Random Predictor Models for Rigorous Uncertainty Quantification: Part 1

Luis G. Crespo, Sean P. Kenny, and Daniel P. Giesy
Dynamic Systems and Control Branch
NASA Langley Research Center, MS 308, Hampton, VA, 23681, USA.

ABSTRACT: This and a companion paper propose techniques for constructing parametric mathematical models describing key features of the distribution of an output variable given input-output data. By contrast to standard models, which yield a single output value at each value of the input, Random Predictors Models (RPMs) yield a random variable at each value of the input. Optimization-based strategies for calculating RPMs having a polynomial dependency on the input and a linear dependency on the parameters are proposed. These formulations yield RPMs having various levels of fidelity in which the mean and the variance of the model's parameters, thus of the predicted output, are prescribed. As such they encompass all RPMs conforming to these prescriptions. The RPMs are optimal in the sense that they yield the tightest predictions for which all (or, depending on the formulation, most) of the observations are less than a fixed number of standard deviations from the mean prediction. When the data satisfies mild stochastic assumptions, and the optimization problem(s) used to calculate the RPM is convex (or, when its solution coincides with the solution to an auxiliary convex problem), the model's reliability, which is the probability that a future observation would be within the predicted ranges, can be bounded tightly and rigorously.

1 INTRODUCTION

Metamodeling refers to the process of creating a mathematical representation of a phenomenon based on input-output data. These models can have a parametric (e.g., polynomial response surfaces, smoothing spline models, polynomial chaos expansions) or non-parametric structure (e.g., Kriging/Gaussian process). In the former case the analyst first prescribes the model structure and then determines the value of the model parameters such that a measure of the discrepancy between observations and predictions is minimized. This step is commonly referred to as model calibration or regression. Model-form uncertainty, measurement noise, and numerical error often inhibit confidently prescribing a fixed constant value for such parameters. Consequently, a family of parameter values is prescribed such that the predictions resulting from evaluating the computational model at any family member accurately represents the observations.

Several model calibration techniques are available in the literature. Many of them assume the structure

$$y = M(x, p) + \eta, \tag{1}$$

where $y \in \mathbb{R}^{n_y}$ is the output, M is the computational model, $x \in \mathbb{R}^{n_x}$ is the input, $p \in \mathbb{R}^{n_p}$ is the parameter, and $\eta \in \mathbb{R}$ is a random variation caused by noise and measurement error. Many model calibration techniques are based on this structure, the assumption of p being a fixed but unknown constant (i.e., the uncertainty in p is epistemic), and the assumption of η being independent and identically distributed (IID) following a normal distribution with zero mean and a fixed variance. A typical regression problem consists of estimating the value of p in M given the set of observations (x_i, y_i), $i = 1, \ldots, N$, where $N > n_p$. This is often carried out by searching for the parameter realization that minimizes the sum of squared residuals, p_{LS}. The precision of this estimate, which prescribes how much it can deviate from its "true value" within an epistemic framework, is often evaluated using confidence intervals (Seber and Wild 2003). The calculation of confidence intervals, prediction intervals (i.e., intervals where future observations are expected to fall) and credible intervals (Kennedy and O'Hagan 2001) require having a probabilistic description of p. This description often requires (i) knowing/assuming a distribution for the prediction error, (ii) M and η taking particular forms (e.g., M depends linearly on p and the noise η is additive) and/or (iii) M being accurately represented by a linear approximation in p. As such, the suitability of the resulting intervals depends tightly on the validity of such assumptions.

A common approach to model calibration is

Bayesian inference. In Bayesian inference the objective is to describe the model's parameters as a vector of possibly dependent random variables using Bayes' rule (Kennedy and O'Hagan 2001). The resulting vector, called the posterior, depends on an assumed prior random vector, and the likelihood function; which in turn depends on the observations, and on the structure of M. Whereas this approach does not make any limiting assumptions on the manner in which M depends on p, nor on the structure of the resulting posterior; it requires that the calibrated variables in p be epistemic. This vector might be comprised of physical epistemic uncertainties and hyperparameters of aleatory variables[1]. Note that the consideration of aleatory uncertainties requires assuming a structure for them, so they can be parameterized in terms of non-physical epistemic variables. The presence of aleatory and model-form uncertainty yields uncertainty characterizations that fail to describe the prediction error. As more data is available, the calibrated p approaches a deterministic quantity, and the model prediction y converges to a single function of the input. The offset between this function and the data is not captured by the model. This deficiency can be mitigated by adding a fictitious discrepancy term to M (Kennedy and O'Hagan 2001). This term, which can have a fixed epistemic or a fixed aleatory structure, is calibrated as if it were part of M. In spite of its high computational demands, and of the potentially high sensitivity of the posterior to the assumed prior, this method is commonly regarded as a benchmark.

In this paper, we do not use an error term such as η, nor do we make prior assumptions about a distribution of p. What is here called a *Random Predictor Model* (RPM) has the general form $y = M(x, p)$, where p is a random vector, so the output, y, is a random process parameterized by x. We do not fully specify the distribution of p. Instead, we only seek to find a mean value, a variance, and, in some cases, a support set for p. These will be determined by solving optimization problems according to the input-output data and a scalar parameter chosen by the analyst. The role of this parameter is to limit the largest number of standard deviations that can separate the data points from the mean prediction. The resulting description of p is chosen to be as tight as possible while satisfying this restriction. We further provide means of identifying outliers in the data set so that eliminating them from the modeling process can result in predictions having a narrower range at the expense of a reduction in the model's reliability. As compared to previous work on interval predictor models (Crespo et al. 2014), the main contribution of this article is the consideration of random descriptions of p, thus of y, having an *arbitrary* structure, e.g., the results apply to p having a Normal, Beta or any other distribution.

