Minimal-Dispersion and Maximum-Likelihood Predictors with a Linear Staircase Structure

Luis G. Crespo[a*], Sean P. Kenny[b], and Daniel P. Giesy[c]

[a,b,c]Dynamic Systems and Controls Branch,
NASA Langley Research Center, Hampton, VA, 23681, USA

Abstract: This paper proposes techniques for constructing computational models describing the distribution of a continuous output variable given input-output data. These models are called Random Predictor Models (RPMs) because the predicted output corresponding to any given input is a random variable. We focus on RPMs having a linear parameter dependency, a bounded support set and prescribed functions for the mean, and the second-, third-, and fourth-order central moments. These constraints are realized by describing the model parameters as staircase random variables. The high versatility of such variables, and their low computational cost enable the efficient generation of possibly skewed and/or multimodal RPMs over an input-dependent interval. Optimization-based strategies for calculating RPMs using several optimality criteria are developed. These criteria include a moment-matching formulation, presented in a companion paper, and a minimal-dispersion and a maximum-likelihood formulations presented herein. The computational demands of the first two approaches, which separate distribution-free and distribution-fixed steps, are kept low by not requiring the simulation of solution candidates during the search for the optimal RPMs.

Keywords: Predictor models, staircase variables, model calibration, moments.

1. INTRODUCTION

Metamodeling [1] is the process of creating a mathematical representation of a phenomenon based on input-output data. Metamodeling techniques can be parametric or non-parametric. In the parametric case, the functional form by which the output depends on the input is first prescribed using a model M, and then the parameters of such a model are characterized. This step is commonly referred to as to model calibration. The approach proposed below falls into this category.

Bayesian inference [2] is often used for model calibration. In Bayesian calibration, the objective is to describe the parameters of a model as a vector of possibly dependent random variables by using Bayes' rule. 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 M. 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. Making the prediction match the observations by adjusting the hyper-parameters of a distribution is a long standing approach used in reliability-based design optimization, moment matching algorithms, and backward propagation of variance [3, 4, 5, 6]. In spite of its high computational demands, which entail simulating the predictor for each candidate combination of the calibrating variables, and of the potentially high sensitivity of the posterior to the assumed prior, this method is regarded as a benchmark in model calibration.

2. PROBLEM STATEMENT

A Data Generating Mechanism (DGM) is postulated to act on a vector of input variables, x, to produce an output, y. In this article, the focus will be on the single-output ($n_y = 1$) multi-input ($n_x \geq 1$) case.

* Corresponding author, Luis.G.Crespo@nasa.gov

The dependency of the output on the input is arbitrary. Assume that N Independent and Identically Distributed (IID) input-output pairs are obtained from a stationary DGM, and denote by $\mathbb{D} = \{x^{(i)}, y^{(i)}\}$, $i = 1, \ldots N$, the corresponding data sequence. The main objective of this article is to generate a model of the DGM based on $\mathbb{D}$. In a parametric model the output depends on both the input and the parameter through an equation. Denote by $y = M(x, p)$ such an equation, where $p \in \mathbb{R}^{n_p}$ are the model parameters. Instead of the standard practice of trying to match all the data as closely as possible with M evaluated at a single point p, the thrust in this work is to characterize p by either a bounded set P or by the joint PDF $f_p(p)$ supported in P. In both cases, the prescription of P must ensure that each data point in $\mathbb{D}$ can be fit exactly by the model evaluated at least one element of p in such a set. For a fixed value of the input x, and as long as P is a connected set and $M(x, p)$ is a continuous function (the only cases we will consider), the propagation of P through M yields an interval of output values. Thus, these models are called Interval Predictor Models (IPM). The desired IPM is a narrow interval of output values where unobserved data will likely fall. Conversely, for a fixed value of the input x, the propagation of $f_p(p)$ through M yields a random variable. Thus, these models are called RPMs. The desired RPM accurately describes the probability distribution governing the DGM.

Optimization-based strategies for calculating RPMs having a linear parameter dependency were developed. These optimality criteria include a moment-matching formulation, presented in the companion paper [7], as well as a minimal-dispersion and maximum-likelihood formulations, presented below. The background supporting the developments that follow is available in [7].

