Random Predictor Models for Rigorous Uncertainty Quantification: Part 2

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, the variance, and the range of the model's parameter, thus of the 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, is bounded rigorously.

1 INTRODUCTION

It is assumed the reader is familiar with the content of the companion paper (Crespo et al. 2015). The introduction, and literature review therein apply to this paper as well but have been omitted here due space limitations. Whereas (Crespo et al. 2015) focuses on Type-1 and Type-2 RPMs, this paper focuses on Type-3 and Type-4 RPMs. The overlap between the papers has been kept to a minimum.

2 PROBLEM STATEMENT

A system is postulated to act on a vector of *input variables* x to produce an *output* y. 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. 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 generated by a *Data Generating Mechanism* (DGM). 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 on the RPM (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} is compliant with such requirements. To continue the example, we want to evaluate the probability that y_{N+1} is less than 2-standard deviations away from the mean prediction.

3 INTERVAL PREDICTOR MODELS

This section introduces concepts from Interval Predictor Models (IPM) that are essential for RPMs. Additional information on IPMs and examples are available in (Crespo et al. 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 input vector $x \in X$ a corresponding outcome interval in Y. 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:

$$I_y(x, P) = \{y = p^\top \varphi(x),\ p \in P\}, \tag{1}$$

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

$$P = \{p : \underline{p} \leq p \leq \overline{p}\}. \quad (2)$$

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

$$\varphi(x) = \left[1, x^{i_2}, x^{i_3}, \ldots, x^{i_n}\right]^\top, \quad (3)$$

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 limits of the IPM prescribed by (1-3) can be explicitly computed as

$$I_y(x, \overline{p}, \underline{p}) = \left[\underline{y}(x, \overline{p}, \underline{p}),\ \ \overline{y}(x, \overline{p}, \underline{p})\right], \quad (4)$$

where

$$\underline{y}(x, \overline{p}, \underline{p}) = \varphi(x)^\top \left(\frac{\overline{p} + \underline{p}}{2}\right) - \varphi(|x|)^\top \left(\frac{\overline{p} - \underline{p}}{2}\right) \quad (5)$$

$$\overline{y}(x, \overline{p}, \underline{p}) = \varphi(x)^\top \left(\frac{\overline{p} + \underline{p}}{2}\right) + \varphi(|x|)^\top \left(\frac{\overline{p} - \underline{p}}{2}\right) \quad (6)$$

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

$$\delta_y(x, \overline{p}, \underline{p}) = \varphi(|x|)^\top (\overline{p} - \underline{p}). \quad (7)$$

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

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

$$p_{\text{LS}} = (A^\top A)^{-1} A^\top [y_1, \ldots, y_N]^\top, \quad (8)$$

$$A_{i,j} = \varphi_j(x_i), \text{ for } i = 1, \ldots, N \text{ and } j = 1, \ldots, n_p.$$

3.1 *Type-1 IPMs*

A Type-1 IPM is given by Equations (1-3) where P is the solution to the following Optimization Problem (OP).

Optimization Problem 1 (OP1): *The limits of P are given by*

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

$$\left. \underline{y}(x_i, p_{\text{b}}, p_{\text{a}}) \leq y_i \leq \overline{y}(x_i, p_{\text{b}}, p_{\text{a}}) \right\}, \quad (9)$$

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 (9) 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 (9), admits a rigorous reliability assessment (see Section 5). This assessment quantifies the probability that a future observation will fall within $I_y(x)$.

The membership of p_{LS} in P can be ensured by replacing the first constraint with $p_{\text{a}} \leq p_{\text{LS}} \leq p_{\text{b}}$, or adding the constraint $p_{\text{a}} + p_{\text{b}} = 2p_{\text{LS}}$. In general, the inclusion of these constraints leads to IPMs with larger expected spreads, with the equality constraint leading to the larger of the two. A formulation resulting from adding either of these two sets of constraints will be called *Augmented*.