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, the methods in this paper do not require any prior description of the uncertainty in p, and the resulting models yield analytical characterizations for both the predicted output and the model's reliability. This paper focuses on (i) computational models that depend linearly on the parameters and polynomially on the state, i.e., $y = p^\top \varphi(x)$ where $\varphi(x)$ is a vector of monomials in the components of x, and (ii) uncertainty sets for p that are hyper-rectangular. The advantage of these sets over all other sets (such as the ellipsoidal sets used in (Campi et al. 2009)), is that each component of p can be selected arbitrarily in its interval independently of the choices made for any of the other parameters. As such, parameter interdependencies are avoided. This independence enables the calculation of RPMs whose parameters are independent random variables. These properties enable an analytical description of the prediction and a formal quantification of its reliability. Extensions to models having other dependencies on x have been made, but this paper will only focus on polynomials.

This paper prescribes formulations for two types of RPMs. Whereas Type-1 RPMs use the entire data set, Type-2 RPMs neglect a fixed percentage of the observations. Such observations, which are identified while the RPM is calculated, are regarded as outliers. The companion paper (Crespo et al. 2015) focuses on two other types of RPMs.

2 PROBLEM STATEMENT

A system is postulated to act on a vector of *inputs* to produce an *output*. The output can depend on the state variables and on some other influences, causing, for instance, intrinsic variability. Let $X \subseteq \mathbb{R}^{n_x}$ be a set of input variables, and $Y \subseteq \mathbb{R}^{n_y}$ be a set of outputs which might result from the system acting on elements of X. In the following, the focus will be on the single-output ($n_y = 1$) multi-input ($n_x \geq 1$) case.

It is desired to build a model of the *Data Generating Mechanism* (DGM) which will predict the output corresponding to unobserved realizations of the input. The presence of intrinsic variability and uncertainty (e.g., the case in which some of the states that prescribe the measured output are unmeasurable or unknown to the analyst) makes it unreasonable to build a mathematical model that predicts a single output for a fixed input. Instead, an Interval Predictor Model (IPM) will predict an interval valued function into which the output from an unobserved input is expected to fall, while an RPM will predict a random process matching key features of the data. Engineering judgment is used to pick a collection of mono-

[1]For instance, if q contains the physical parameters of the model M, where q_1 is epistemic and q_2 is aleatory having a normal distribution with mean μ and standard deviation σ, the vector $p = [q_1, \mu, \sigma]^\top$ contains three epistemic variables, one physical and two non-physical.

mials in the state variables, $\varphi(x)$, to use as basis functions. 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 each data point (except, possibly, for a few outliers neglected purposely by the analyst) can be fit *exactly* by *at least one* element in such a set. One restriction to be considered is for p to belong to a set P. For a fixed value of the state x, the propagation of P through M yields an interval of output values. Thus, these models are called Interval Predictor Models. The objective here is to choose P to make the corresponding y intervals as small as possible and still allow each data point (x_i, y_i) to be modeled as $y_i = p^\top \varphi(x_i)$ for some $p \in P$. The other form of restriction considered is to describe p as a random vector. For a fixed value of the state x, the propagation of this vector through M yields a random variable $R_y(x)$ for the outcome y at x. Various properties of $R_y(x)$, such as mean, variance, and support set, are determined by those of p. The thrust here is to choose a random vector that leads a prediction matching key features of the data.

In this setting the two main problems of interest can be stated as follows. Let $\mathsf{z} = \{z_i\} = \{(x_i, y_i)\}$, for $i = 1, \ldots, N$, be a sequence of observations. First, we want to find an empirical model that, when evaluated at a new value x_{N+1} of the state, returns an informative prediction of the unobserved output y_{N+1}. An informative prediction can be interpreted as a prediction that is consistent with salient features of the data comprising z. These features, which are cast by the analyst as design requirements (for example, we might want all observed outcomes to be less than 2-standard deviations from the mean prediction), are cast as inequality constraints in the optimization problems used to calculate the model. Second, we want to quantify the probability that y_{N+1} be compliant with such requirements (in the previous example, we want to evaluate the probability that y_{N+1} be less than 2-standard deviations away from the mean prediction).

3 INTERVAL PREDICTOR MODELS

This section introduces basic concepts from IPMs that are essential for the construction of RPMs. Additional information on IPMs and examples are available in (Crespo, Kenny, & Giesy 2014). An IPM is simply a mapping that assigns an output interval for each value of the input. In the context of this paper, an IPM assigns to each instance vector $x \in X$ a corresponding outcome interval in Y. That is, an IPM is a set-valued map

$$I : x \to I_y(x) \subseteq Y, \tag{2}$$

where x is a state vector, and $I_y(x)$ 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 outputs y corresponding to all values of p in P:

$$I_y(x, P) = \{y = M(x, p),\ p \in P\}. \tag{3}$$

$I_y(x, P)$ 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 satisfy these restrictions. Attention will be limited to the IPM given by

$$I_y(x, P) = \{y = p^\top \varphi(x),\ p \in P\}. \tag{4}$$

where $\varphi(x)$ is a vector of monomials, and

$$P = \{p : \underline{p} \le p \le \overline{p}\}. \tag{5}$$

The analyst is free to choose which monomials are relevant to the particular application. A general representation of a multivariate polynomial basis is

$$\varphi(x) = \left[1, x^{i_2}, x^{i_3}, \ldots, x^{i_n}\right]^\top, \tag{6}$$

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[2].