3. RANDOM PREDICTOR MODELS

An RPM is a mapping that assigns to each input vector $x \in X$ a corresponding random variable $R_y(x)$. A non-parametric RPM is the random variable-valued map given by

$$R_y(x) = \{f_{y(x)}(y),\ y(x) \in \Delta_y(x)\}, \tag{1}$$

where $f_{y(x)}$ is the PDF of y at $x \in X$ having the support set $\Delta_y(x) = [\underline{y}(x), \overline{y}(x)] \subseteq Y$. By contrast, 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 PDF $f_p(p)$ supported in Δ_p, so

$$R_y(x, f_p) = \{y = M(x, p),\ p \sim f_p(p),\ p \in \Delta_p\}. \tag{2}$$

Attention will be limited to the case in which the output depends linearly on p and arbitrarily on x, i.e.,

$$y = p^\top \varphi(x), \tag{3}$$

where $\varphi(x)$ is an arbitrary basis. This structure enables the analytical description of the moments of the output in terms of the moments of the parameter. These expressions, fully prescribed in [7], can be written as

$$\mu_{y(x)} = h_\mu(\mu, x), \tag{4}$$

$$m_{2,y(x)} = h_{m_2}(\mu, m_2, x), \tag{5}$$

$$m_{3,y(x)} = h_{m_3}(\mu, m_2, m_3, x), \tag{6}$$

$$m_{4,y(x)} = h_{m_4}(\mu, m_2, m_3, m_4, x). \tag{7}$$

Means to characterize $f_p(p)$ in (2) such that the resulting RPM accurately describes the distribution of the data are presented next.

3.1. Minimal-Dispersion RPMs

This section presents a strategy for generating RPMs having minimal dispersion from the data. As in the moment-matching RPMs, we will first perform a search for the optimal moments of the parameters in a distribution-free setting, and then prescribe a staircase random variable for each parameter that matches such moments. In contrast to the moment-matching approach, however, the minimal dispersion RPMs do not require setting target moment functions. The formulation for calculating minimal dispersion RPM is presented next.

Assume that the n_p parameters in (3) are independent random variables realizing the support bounds and moments given by $\theta_{p_1}, \dots \theta_{p_{n_p}}$ respectively. A minimal dispersion RPM is constrained to satisfy

$$\left\langle \hat{\theta}_{p_1}, \dots \hat{\theta}_{p_{n_p}} \right\rangle = \underset{\theta_{p_1},\dots,\theta_{p_{n_p}}}{\arg\min} \left\{ \frac{\|c\|}{N} : g(\theta_{p_i}) \leq 0,\ y^{(j)} \in I_\alpha\left(x^{(j)}\right),\ i = 1,\dots n_p,\ j = 1,\dots N \right\}, \tag{8}$$

where $c \in \mathbb{R}^N$ is given by

$$c_j = \int_{\underline{y}(x^{(j)})}^{\bar{y}(x^{(j)})} \left(y^{(j)} - y\right)^2 f_{y(x^{(j)})}(y)\, dy = \left(y^{(j)} - \mu_{y(x^{(j)})}\right)^2 + m_{2,\, y\left(x^{(j)}\right)}, \tag{9}$$

$f_{y(x)}(y)$ is the PDF of the RPM at x, $I_\alpha(x) = [y_\alpha(x), y_{1-\alpha}(x)]$ is an approximation[1] to the prediction interval bounded by the α and $1-\alpha$ quantiles, and $\mu_{y(x)}$ and $m_{2,\, y(x)}$ are given by (4) and (5).

Hence, we seek an RPM that minimizes the average dispersion of the prediction from the data, measured by c, while enclosing the data within the high-probability region $I_\alpha(x)$. Note that the less dispersed $f_{y(x)}(y)$ is about the observed output y_j, the smaller c_j. An ideal prediction will have a dual effect: it will place the mean at the observed output while minimizing the variance. The first set of inequality constraints ensures feasibility, whereas the second set enforces data containment by $I_\alpha(x)$. Hence, the minimization of the cost reduces the dispersion of the process from the data, whereas the second set of constraints ensure that such a dispersion is sufficiently large.