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

$$R : x \to R_y(x) \subseteq Y, \quad (10)$$

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

$$R_y(x) = \{y = p^\top \varphi(x),\ p : F_p(p),\ p \in P\}. \quad (11)$$

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 p respectively. 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 (2) must satisfy the consistency equations:

$$\underline{p} \ \leq \ \mu \ \leq \ \overline{p}, \quad (12)$$

$$0 \ \leq \nu \ \leq \ (\mu - \underline{p}) \odot (\overline{p} - \mu), \quad (13)$$

$$-1 \ \leq c \ \leq \ 1, \quad (14)$$

$$C(\nu, c) \succeq 0. \quad (15)$$

The symbols $\odot$ and $\succeq$ denote the component-wise product of vectors, and positive semidefiniteness respectively.

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

$$\nu_y(x,\nu,c) = \varphi(x)^\top C(\nu,c)\varphi(x), \tag{16}$$

and the output's range is the interval value function (4). When the components of p are uncorrelated, (16) reduces to[1]

$$\nu_y(x,\nu) = \nu^\top \varphi^2(x). \tag{17}$$

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

$$l(x,\mu,\tau,\nu,c) = \mu^\top \varphi(x) + \tau\sqrt{\nu_y(x,\nu,c)}. \tag{18}$$

The σ-volume, defined as

$$I_\sigma(x,\mu,\tau,\nu) = [l(x,\mu,-\tau,\nu,c), l(x,\mu,\tau,\nu,c)], \tag{19}$$

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

$$\underline{y}(x,\overline{p},\underline{p}) \le l(x,\mu,\tau,\nu,c) \le \overline{y}(x,\overline{p},\underline{p}). \tag{20}$$

Equation (20) ensures that the support of the process contains outcomes that are up to τ standard deviations from the mean prediction. Note that the range of standard deviation values satisfying these inequalities is a function of x.

The formulations that follow prescribe key statistics 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. Four types of RPMs are proposed. Type-1 RPMs and Type-2 RPMs, covered in detail and exemplified in (Crespo et al. 2015), prescribe the mean and variance of p. Conversely, Type-3 and Type-4 RPMs also prescribe the range of the output. Whereas Type-3 RPMs emphasize the tightness of the outputs' range, Type-4 RPMs emphasize the tightness of the σ-volume. In contrast to Type-1 and Type-2 RPMs, which only require solving a single OP, Type-3 and Type-4 RPMs require solving a pair of interdependent OPs. The formulations below only consider $c = 0$. 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. The section that follows covers the essentials of Type-1 RPMs and Type-2 RPMs needed to calculate Type-3 and Type-4 RPMs.

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

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 (3, 11), 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 (OP2): *The variance of p is equal to*

$$\hat{\nu} = \underset{\nu \ge 0}{\operatorname{argmin}} \{E_x[\nu_y(x,\nu)] : l(x_i,\mu,-\sigma_{max},\nu) \le y_i \le l(x_i,\mu,\sigma_{max},\nu) \text{ for } i = 1,\ldots,N\}, \tag{21}$$

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 $\sigma_{\max}$ standard deviations away from the mean prediction, i.e., all observations are within the σ-volume $I_\sigma(x,\mu,\sigma_{\max},\hat{\nu})$.

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 of input dimensions. 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 (12-15) 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 OP, they admit a rigorous reliability assessment. This assessment, presented in Section 5, quantifies the probability that a future observation will fall within the σ-volume $I_\sigma(x,\mu,\sigma_{\max},\hat{\nu})$.

4.1.1 *Outliers in the Data Set*

The presence of outliers in the data yields undesirably large σ-volumes and uncertainty sets, 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. The outliers, whose removal from the data set

will lead to smaller predicted variances, can be identified using anyone of several figures of merit. This paper will use the figure of merit

$$\kappa_i(\mu, \nu, c) = \frac{\left(y_i - \mu^\top \varphi(x_i)\right)^2}{\nu_y(x_i, \nu, c)}, \tag{22}$$

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 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 $\sigma_{\max}$.

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.