The limits of the output of the IPM prescribed by (4-6) can be explicitly computed as

$$I_y(x, \overline{p}, \underline{p}) = \left[\, \underline{y}(x, \overline{p}, \underline{p}),\ \ \overline{y}(x, \overline{p}, \underline{p}) \,\right], \tag{7}$$

where

$$\underline{y}(x, \overline{p}, \underline{p}) = \varphi(x)^\top \left(\frac{\overline{p} + \underline{p}}{2}\right) - \varphi(|x|)^\top \left(\frac{\overline{p} - \underline{p}}{2}\right) \tag{8}$$

$$\overline{y}(x, \overline{p}, \underline{p}) = \varphi(x)^\top \left(\frac{\overline{p} + \underline{p}}{2}\right) + \varphi(|x|)^\top \left(\frac{\overline{p} - \underline{p}}{2}\right) \tag{9}$$

Therefore, the envelopes of the interval valued function I_y, are linear functions of $\underline{p}$ and $\overline{p}$, and piecewise polynomial functions of the input.The spread of I_y, which is the separation between its limits, is

$$\delta_y(x, \overline{p}, \underline{p}) = \varphi(|x|)^\top (\overline{p} - \underline{p}). \tag{10}$$

Note that the spread depends on the size of the uncertainty box P, but is independent of its geometric center.

Commonly, the *Least Squares* (LS) prediction, $y = p_{\text{LS}}^\top \varphi(x)$, where p_{LS} is given by

$$p_{\text{LS}} = (A^\top A)^{-1} A^\top [y_1, \ldots, y_N]^\top, \tag{11}$$

[2]The inclusion of 1 in $\varphi(x)$ guarantees that every (x, y) pair will be interpolated using some p even if $x = 0$.

for $A_{i,j} = \varphi_j(x_i)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_p$ is used to approximate the DGM. p_{LS} minimizes the sum of the squares of the predicted errors.

3.1 *Type-1 IPMs*

A Type-1 IPM is given by Equations (4-6) where P is the solution to the following Optimization Problem (OP).

Optimization Problem 1. *The limits of P are given by*

$$\left\langle \hat{\overline{p}}, \hat{\underline{p}} \right\rangle = \underset{p_{\text{b}},\, p_{\text{a}}}{\operatorname{argmin}} \left\{ E_x[\delta_y(x, p_{\text{b}}, p_{\text{a}})] : p_{\text{a}} \leq p_{\text{b}}, \right.$$

$$\left. \underline{y}(x_i, p_{\text{b}}, p_{\text{a}}) \leq y_i \leq \overline{y}(x_i, p_{\text{b}}, p_{\text{a}}) \right\}, \quad (12)$$

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

Therefore, a Type-1 IPM yields a P that minimizes the expected interval spread such that all the observed outputs are within $I_y(x)$. When x is a random vector of known distribution, the cost function in (12) can be calculated analytically. Otherwise, the sample mean based on the data can be used to approximate it. The resulting IPM, which is calculated by solving the convex optimization problem in (12), admits a rigorous reliability assessment (see Section 5). This assessment quantifies the probability that a future observation will fall within $I_y(x)$.

4 RANDOM PREDICTOR MODELS

A RPM is a mapping that assigns to each input vector $x \in X$ a corresponding random variable in the output space Y. That is, an RPM is a random variable-valued map

$$R : x \to R_y(x) \subseteq Y, \quad (13)$$

where x is the input, and $R_y(x)$ is a random process whose support lies in Y. A parametric RPM is obtained by associating to each $x \in X$ the set of outputs y corresponding to all values of p described by a random vector with joint Cumulative Distribution Function (CDF) $F_p(p)$ having the support set P. As before, attention will be limited to the case where the output is a linear function of the parameter p, and a polynomial function of x. This leads to

$$R_y(x) = \{ y = p^\top \varphi(x),\ p : F_p(p),\ p \in P \}. \quad (14)$$

Denote by $\mu \in \mathbb{R}^{n_p}$, $\nu \in \mathbb{R}^{n_p}$, and $c \in \mathbb{R}^{n_p(n_p-1)/2}$ the mean, variance and correlation of the random vector. The variance and correlation fully prescribe the covariance matrix $C(\nu, c) \in \mathbb{R}^{n_p \times n_p}$. It can be shown that any random vector with a support set P as in (5) must satisfy the consistency equations[3]

$$\underline{p} \leq \mu \leq \overline{p}, \quad (15)$$

$$0 \leq \nu \leq (\mu - \underline{p}) \odot (\overline{p} - \mu), \quad (16)$$

$$-1 \leq c \leq 1, \quad (17)$$

$$C(\nu, c) \succeq 0. \quad (18)$$

The operator symbol in (16) denotes the componentwise product of vectors, and the symbol in (18) denotes positive semidefiniteness.