Multiple approximations to $I_\alpha(x)$ can be used. In this paper we use

$$y_\alpha(x) = \mu_{y(x)} - n_1 \sqrt{m_{2,\, y(x)}} - n_2 \sqrt[3]{m_{3,\, y(x)}}, \tag{10}$$

$$y_{1-\alpha}(x) = \mu_{y(x)} + n_1 \sqrt{m_{2,\, y(x)}} - n_2 \sqrt[3]{m_{3,\, y(x)}}, \tag{11}$$

where $n_1 > 0$, $n_2 \geq 0$, $\mu_{y(x)}$ is given by (4), $m_{2,\, y(x)}$ by (5), and $m_{3,\, y(x)}$ by (6). If $n_2 = 0$, only the support bound, the mean and the variance of each parameter will be solved for. If $n_2 > 0$, we will also be solving for the third-order central moment. The most general case, in which n_1 and n_2 are arbitrary functions of all the components of θ, requires using the full set of feasibility constraints. Being able to accurately describe the high-probability region $I_\alpha(x)$ in terms of only the design variables is instrumental in making Equation (8) distribution-free. Equation (9) indicates that a minimal dispersion RPM implicitly assumes the targets $\tilde{\mu}_{y(x^{(i)})} = y^{(i)}$ and $\tilde{m}_{2,\, y(x^{(i)})} = 0$.

The likelihood of the data for a minimal dispersion RPM might not be maximal. For instance, consider the degenerated case in which there is no input dependency, and the RPM is a random variable based on the output data $y^{(1)}, \dots y^{(N)}$. In this case the PDF will peak at the sample mean, where no data might

[1] This interval-valued function must only depend on the components of θ for the formulation to be distribution-free.

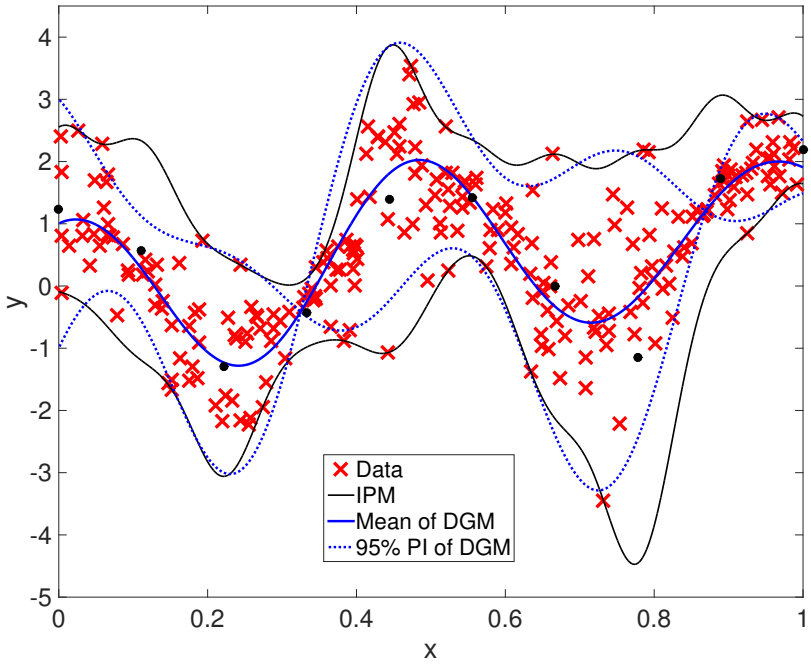

Figure 1: Data points, 95% prediction interval of the DGM, and corresponding IPM.

be present. As such, this technique is well suited for cases in which most of the probability associated with the DGM concentrates about its mean function. In such a case, the formulation in (8) will render predictors similar to those found via Maximum Likelihood Estimation (MLE). The notable difference between both approaches however, will be not having to simulate the process during the search for a minimal dispersion RPM. With the solution to (8) at hand, staircase variables that realize the resulting values for $\hat{\theta}_{p_1}, \ldots \hat{\theta}_{p_{n_p}}$ can be readily calculated. Note that (8) is a distribution-free step whereas the latter step is not.