In particular, a Type-2 RPM is given by Equations (3, 11), 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 (OP3): *The variance of p is*

$$\hat{\nu} = \underset{\nu \geq 0}{\operatorname{argmin}} \left\{ E_x[\nu_y(x, \nu)] : F_{\kappa(\nu)}\left(\sigma^2_{max}\right) \geq \lambda \right\}, \tag{23}$$

where $\sigma_{max} > 0$ is a parameter prescribed by the analyst, $F_{\kappa(\nu)}$ is the empirical CDF of $\kappa(\nu)$ in (22) 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 a λ fraction of the observations are no more than $\sigma_{\max}$ standard deviations apart from the mean prediction. The tightening of the prediction for such a fraction yields a σ-volume $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$ that does not enclose the remaining $1 - \lambda$ fraction. This shows that (23) 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. When $\lambda < 1$, 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 (23) based on the data sequence z. This relationship enables performing a reliability assessment of Type-2 RPMs. This assessment, presented in Section 5, formally quantifies the probability that a future observation will be within $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$.

4.3 *Type-3 RPMs*

Type-3 RPMs not only prescribe the mean and variance of p, thus of $R_y(x)$, but also their ranges. Type-3 RPMs optimize the tightness of both the range of the output, and of the σ-volume prioritizing the former. In contrast to Type-1 and Type-2 RPMs, the calculation of a Type-3-RPM entails solving a sequence of two OPs.

A Type-3 RPM is defined by Equations (3, 11), where μ is given by the LS parameter estimate in (8). The support set P is prescribed by an Augmented version of (9), and the variance ν is the solution to the following OP:

Optimization Problem 4 (OP4): *The variance of p is*

$$\hat{\nu} = \underset{0 \leq \nu \leq \nu_{max}}{\operatorname{argmin}} \left\{ E_x[\nu_y(x, \nu)] : F_{\kappa(\nu)}\left(\sigma^2_{max}\right) \geq \lambda \right\}, \tag{24}$$

where $\nu_{max} = (\mu - \underline{\hat{p}}) \odot (\hat{\overline{p}} - \mu)$, $\hat{\overline{p}}$ and $\underline{\hat{p}}$ are given by (9), and $F_{\kappa(\nu)}$ is the empirical CDF of $\kappa(\nu)$ in (22) based on the N observations in z*. The parameters σ_{max} and λ, to be set by the analyst and defined earlier, must satisfy*

$$\sigma_{max} > \sigma^*_{max} = \max_{1 \leq i \leq N} \left\{ \frac{|y_i - \mu^\top \varphi(x_i)|}{\sqrt{\nu_{max}^\top \varphi^2(x_i)}} \right\}, \tag{25}$$

and $0 < \lambda \leq 1$.

Hence, a Type-3 RPM minimizes the expected variance of the random process $R_y(x)$ given that (i) the σ-volume associated with $\sigma_{\max}$ contains a λ fraction of the observations, (ii) the relationships among the mean, the variance and the support set satisfy the consistency Equations (12-15), and (iii) the range of outputs $I_y(x)$ has minimal expected spread while containing all N observations. While Augmented OP1 is convex, the first inequality constraint in (24) makes OP4 non-convex. Notice that extreme observations prescribe the support set P in OP1, whereas the floor$(N\lambda)$ observations attaining the smallest κ_i values prescribe $\hat{\nu}$ in OP4.

The solution to (9) enters (24) via the upper bound on ν, $\nu_{\max}$. The constraint (25) ensures the feasible design space is non-empty. The ith component of the vector at the right hand side of (25) is the value of τ_i for which $y_i = l(x_i, \mu, \tau_i, \nu_{\max})$. Hence, τ_i is the smallest number of standard deviations that can separate (x_i, y_i) from the mean prediction without letting ν exceed $\nu_{\max}$.

When $\lambda = 1$, the constraints in (24) can be written as a set of convex constraints. When $\lambda < 1$, the constraints in (24) are equivalent to a subset of the convex constraints. This subset is given by all the elements in z satisfying $F_{\kappa(\hat{\nu})}(\kappa_i) \leq \lambda$. The floor$(N\lambda)$ observations satisfying this condition constitute the data sequence w. Therefore, OP4 based on the data sequence z renders the same empirical model as a convex-program based on the data sequence w. This is the basis used for evaluating the reliability of Type-3 RPMs. To this end (See Theorem 2), it is useful to determine if $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda)) \subseteq I_y(x, \hat{\overline{p}}, \hat{\underline{p}})$ holds, i.e., the σ-volume associated with $\sigma_{\max}$ is within the range of $R_y(x)$. This is the case if and only if the *Containment Condition*

$$\varphi(|x|)^\top(\hat{\overline{p}} - \hat{\underline{p}}) - |\varphi(x)^\top(\hat{\overline{p}} + \hat{\underline{p}} - 2\mu)| - 2\sigma_{\max}\sqrt{\hat{\nu}^\top\varphi^2(x)} \geq 0, \quad (26)$$

holds for all $x \in X$. This semi-infinite constraint can be evaluated rigorously using Bernstein polynomials and interval analysis. Type-3 RPMs satisfying (26) allow for a tighter reliability assessment. Enforcing this condition by design requires incorporating (26) into (24). This practice, however, will not be considered in this paper.