The random process $R_y(x)$ is fully prescribed by the CDF of p. Naturally, statistics of the output y, such as the mean $\mu_y(x) = E_p[y(x,p)]$, the variance $\nu_y(x) = E_p[(y(x,p) - \mu_y(x))^2]$, and the range $I_y(x) = [\min_p y(x,p), \max_p y(x,p)]$, vary with x. In particular, the mean prediction is $\mu_y(x, \mu) = \mu^\top \varphi(x)$, the output's variance is

$$\nu_y(x, \nu, c) = \varphi(x)^\top C(\nu, c) \varphi(x), \quad (19)$$

and the output's range is the interval value function (7). When the components of p are uncorrelated, (19) reduces to [4]

$$\nu_y(x, \nu) = \nu^\top \varphi^2(x). \quad (20)$$

A few metrics for characterizing $R_y(x)$ are introduced next. The σ-surface, which connects all the outputs y that are τ standard deviations from the mean prediction, is defined by

$$l(x, \mu, \tau, \nu, c) = \mu^\top \varphi(x) + \tau \sqrt{\nu_y(x, \nu, c)}. \quad (21)$$

The σ-volume, defined as

$$I_\sigma(x, \mu, \tau, \nu) = \left[l(x, \mu, -\tau, \nu, c), l(x, \mu, \tau, \nu, c) \right], \quad (22)$$

contains all the outputs y that are no more than τ standard deviations away from the mean prediction $\mu_y(x)$. For the value of τ to be feasible (i.e., for the σ-surface to be within the support of R_y), it must satisfy

$$\underline{y}(x, \overline{p}, \underline{p}) \leq l(x, \mu, \tau, \nu, c) \leq \overline{y}(x, \overline{p}, \underline{p}). \quad (23)$$

Equation (23) ensures that the support of the process contains outcomes that are up to τ standard deviations from the mean prediction.

The formulations that follow prescribe key statis-

[3] The upper bound in (16) results from applying the expected value operator $E_{p_i}[\cdot]$ to both sides of $p_i^2 \leq (\underline{p}_i + \overline{p}_i) p_i - \underline{p}_i \overline{p}_i$, which holds for all $p_i \in [\underline{p}_i, \overline{p}_i]$, and using $\nu_i = E_{p_i}[p_i^2] - \mu_i^2$ for $i = 1, \ldots, n_p$.

[4] When the correlation c is zero, the corresponding argument of any function depending on it will be dropped from the notation.

tics of p, thus of the random output $y(x)$, based on input-output data. As such they encompass all RPMs that conform to these statistics. The first two of four types of RPMs are proposed here. Type-1 RPMs prescribe the mean and variance of $R_y(x)$ when the entire data set is used. Type-2 RPMs prescribe the same statistics after eliminating the effects of a fixed percentage of the observations. Such observations, which can be regarded as outliers, are worst-case in the sense that their removal tightens the predicted range of the σ-volume the most. The formulations below only consider the uncorrelated case $c = 0$. Extensions to the correlated case can easily be made. Furthermore, the selection of μ as p_{LS} is arbitrary, and any other value can be used. In the developments that follow, the *performance* of an RPM refers to the property evaluated by the cost function in the corresponding OP.

4.1 *Type-1 RPMs*

Type-1 RPMs prescribe the mean and variance of $R_y(x)$ when the entire data set in z is used. A Type-1 RPM is given by Equations (6, 14), where p is a vector of uncorrelated random variables with expected value $\mu = p_{LS}$, and a variance $\nu = \hat{\nu}$, given by the solution to the following program.

Optimization Problem 2. *The variance of p is equal to*

$$\hat{\nu} = \underset{\nu \geq 0}{\operatorname{argmin}} \{E_x[\nu_y(x,\nu)] : l(x_i,\mu,-\sigma_{max},\nu) \leq y_i \leq$$

$$l(x_i,\mu,\sigma_{max},\nu) \text{ for } i = 1,\ldots,N\}, \quad (24)$$

where $\sigma_{max} > 0$ is a parameter prescribed by the analyst, and (x_i, y_i) for $i = 1,\ldots,N$ are the observations in z.

Hence, a Type 1-RPM minimizes the expected variance of the random process $R_y(x)$ such that all observations are no more than σ_{max} standard deviations away from the mean prediction; i.e., all observations are within the σ-volume $I_\sigma(x,\mu,\sigma_{\text{max}},\hat{\nu})$.

The dependence of $\hat{\nu}$ on σ_{max} is studied next. Equation (24), which is subject to $2N + n_p$ inequality constraints, is equivalent to the linear program

$$\hat{\nu} = \underset{\nu}{\operatorname{argmin}} \left\{\nu^\top E_x\left[\varphi^2(x)\right] : \sigma^2_{\text{max}}\nu^\top\varphi^2(x_i) \geq \right.$$

$$\left.\left(y_i - \mu^\top\varphi(x_i)\right)^2 \text{ for } i = 1,\ldots,N,\ \nu \geq 0\right\}, \quad (25)$$