Example 1: Next we consider a single-input single-output Gaussian DGM having input-dependent mean and variance functions in $X = [0,1]$. Figure 1 shows the mean function as well as the prediction interval corresponding to the 95% quantiles. Notice that the variance reduces to zero at $x = 0.32$ and $x = 0.87$. An IPM based on $N = 250$ data points, shown as red $\times$'s, was derived first. The IPM uses the basis $\varphi_i(x) = e^{-(x-u_i)^2/v_i}$, $i = 1, \ldots, n_p$, where $u_i \in \mathbb{R}$ is a center and $v_i \in \mathbb{R}$ is a length-scale parameter. We choose uniformly distributed centers over X, and length scale parameters equal to 1/15. Note that the IPM boundaries, shown as black solid lines, are driven by the outer-most data points of the ensemble, and whereas some regions within the IPM have low/zero-probability, e.g., the prediction at $x = 0.87$, some regions outside the IPM are likely to occur, e.g., the prediction at $x = 0$.

The cost in (8) corresponding to an RPM having a uniformly distributed p over the parameter box $\hat{P}$ is 1.4586. This RPM however, violates the data containment constraint.

A minimal dispersion RPM was derived using the same data sequence and basis used for the IPM. Furthermore, we fixed the bound of the support set Ω_p to

$$\Omega_p(\gamma) = \left\{ p : -\gamma(\hat{\overline{p}} - \hat{\underline{p}}) \leq 2p - \hat{\underline{p}} - \hat{\overline{p}} \leq \gamma(\hat{\overline{p}} - \hat{\underline{p}}) \right\}, \tag{12}$$

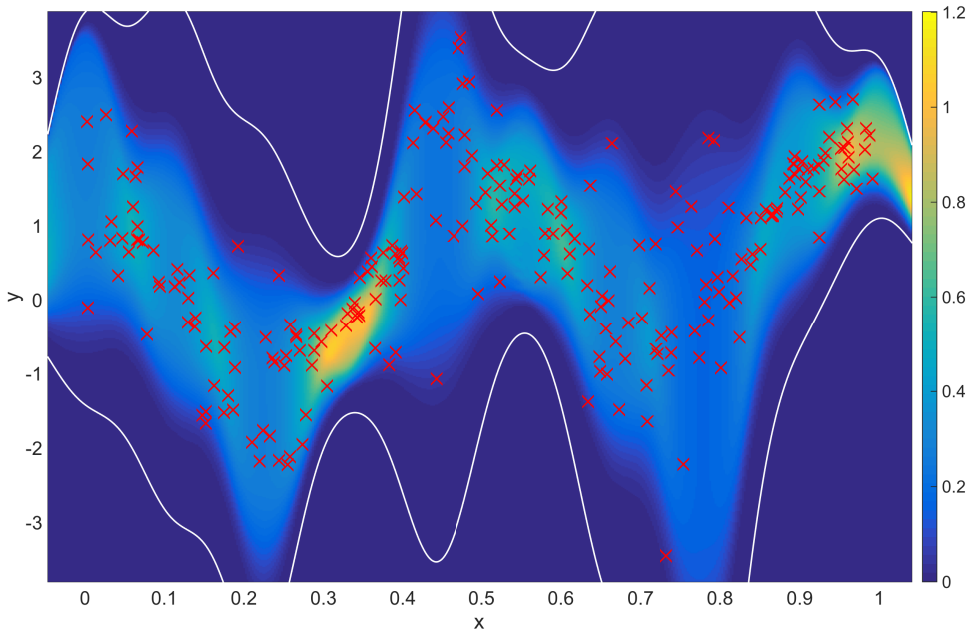

Figure 2: Minimal-dispersion RPM.