Note that Type-3 RPMs for the case in which $\lambda = 1$ can be found by solving a sequence of two convex OPs. This structure allows considering problems with hundreds of thousands of observations. In such a case outliers can be dealt with by identifying them and removing them from the data sequence in advance as explained in Section 4.1.1.

Example 1: Two Type-3 RPMs based on the same data sequence used in (Crespo et al. 2015) are derived next. Whereas the two RPMs differ in the value of λ used to calculate $\hat{\nu}$, both use the same set P. This set is calculated via an Augmented OP1 with $\underline{p} \leq p_{\text{LS}} \leq \overline{p}$. This leads to $\hat{\underline{p}} = [-12.9837, -1.1488, -0.8339, 0.0013, -0.0379, -0.0001, 0.0032]^\top$, and $\hat{\overline{p}} = [7.2080, -1.1488, -0.8339, 0.0013, -0.0379, -0.0001, 0.0034]^\top$. These values, in turn, yield an upper bound for ν where the only significant component is $\nu_{\max,1} = 90.8037$. This IPM's performance is $E_x[\delta_y] = 10.4942$. The bound on $\sigma_{\max}$ resulting from (25) yields $\sigma^*_{\max} = 1.4094$. Thus, we selected $\sigma_{\max} = 1.5$.

A Type-3 RPM for $\lambda = 1$ is calculated first. Therefore, we require that all 150 observations be no more than 1.5 standard deviations from the mean prediction. The resulting RPM, to be referred to as RPM D, leads to a variance $\hat{\nu}$ for which the only significant term is $\hat{\nu}_1 = 80.1699$. The performance of RPM D is given by both $E_x[\delta_y] = 10.4942$ and $E_x[\nu_y] \approx \hat{\nu}_1$. Figure 1 shows RPM D. Whereas the limits of the range, $I_y(x)$, are shown as dashed lines, σ-surfaces separated by 0.5 units are shown as dashed-dotted lines. Note that the augmented constraint yielded a skewed random process with respect to its mean prediction. Further notice that the lower limit of the support coincides with the σ-surface associated with $\sigma = -1.5$, whereas the values of σ reaching the upper limit of the range vary. Even though the portions of the σ-surfaces spreading outside $I_y(x)$ are infeasible (e.g., almost the entire $\sigma = 1.5$ surface), they have been plotted for clarity. The feasible range of σ values at each value of x is given by (20). Because the majority of the observations are close to the mean prediction, we can expect that neglecting a few extreme observations will considerably improve the model's performance.

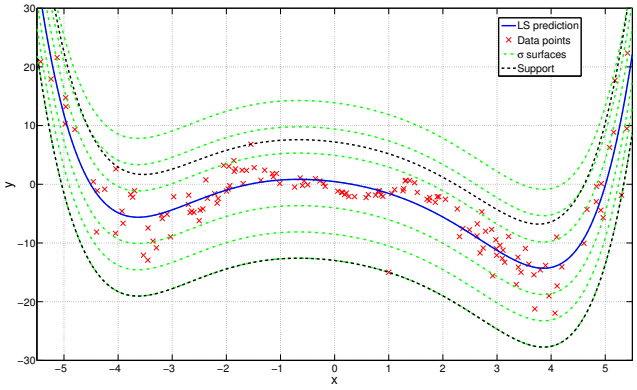

Figure 1: RPM D: Type-3 RPM for $\sigma_{\max} = 1.5$ and $\lambda = 1$.