which is subject to $N + n_p$ constraints. The constraint set in (25) scales inversely with σ^2_{max}, so the scaled optimal objective value $\sigma^2_{\text{max}}\hat{\nu}^\top E_x[\varphi^2(x)]$, is constant as σ_{max} varies. It follows that that the larger σ_{max}, the smaller $\|\hat{\nu}\|$, and the larger the number of standard deviations separating any given point (x, y) from the corresponding mean prediction. This observation has consequences for the I_σ resulting from this formulation. If $\hat{\nu}_1$ is the solution to (25) corresponding to $\sigma_{\text{max},1}$, and $\hat{\nu}_2 = \alpha\hat{\nu}_1$ where $\alpha = (\sigma_{\text{max},1}/\sigma_{\text{max},2})^2$, then $\hat{\nu}_2$ is the solution to (25) corresponding to $\sigma_{\text{max},2}$. Consequently, the σ-volumes $I_\sigma(x,\mu,\sigma_{\text{max},1},\hat{\nu}_1)$, and $I_\sigma(x,\mu,\sigma_{\text{max},2},\hat{\nu}_2)$ are equal. Hence, the σ-volume, $I_\sigma(x,\mu,\sigma_{\text{max}},\hat{\nu})$ is independent of the choice of σ_{max}.

Note that both Type-1 IPMs and Type-1 RPMs require solving a convex OP. As such they can efficiently handle hundreds of thousands of data points, thus, many more input dimensions than alternative metamodels. This is in sharp contrast to Gaussian Processes which are limited to a few thousand data points before becoming numerically intractable.

A Type-1 RPM does not prescribe the support of p, thus, of $R_y(x)$. Any random vector satisfying the consistency Equations (15-18) for $\mu = p_{\text{LS}}$ and $\nu = \hat{\nu}$ is a valid characterization of $F_p(p)$. Since Type-1 RPMs are calculated by solving a convex optimization problem, they admit a rigorous reliability assessment. This assessment, presented in Section 5, quantifies the probability that a future observation will fall inside σ-volume $I_\sigma(x,p_{\text{LS}},\sigma_{\text{max}},\hat{\nu})$.

Example 1: Consider the DGM $y = x^2\cos(x) - \sin(3x)e^{-x^2} - \cos(x^2) + x(g-1)$, where $x \in \mathbb{R}$ is an IID sequence of random variables with uniform distribution over $X = [-5.5, 5.5]$, and g is IID with a standard normal distribution. Note that no knowledge on the structure of the DGM is required to calculate the RPMs. A data sequence z for $N = 150$ observations was generated. In the developments that follow we assume that n_p=7. In (Crespo et al. 2014) we calculate several IPMs based on the same data sequence. The LS solution is $p_{\text{LS}} = [-0.8734, -1.1059, -0.9926, 0.0026, -0.0228, -0.0004, 0.0028]^\top$.

A Type-1 RPM for $\sigma_{\text{max}} = 1$, to be referred to as RPM A, is shown in Figure 1. This figure shows the observations ($\times$'s), the mean prediction $\mu_y(x)$ (solid line), as well as σ-surfaces (green dashed-dotted lines) in increments of 0.5 standard deviations. Note that the observation near $(1, -15)$ limits the σ-volume from below. The only significant variance in $\hat{\nu}$ is $\hat{\nu}_1 = 180.3824$. The performance of RPM A, $E_x[\nu_y]$ is practically equal to $\hat{\nu}_1$. Note that 143 out of the 150 observations are within the $\sigma = 0.5$ volume. Further notice that the number of standard deviations between an arbitrary point (x, y) and the mean prediction for the same value of x, can be reduced by enlarging ν. This can be attained by reducing σ_{max}.

4.1.1 *Identification of Outliers*

The presence of outliers in the data yields undesirably large σ-volumes and output ranges, diminishing the RPMs performance. Whereas the limits of the optimal I_σ might be prescribed by a few observations, the majority of them might be much closer to the mean prediction, e.g., RPM A. The outliers, whose removal from the data set will lead to smaller predicted variances, can be identified using anyone of several fig-

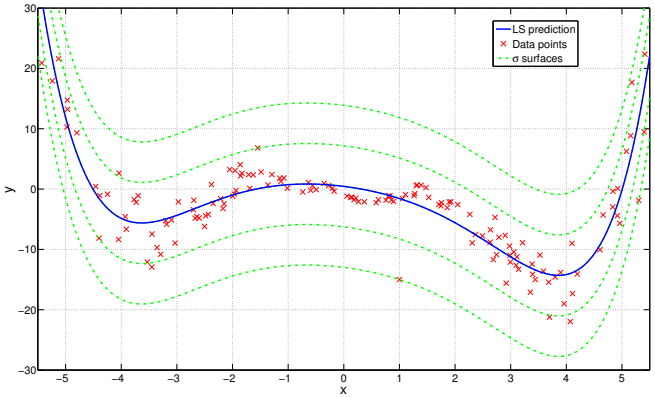

Figure 1: RPM A: Type-1 RPM for for $\sigma_{\max} = 1$.

