ can accommodate situations where the fluctuation in the outcome is caused by sources other than x. No assumption is made on $\mathbb{P}$ so that the functional form relating x and y can be arbitrary.

The reliability of the IPM $\mathcal{E}$ is defined as

$$r(\mathcal{E}) = \mathrm{Prob}_{\mathbb{P}}\left[(x,y) \in I_y\left(x, \hat{p}_{\max}(\lambda), \hat{p}_{\min}(\lambda)\right)\right], \quad (15)$$

where $\mathrm{Prob}_{\mathbb{P}}[\cdot]$ is the probability operator based on the distribution $\mathbb{P}$. Hence, $r(\mathcal{E})$ is the probability that the unobserved state-outcome pair (x, y) will fall within the limits of I_y. Recall that $\lambda = 1$ corresponds to Type-1 IPMs whereas $\lambda < 1$ corresponds to Type-2 IPMs. The following theorem, taken from [3], permits quantifying the reliability of an empirical model whenever the optimization problem used for its calculation is convex.

Theorem 1: Let $\mathsf{z} = \{z_i\} = \{(x_i, y_i)\}$, $i = 1, \ldots, N$ *be an independent data sequence resulting from a stationary discrete-time data generating process. Suppose the IPM* $\mathcal{E}$ *is calculated by solving a convex constrained optimization problem having a unique solution. Furthermore, assume that* k *observations (outliers) out of the* N *available have been discarded when calculating the model. Then, for any* $\epsilon \in (0, 1)$ *and* $k < N - d$*, where* d *is the number of optimization variables, it holds that*

$$\mathrm{Prob}_{\mathbb{P}^N}\left[r(\mathcal{E}) \geq 1 - \epsilon\right] > 1 - \beta, \quad (16)$$

where

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

and $\mathbb{P}^N$ *is the probability of obtaining the data sequence.*

This Theorem provides an assessment on unobserved data. The theorem states that the reliability of $\mathcal{E}$ is no worse than $1-\epsilon$ with probability greater than $1-\beta$. As for the probability $1 - \beta$, one should note that $\mathcal{E}$ is a random element that depends on N observations of $\mathbb{P}$. Therefore, its reliability can be greater than or equal to $1-\epsilon$ for some random observations but not for others, and β refers to the probability $\mathbb{P}^N = \mathbb{P} \times \cdots \times \mathbb{P}$ of observing a bad set of N samples such that the reliability of the model is less than $1 - \epsilon$. Parameter ϵ is referred to as the reliability parameter while β is the confidence parameter. The confidence probability $1 - \beta$ is key for obtaining results that are guaranteed independently of the data-generating mechanism. It is worth noting that the confidence parameter can be made very small such that it losses any practical significance and $r(\mathcal{E}) \geq 1 - \epsilon$. This can be done without letting N be too large because β vanishes exponentially fast with N. Note that assessing the reliability of the model does not require knowing $\mathbb{P}$.

The convexity of the OP1 enables the direct application of Theorem 1 to Type-1 IPMs. This includes the cases in which none ($k = 0$) and some ($k > 0$) of the observations are removed. In contrast to OP1, OP2 is non-convex, thus Theorem 1 cannot be applied directly to Type-2 IPMs. However, the reliability of such models can be evaluated by using a *Principle of Equivalence*. This principle is based on identifying an auxiliary convex formulation that will result in the very same empirical model found by solving the non-convex formulation. If this is attained, the reliability of the empirical model, which is independent of the means used to calculate it, can be rigorously evaluated via the auxiliary formulation. This approach can be applied to Type-2 IPMs. In particular, the solution to OP2 using the the data sequence z is equivalent to the solution of OP1 for the data sequence w[2]. Because only the $N - k^*$ elements in w, where

$$k^* = \mathrm{floor}[N(1-\lambda)], \quad (17)$$

are required by the auxiliary program, the reliability of Type-2 IPMs is given by (16) with $k = k^*$. These k^* observations fall outside I_y and satisfy $F_{\rho(\hat{p}_{\max}, \hat{p}_{\min})}(\rho_i) > \lambda$.

Example 3: The reliability of IPM A, for which $N = 150$, $d = 14$, and $k = 0$, is no less than $1 - \epsilon = 0.6984$ with confidence $1-\beta = 0.99$. The reliability of IPM B, for which $N = 150$, $d = 14$ and $k = 5$, is no less than $1 - \epsilon = 0.614$ with confidence $1 - \beta = 0.99$. Hence, the exclusion of five outliers rendered an improvement in the system performance of 32% at the expense of a reduction in the reliability of 8.4%. As for IPM C, note that $N = 150$, $d = 14$ and $k = k^* = 5$. While the reliability of IPM B and IPM C are the same, the performance of the latter is 10% better. The above results illustrate the typical trade-off between performance and reliability. These two figures of merit can be traded off by changing the number of outliers (i.e., λ) or the the model's structure (i.e., n_p).

A few of remarks on the significance and practical use of IPMs are now in order. In real applications, the difference between outliers and meaningful data is often unclear. In this regard, the proposed approach provides a probabilistic certificate of the predicted range of outputs regardless of both the value of λ and the nature of the discarded data (outliers or not). Furthermore, note that the uncertainty in p captures the discrepancy between the observations and the model prediction regardless of its origin. In the example above this discrepancy is caused by measurement noise (i.e., $g(t)$) and model-form uncertainty (i.e., describing (13) as a polynomial). The effects of model-form and parametric uncertainty, numerical and approximation error, measurement noise and biases are all lumped into P.

REFERENCES

[1] G. A. Seber and C. J. Wild, *Nonlinear Regression*. Hoboken, New Jersey, USA: JohnWiley & Sons, 2003.

[2] M. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society B*, vol. 63, no. 3, pp. 425–464, 2001.

[3] M. Campi, G. Calafiore, and S. Garatti, "Interval predictor models: Identification and reliability," *Automatica*, vol. 45, no. 2, pp. 382–392, 2009.

[4] A. Charnes, W. W. Cooper, and G. H. Symonds, "Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil," *A Journal of the Institute for Operations Research and the Management Sciences*, vol. 4, no. 3, 1958.

[5] G. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on automatic control*, vol. 51, no. 1, pp. 742–753, 2006.

[6] T. Alamo, A. Luque, D. Rodriguez, and R. Tempo, "Randomized control design through probabilistic validation," in *American Control Conference*, 2012.

[2]When $E_x[\delta_y]$ is evaluated by the sample mean, equivalence is attained by using w to evaluate the constraints, and z to evaluate the cost function.