A Type-3 RPM for $\lambda = 143/150$ is derived next. Therefore, we require that 143 observations be no more than 1.5 standard deviations from the mean prediction. This model, to be referred to as RPM E, leads to a variance $\hat{\nu}$ for which $\hat{\nu}_1 = 22.2497$ is the only significant term. This indicates that the CDF of p corresponding to RPM E is about four times more concentrated about the mean than that of RPM D. The performance of RPM E is given by $E_x[\delta_y] = 10.4942$ as before, and by $E_x[\nu_y] \approx \hat{\nu}_1$. In terms of the latter metric, RPM E is 72% better than RPM D. Figure 2 shows σ-surfaces corresponding to RPM E. The same line conventions used before apply. A comparison between Figures 1 and 2 indicates that RPM E yields a tighter probabilistic description for $100\lambda\%$ of the observations than RPM D. The containment condition in (26), which will be required to calculate the models' reliability, is s not satisfied for either RPM D or RPME. This is reflected in Figures 1 and 2, where the σ-surface corresponding to $\sigma_{\max}$ is above $\overline{y}(x, \hat{\overline{p}}, \hat{\underline{p}})$ for some x in X.

The sequential construction of a Type-3 RPM,

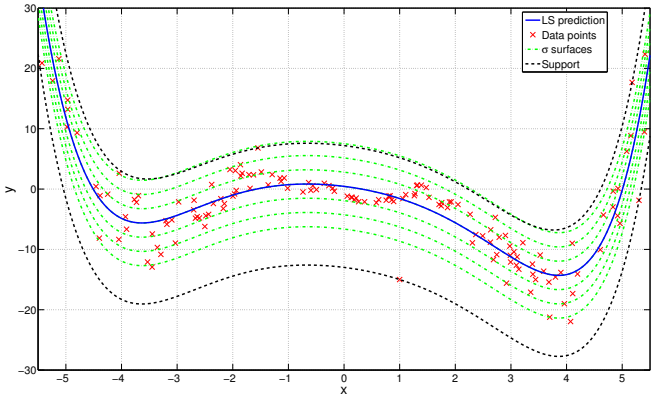

Figure 2: RPM E: Type-3 RPM for $\sigma_{\max} = 1.5$ and $\lambda = 143/150$.

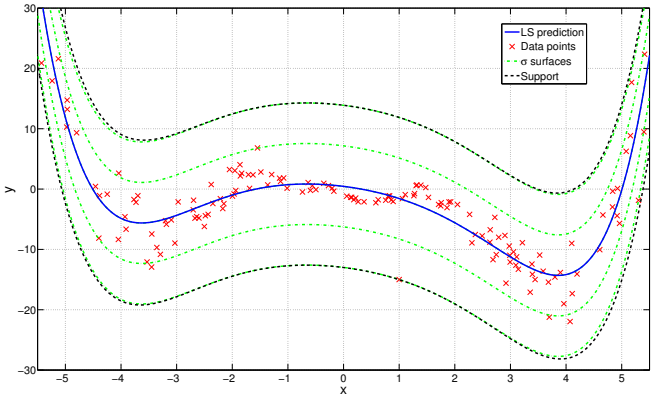

Figure 3: RPM F: Type-4 RPM for $\lambda = 1$ and $\sigma_{\max} = 1$.

where the variance ν is solved for after solving for the support set P, restricts its probabilistic performance (i.e., the variance is calculated given an optimal support set). This restriction manifests in the lower bound (25) to admissible values of $\sigma_{\max}$. A sequential approach reversing the priority order is presented next (i.e., P is calculated given an optimal variance).

4.4 *Type-4 RPMs*

A Type-4 RPM is given by by Equations (3, 11), where the expected value μ is given by the LS solution in (8), the variance ν is given by (23), and the support P is given by the following OP.