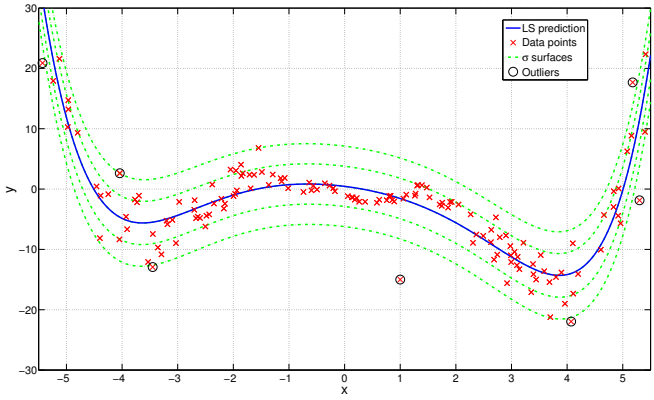

Figure 2: RPM B: Type-1 RPM after the removal of outliers.

ures of merit. In this paper we will use the figure of merit

$$\kappa_i(\mu, \nu, c) = \frac{\left(y_i - \mu^\top \varphi(x_i)\right)^2}{\nu_y(x_i, \nu, c)}, \tag{26}$$

where ν is the variance of p. κ_i is a variance-normalized distance squared between the ith observed output and the mean prediction at the corresponding input. Outliers will be identified by determining the data points corresponding to the largest percentiles of the empirical CDF of κ, $F_{\kappa(\hat{\nu})}(\kappa)$, for $i = 1, \ldots, N$, i.e., (x_i, y_i) is an outlier if $F_{\kappa(\hat{\nu})}(\kappa_i) > \lambda$ where $0 \ll \lambda < 1$. Once the outliers are identified, they can be removed from the data sequence and a new Type-1 RPM will be calculated. The resulting RPM will attain tighter predictions for a λ fraction of the observations in z, while the prediction for the remaining $1 - \lambda$ fraction might be considerably degraded. The outliers found by this procedure will be the same regardless of the value of $\sigma_{\max}$. This is a consequence of the following observation. If $(\kappa_i, F_\kappa(\kappa_i))$ are points on the optimal CDFs corresponding to $\sigma_{\max,1}$, the points on the optimal CDF corresponding to $\sigma_{\max,2}$ are $(\alpha\kappa_i, F_\kappa(\kappa_i))$, where α was defined earlier.

Example 2: We now derive a Type-1 RPM for $\sigma_{\max} = 1$ after removing seven outliers from the original data set. These outliers attain the largest values of κ_i. The resulting RPM, to be referred to as RPM B, is shown in Figure 2. In this case there are seven observations outside the $\sigma = 1$ volume by design (shown with circled cross symbols), 114 within the $\sigma = 0.5$ volume, and the remaining 29 are inside the $\sigma = 1$ volume and outside the $\sigma = 0.5$ volume. The only sizable variances for RPM B are $\hat{\nu}_1 = 44.5139$, and $\hat{\nu}_2 = 0.5194$. The performance of RPM B, $E_x[\nu_y] = 49.2469$, is 72.7% better than that of RPM A.

4.2 *Type-2 RPMs*

A formulation leading to an alternative RPM is presented next. In contrast to Type-I RPMs, this approach searches for ν by using only a fixed percentage of the N observations available. The observations comprising the removed set are worst-case in the sense that their removal tightens the optimal σ-volume the most. Whereas the outliers removed to construct RPM B are worst-case for the value of $\hat{\nu}$ corresponding to RPM A only, those neglected in a Type-2 RPMs are worst-case for the varying value of ν being considered during the optimization. This will be carried out without removing any point in the data sequence.

In particular, a Type-2 RPM is given by Equations (6, 14), where p is a vector of uncorrelated variables with expected value $\mu = p_{\text{LS}}$, and a variance $\nu = \hat{\nu}$ given by the following OP.

Optimization Problem 3. *The variance of p is equal to*

$$\hat{\nu} = \underset{\nu \geq 0}{\operatorname{argmin}} \left\{ E_x[\nu_y(x, \nu)] : F_{\kappa(\nu)}\left(\sigma^2_{max}\right) \geq \lambda \right\}, \tag{27}$$

where $\sigma_{max} > 0$ is a parameter prescribed by the analyst, $F_{\kappa(\nu)}$ is the empirical CDF of $\kappa(\nu)$ in (26) based on the N observations in z, *and $0 < \lambda \leq 1$, another parameter to be chosen by the analyst, is the proportion of observations to be contained by $I_\sigma(x, \mu, \sigma_{max}, \hat{\nu}(\lambda))$.*

Hence, a Type-2 RPM minimizes the expected variance of the random process $R_y(x)$ such that $100\lambda\%$ of the observations are no more than $\sigma_{\max}$ standard deviations apart from the mean prediction. The tightening of the prediction for $100\lambda\%$ of the observations caused by (27) yields a σ-volume $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$ that does not enclose the remaining $100(1 - \lambda)\%$. This shows that (27) is a chance-constraint formulation (Charnes et al. 1958), in which one is willing to accept the occurrence of unfavorable low-probability events (probability $1 - \lambda$) for the sake of an improved performance for high-probability events (probability λ). As with Type-1 RPMs, $\sigma_{\max}$ is essentially a scaling factor.

OP3 is a non-convex formulation, which for $\lambda = 1$ yields the same RPM as OP2 [5]. When $\lambda < 1$,

[5] Note that, if $E_x[\cdot]$ is calculated based on a sample mean, the entire z sample must be used to make the convex formulation equivalent to (27).

a fixed number of observations (outliers) are neglected as the RPM is being calculated. Outliers can be easily identified by finding the data points for which $F_{\kappa(\hat{\nu})}(\kappa_i(\hat{\nu})) > \lambda$. The points not satisfying this condition, which are the elements of z within $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$, constitute the sequence w. A Type-1 RPM based on the data sequence w is equivalent to the Type-2 RPM in (27) based on the data sequence z. This relationship enables performing a reliability assessment of Type-2 RPMs. This assessment, presented in Section 5, quantifies the probability that a future observation will be within $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$.