Optimization Problem 5 (OP5): *The support set P of the random vector p, having expected value μ and variance $\hat{\nu}(\sigma_{max}, \lambda)$, is given by*

$$\langle \hat{\overline{p}}, \hat{\underline{p}} \rangle = \underset{p_b,\, p_a}{\operatorname{argmin}} \{ E_x[\delta_y(x, p_b, p_a)] : p_a \leq \mu \leq p_b,$$

$$\underline{y}(x_i) \leq y_i \leq \overline{y}(x_i) \ \textit{for} \ i = 1, \ldots, N,$$

$$\hat{\nu} \leq (\mu - p_a) \odot (p_b - \mu) \} . \qquad (27)$$

Hence, a Type-4 RPM minimizes the expected spread of the random process given that (i) P contains μ, (ii) the range contains all the observations, (iii) the relationship between $\hat{\nu}$ and P satisfies consistency condition (13), and (iv) the σ-volume $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$ contains $100\lambda\%$ of the observations. Note that the solution to OP3 enters into OP5 via the lower bound of the last constraint. Further notice that OP3, used to calculate $\hat{\nu}$, is non-convex, whereas OP5, used to calculate P, is convex. This is the case even though the feasible design space associated with the bilinear constraints in (27) is non-convex. The equivalence between OP3 and OP2 covered in Section 4.2, allows performing a rigorous reliability analysis of Type-4 RPMs. This analysis quantifies the probability that a future observation will be inside both the σ-volume $I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$ and the range $I_y(x, \hat{\overline{p}}, \hat{\underline{p}})$. As before, the containment condition in (26) plays a key role in the evaluation of the model's reliability.

Example 2: Next we derive two Type-4 RPMs for $\sigma_{\max} = 1$ and the same data used earlier. The two RPMs differ in the value of λ used to calculate $\hat{\nu}$. Because $\sigma_{\max} < \sigma^*_{\max} = 1.4094$, there is no Type-3 RPM able to satisfy this requirement. This illustrates the limitations on the probabilistic performance resulting from Type-3 RPMs.

The first RPM, to be referred to as RPM F, uses $\lambda = 1$. Hence, we will require that all 150 observations will be less than one standard deviation from the LS prediction. The only significant term in the solution is $\hat{\nu}_1 = 180.3824$. With $\hat{\nu}$ available, we then solve for $\hat{\overline{p}}$ and $\hat{\underline{p}}$ using (27). This leads to a support set with limits $\hat{\underline{p}} = [-12.9981, -1.1488, -0.8339, 0.0012, -0.0379, -0.0006, 0.0032]^\top$, and $\hat{\overline{p}} = [13.8920, -1.1488, -0.8339, 0.0012, -0.0379, 0.0001, 0.0032]^\top$. Therefore, whereas the first and sixth component of p vary in a range, the other ones can be treated as fixed constants. The performance of RPM F is given by both $E_x[\nu_y] = 180.3824$ and $E_x[\delta_y] = 13.4714$, which are 1620% larger and 83% smaller than those of RPM D. Figure 3 shows RPM F. Note that the containment condition $I_\sigma \subseteq I_y$ holds for all $x \in X$. The σ-volumes and limits of I_y appear to be centered about the LS prediction. This is not the case for other values of $\sigma_{\max}$ (not shown). Because most of the observations are close to the mean prediction, it is natural to expect that neglecting a few observations will considerably improve the model's performance.

We now derive a Type-4 RPM for $\lambda = 143/150$. The solution to OP3 indicates that the most significant variances are $\hat{\nu}_1 = 6.3378$, and $\hat{\nu}_2 = 3.1509$. With $\hat{\nu}$ available, we then now solve for $\hat{\overline{p}}$ and $\hat{\underline{p}}$ using (27). OP5 yields a support set P with limits $\hat{\underline{p}} = [-10.2605, -3.8559, -0.8480, 0.0002, -0.0420, -0.0002, 0.0032]^\top$, and $\hat{\overline{p}} = [2.9670, 0.016, -0.8200, 0.0022, -0.0338, -0.0001, 0.0032]^\top$. Therefore, according to their ranges' size; the first, second and fifth components of p contribute the most to the spread in the predicted output. The perfor-

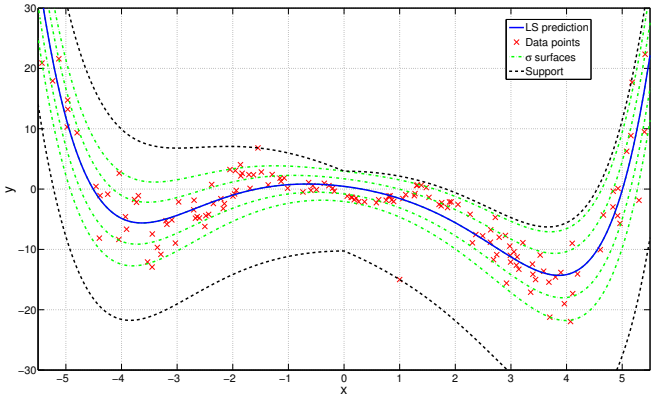

Figure 4: RPM G: Type-4 RPM for $\lambda = 143/150$ and $\sigma_{\max} = 1$.