Example 3: We now derive a Type-2 RPM for the same observations used earlier having $\lambda = 143/150$ and $\sigma_{\max} = 1$. As with RPM B, we search for a σ-volume for which 143 observations are less than one standard deviation from the mean prediction. The resulting RPM, shown in Figure 3, will be referred to as RPM C. Note that the process is more concentrated about the LS prediction than either RPM A or RPM B. The only sizable components of $\hat{\nu}$ are $\hat{\nu}_1 = 6.0124$, and $\hat{\nu}_2 = 3.2985$. Note that the outliers, which are the observations outside $I_\sigma(x, \mu, 1, \hat{\nu}(\lambda))$, differ from those corresponding to RPM B. The performance of RPM C, $E_x[\nu_y] = 37.7676$, is 23% better than that of RPM B.

Figure 4 shows the empirical CDFs of $\kappa(\hat{\nu})$ for RPM A, B and C. The horizontal line $\lambda = 143/150$ is shown in green. Recall that the larger the expected value of κ, the more concentrated $F_p(p)$, and the better the RPM. Using the figure of merit $E[\kappa|\kappa < \sigma_{\max}]$, which is the area between the constant function λ and the CDF in the domain $\kappa \in [0, \sigma_{\max}]$, the ranking from best to worst is RPM C, RPM B, and RPM A. The advantage of RPM C is also reflected in the values of $E_x[\nu_y]$ listed above. The vertical jumps in the CDFs at $\kappa = \sigma_{\max}$ are the result of obtaining an optimum for which several observations are on the boundary of the $\sigma_{\max}$-volume.

The evaluation of the same figure of merit above over the domain $\kappa \in [\sigma_{\max}, \infty]$ indicates that RPM B is the best model of the three. This can be inferred from Figure 4 by noting that the CDF of RPM B assumes the largest values of κ for most of the probability values exceeding λ. This illustrates that (27) is a chance-constraint formulation, in which one is willing to accept the occurrence of unfavorable low-probability events (i.e., those in the $\kappa \in [\sigma_{\max}, \infty]$ range) for the sake of an improved performance for high-probability events (i.e., those in the $\kappa \in [0, \sigma_{\max}]$ range).

5 MODEL'S RELIABILITY

This section presents a framework for rigorously evaluating the reliability of the predictor models proposed above. The reliability of model $\mathcal{E}$, $r(\mathcal{E})$, is the probability that a future observation will be compliant with the requirements imposed upon the calculation of the

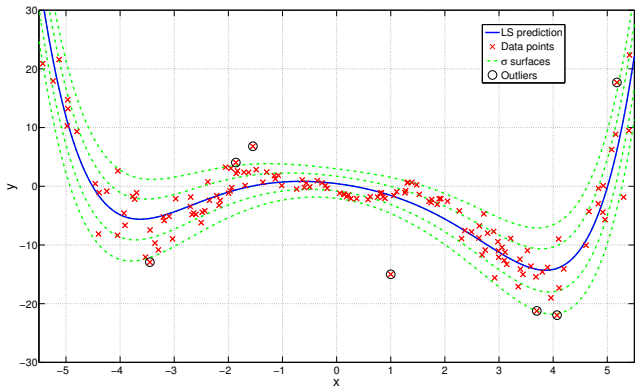

Figure 3: RPM C: Type-2 RPM for $\lambda = 143/150$ and $\sigma_{\max} = 1$.

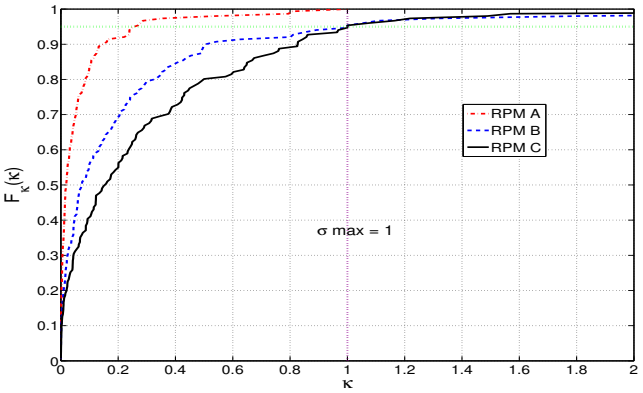

Figure 4: CDFs of $\kappa(\hat{\nu})$ for RPM A (red, dashed-dotted), B (blue, dashed) and C (black, solid).

model. These requirements are cast in terms of an output y belonging to a σ-volume $I_\sigma(x)$ for both Type-1 and Type-2 RPMs. The developments that follow are based on the *Scenario Approach* (Calafiore & Campi 2006).

Denote by $\mathbb{P}$ the *unknown* distribution of the DGM from which the points of the data sequence z are obtained. $\mathbb{P}$ can be interpreted as a probabilistic cloud in the $X \times Y$-space. The case in which y is a deterministic 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 y 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 following theorem, taken from (Campi, Calafiore, & Garatti 2009), permits quantifying the reliability of an empirical predictor model whenever the OP used for its calculation is convex.