mances of the resulting empirical model, shown in Figure 4 and called RPM, are $E_x[\nu_y] = 36.3341$ and $E_x[\delta_y] = 12.3649$. These values are 246% larger and 39% smaller than those of RPM E, respectively. The containment condition $I_\sigma \subseteq I_y$, which will be used to quantify the models reliability, does not hold at $x = 0$ (not seen in Figure 4). The support set of the CDF of p_1 is not centered about its mean value of -0.8734 (Crespo et al. 2015). This causes a sizable offset between the mean prediction and the midpoint function $(\underline{y}(x) + \overline{y}(x))/2$. Further notice that the high-probability region of the random process contains most of the observations whereas the outliers only affect $I_y(x)$. The limits $I_y(x)$, which have a derivative discontinuity at $x = 0$, do not coincide with any σ-surface. As expected, the comparison of RPM D with RPM F; and of RPM E with RPM G, indicate that the improvements in probabilistic performance $E_x[\nu_y]$ cause a degradation of the non-probabilistic performance $E_x[\delta_y]$.

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 model. These requirements are cast in terms of an output y belonging to a σ-volume $I_\sigma(x)$ for Type-1 and Type-2 RPMs, and also to the range $I_y(x)$ for Type-3 and Type-4 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 output 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 et al. 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*

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

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

The reliability of Type-1 IPMs, to be denoted as $\mathcal{I}$, is defined as

$$r(\mathcal{I}) = \text{Prob}_{\mathbb{P}}\left[(x, y) \in I_y\left(x, \hat{\overline{p}}, \hat{\underline{p}}\right)\right]. \tag{30}$$

The convexity of the OP1 enables the direct application of Theorem 1. The reliability of Type-1 and Type-2 RPMs, denoted by $\mathcal{R}$, is defined as

$$r\left(\mathcal{R}\right) = \text{Prob}_{\mathbb{P}}\left[(x, y) \in I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))\right]. \tag{31}$$

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 (23) 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

$$k^* = \text{floor}[N(1 - \lambda)], \tag{32}$$

are required by the auxiliary program, the reliabil-

ity of Type-2 RPMs is given by Theorem 1 with $k = k^*$. These k^* observations fall outside the optimal σ-volume and satisfy $F_{\kappa(\hat{\nu})}(\kappa) > \lambda$.

The reliability of Type-3 and Type 4 RPMs is considered next. Denote by $\hat{\mathcal{R}}$ any of such RPMs. The reliability of $\hat{\mathcal{R}}$ is defined as

$$r(\hat{\mathcal{R}}) = \text{Prob}_{\mathbb{P}}\left[(x, y) \in \mathcal{S}\right], \tag{33}$$

where $\mathcal{S} = I_y(x, \hat{\overline{p}}, \hat{\underline{p}}) \cap I_\sigma(x, \mu, \sigma_{\max}, \hat{\nu}(\lambda))$. The following theorem enables calculating $r(\hat{\mathcal{R}})$.

Theorem 2 *: 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 and* $\hat{\mathcal{R}}$ *be a Type-3 or Type-4 RPM. When the containment condition (26) holds, the reliability of* $\hat{\mathcal{R}}$ *is given by (28) with* $d = n_p$ *and* $k = k^* < N - d$*. Otherwise, the reliability of* $\hat{\mathcal{R}}$ *is given by (28) with* $\epsilon = \epsilon_1 + \epsilon_2$*, where* ϵ_1 *is given by (29) for* $d = 2n_p$ *and* $k = 0$*; and* ϵ_2 *is given by (29) for* $d = n_p$ *and* $k = k^* < N - d$*.*