Theorem 1. *Let* $\mathsf{z} = \{z_i\} = \{(x_i, y_i)\}$, *for* $i = 1, \ldots, N$, *be an independent data sequence resulting from a stationary discrete-time data generating process. Suppose the model* $\mathcal{E}$ *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 assuming* $k < N - d$, *where* d *is the*

number of optimization variables used to calculate $\mathcal{E}$, it holds that

$$\text{Prob}_{\mathbb{P}^N}\left[r(\mathcal{E}) \geq 1-\epsilon\right] > 1-\beta, \text{ where} \tag{28}$$

$$\beta = \frac{N!(1-\epsilon)^{N-d}}{(N-d)!d!}\sum_{i=0}^{k}\frac{(N-d)!}{(N-d-i)!i!}\frac{\epsilon^i}{(1-\epsilon)^i}. \tag{29}$$

This theorem provides an assessment of 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 model by virtue of the randomness in $\mathbb{P}$ prescribing z. 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. It is worth noting that the confidence parameter can be made small enough 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 with N.

The reliability of a Type-1 IPM, to be denoted as $\mathcal{I}$, is defined as $r(\mathcal{I}) = \text{Prob}_{\mathbb{P}}\left[(x,y)\in I_y\left(x,\hat{\underline{p}},\hat{\underline{p}}\right)\right]$. Hence, $r(\mathcal{I})$ is the probability that an unobserved input-output pair (x,y) will fall within the range $I_y(x)$. The convexity of the OP1 enables the direct application of Theorem 1. The reliability of any Type-1 or Type-2 RPM, to be denoted as $\mathcal{R}$, is defined as $r\left(\mathcal{R}\right) = \text{Prob}_{\mathbb{P}}\left[(x,y)\in I_\sigma(x,\mu,\sigma_{\max},\hat{\nu}(\lambda))\right]$. Hence, $r(\mathcal{R})$ is the probability that an unobserved input-output pair (x,y) will fall in the optimal σ-volume corresponding to $\sigma_{\max}$.

The convexity of OP2 enables the direct application of Theorem 1 to Type-1 RPMs. This includes the cases in which none ($k=0$) and some ($k>0$) of the observations are removed from the data set in advance. In contrast to OP2, OP3 is non-convex. This opens the possibility of (27) having multiple optima. Multiple optima may result from the possibility of obtaining the same RPM for different sets of outliers. Because Type-2 RPMs are calculated by solving a non-convex program, Theorem 1 cannot be applied directly. However, the reliability of such models can be establish by using the *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 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 RPMs. In particular, the solution to OP3 using the original the data sequence z for a given value of λ is equivalent to the solution of OP2, which is a convex program, with the data sequence w. Because only the $N-k^*$ elements in w, where

$$k^* = \text{floor}[N(1-\lambda)], \tag{30}$$

are required by the auxiliary program, the reliability of Type-2 RPMs is given by (28) with $k=k^*$ in (29). These k^* observations fall outside the optimal σ-volume and satisfy $F_{\kappa(\hat{\nu})}(\kappa) > \lambda$.

Example 6: The reliability of RPM A, B and C, for which (28) is directly applicable, is considered first. The reliability of RPM A, which is a Type-1 RPMs calculated using $N=150$, $k=0$ and $d=7$, is no less than $1-\epsilon = 0.8050$ with confidence $1-\beta = 0.99$; while the reliability of RPM B, which is also a Type-1 RPM for which $N=150$, $k=7$ and $d=7$, is no less than $1-\epsilon = 0.6984$ with confidence $1-\beta = 0.99$. Hence, the exclusion of seven outliers rendered an improvement in the system performance of 72.7% at the expense of a reduction in the model's reliability of 10.66%. The reliability of RPM C, which is a Type-2 RPM for which $k=k^*=7$, and the reliability of RPM B are the same, even though the performance of RPM C is 23% better.

6 CONCLUSIONS

This and the companion paper (Crespo et al. 2015) develop techniques for constructing random predictor models based on data. The formulations proposed enable a rigorous characterization of key features of the predicted output, and of the reliability of the prediction. These articles set forth a new paradigm for the construction of empirical models in which the models performance and reliability can be evaluated and traded off using rigorous means.

REFERENCES

Calafiore, G. & M. C. Campi (2006). The scenario approach to robust control design. *IEEE Transactions on automatic control 51*(1), 742–753.

Campi, M., G. Calafiore, & S. Garatti (2009). Interval predictor models: Identification and reliability. *Automatica 45*(2), 382–392.

Charnes, A., W. W. Cooper, & G. H. Symonds (1958). 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 4*(3).

Crespo, L. G., S. P. Kenny, & D. P. Giesy (2014, December). Interval predictor models with a formal characterization of uncertainty and reliability. In *53 IEEE Conference on Decision and Control*, Los Angeles, CA, USA, pp. 1–26.

Crespo, L. G., S. P. Kenny, & D. P. Giesy (2015, September, 7-10). Random predictor models for rigorous uncertainty quantification: Part 2. In *ESREL 2015*, Zurich, Switzerland.

Kennedy, M. & A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society B 63*(3), 425–464.

Seber, G. A. & C. J. Wild (2003). *Nonlinear Regression*. Hoboken, New Jersey, USA: JohnWiley & Sons.