Proof. When the containment condition holds, the two events defining the model's reliability are dependent and $r(\hat{\mathcal{R}}) = \text{Prob}_{\mathbb{P}}\left[(x, y) \in I_\sigma\right]$. In this case the reliability is given by Theorem 1 after applying the Principle of Equivalence to the non-convex formulations (24) or (23) to Type-3 and Type-4 RPMs respectively. In both cases $k = k^*$ as defined in (32). When set containment does not hold, use the bound $r(\hat{\mathcal{R}}) \geq \text{Prob}_{\mathbb{P}}\left[(x, y) \in I_y\right] + \text{Prob}_{\mathbb{P}}\left[(x, y) \in I_\sigma\right] - 1$. This bound is generally loose, so the actual model's reliability is probably larger. Each of the two events in (33) will be considered separately. Since the event $(x, y) \in I_y(x)$ is enforced by solving the convex program in (9) or (27) with N observations, we can readily bound its probability using Theorem 1 for $d = 2n_p$ and $k = 0$. This leads to ϵ_1. Conversely, the event $(x, y) \in I_\sigma(x)$ is enforced by solving the non-convex programs in (24) for a Type-3 RPM, and (23) for a Type-4 RPM. The principle of equivalence enables evaluating the probability of this event by considering an auxiliary convex program for which k^* out of the N observations are discarded. This leads to ϵ_2. Theorem 2 results from substituting these expressions into Theorem 1. □

Example 3: The reliability of RPM D and E, which are Type-3 RPMs, is considered first. Since neither model satisfies the Containment Condition (26), the reliability of each event must be added. Whereas the first event in (33), for which $N = 150$, $k = 0$ and $d = 14$, yields $1 - \epsilon_1 = 0.6984$ with confidence $1 - \beta = 0.99$; the second event, for which $N = 150$, $k = 0$ and $d = 7$, leads to $1 - \epsilon_2 = 0.8050$ with the same confidence. Therefore, the reliability of RPM D is no less than $1 - \epsilon_1 - \epsilon_2 = 1 - \epsilon = 0.503$ with confidence $1 - \beta = 0.99$. In the case of RPM E we have the same value for ϵ_1 as that for RPM D, whereas for the second event, for which $N = 150$, $k = 7$ and $d = 7$, leads to $1 - \epsilon_2 = 0.6984$ with confidence $1 - \beta = 0.99$. Therefore, the reliability of RPM D is no less than $1 - \epsilon_1 - \epsilon_2 = 1 - \epsilon = 0.3968$ with confidence $1 - \beta = 0.99$. Hence, discarding seven outliers improved performance by 74% at the expense of a reduction in the reliability of 10%. Finally, we will evaluate the reliability of RPM F and G, which are Type-4 RPMs. The containment condition holds for RPM F but not for RPM G. The reliability of RPM F, for which $N = 150$, $k = 0$ and $d = 7$, is no less than $1 - \epsilon = 0.8050$ with confidence $1 - \beta = 0.99$. In the case of RPM G, the first event in (33), for which $N = 150$, $k = k^* = 7$ and $d = 7$, leads to $1 - \epsilon_1 = 0.6984$ with confidence $1 - \beta = 0.99$; while the second event, for which $N = 150$, $k = 0$ and $d = 14$, leads to $1 - \epsilon_2 = 0.6984$ with confidence $1 - \beta = 0.99$. Therefore, the reliability of RPM G is no less than $1 - \epsilon_1 - \epsilon_2 = 1 - \epsilon = 0.3968$ with confidence $1 - \beta = 0.99$. The 30% reliability reduction of RPM G relative to RPM F is affected by the conservatism in Theorem 2. This illustrates the benefits of satisfying the containment condition. These results illustrate the typical trade-off between performance and reliability. These figures of merit should be traded off until the desired balance is reached. This balance can be reached by increasing the number of observations N, of outliers via λ, or by changing the model's structure via n_p, which prescribes d.

6 CONCLUSIONS

This and the companion paper (Crespo et al. 2015) present techniques for constructing random predictor models via optimization. These models enable a rigorous characterization of key features of the prediction, and of its reliability. Models with various degrees of fidelity are developed. This mathematical framework sets forth a new paradigm for the construction of empirical models in which the model's performance and reliability can be rigorously evaluated and traded-off.

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 1. In *ESREL 2015*, Zurich, Switzerland.