where $\underline{\hat{p}}$ and $\overline{\hat{p}}$ are the defining vertices of the uncertainty box $\hat{P}$ of a data-enclosing IPM [7]. Data containment is achieved as long as $\gamma \geq 1$. Therefore, we will be searching for the optimal moments of the parameters given an outer bound on their support set. The high-probability region $I_\alpha(x)$ was built using Equations (10) and (11) for $n_1 = 2$, which corresponds to the prediction interval of a Gaussian random variable for $\alpha = 0.025$, and $n_2 = -1$. Furthermore, we assume the staircase structure $S_p(\theta_p, 500, E)$ for all $n_p = 10$ parameters and $\gamma = 2$. Figure 2 shows the PDFs of the resulting RPM. The limits of $\Omega_{y(x)}$ (white lines) and the data (red $\times$'s) are also shown. Note that the PDF peaks in the areas of high concentration of data (regions in yellow), while exhibiting enough probability dispersion to cover the entire data ensemble. This dispersion captures well the strong input-dependency of the data. This figure shows that even though the predictor is not Gaussian, the RPM accurately represents the Gaussian DGM. The cost of the minimal dispersion RPM is 1.6304, which is about 13% greater than that for the uniform RPM referred to above. This increase in the cost is needed to satisfy the data enclosing constraint.

Figure 3 shows the staircase variables corresponding to $\Omega_p = \Delta_p$. The assumed form for $I_\alpha(x)$ makes Equation (8) independent of the fourth-order moments. Therefore, these moments are prescribed indirectly by choosing the entropy E as the cost J.

Values of γ in (12) near one yield RPMs that not only attain a greater dispersion/cost but might also have a multimodal distribution. This is a consequence of using an overly tight and suboptimal Ω_p in Equation (8). By fixing Ω_p upfront we are artificially restricting the probabilistic performance the RPM can attain. The optimal/smallest dispersion is obtained when the six components of all the θ's are used during optimization. This practice however, will not prevent the occurrence of multimodality.

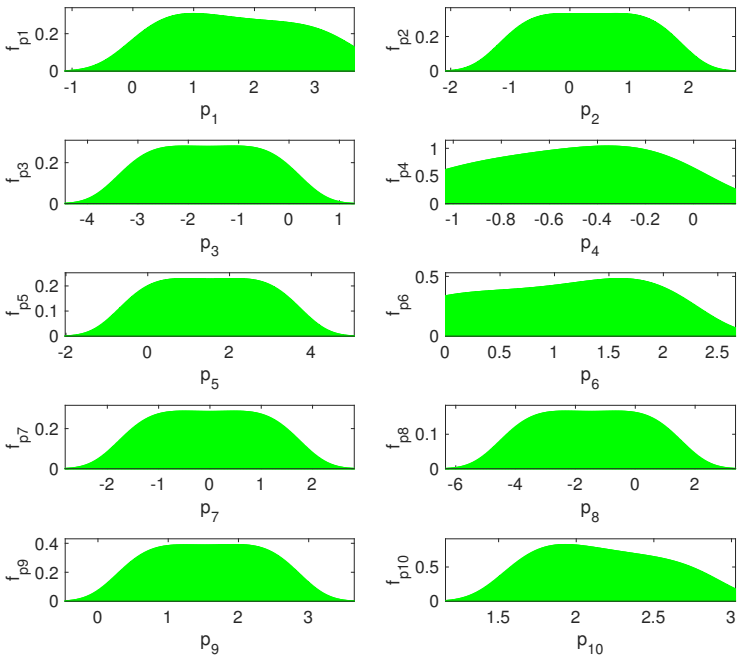

Figure 3: PDFs of the parameters of the minimal-dispersion RPM.

3.2. Maximum-Likelihood RPMs

This section presents a technique in which the hyper-parameters of the variables describing p are tuned using a maximum likelihood technique. To this end, we define the Zero-Prediction-Error (ZPE) manifold corresponding to the data point $\{x^{(i)}, y^{(i)}\} \in \mathbb{D}$ as

$$Z_i = \left\{ p : y^{(i)} = M(x^{(i)}, p) \right\}. \tag{13}$$

The uncertainty set P corresponding to any data containing IPM/RPM must intersect all ZPE manifolds $Z^{(1)}, \ldots Z^{(N)}$. Unless the intersection between P and Z_i is at one of its verteces, there are infinitely many parameter realizations, thus infinitely many model predictions $y = M(x, p)$, $p \in P$, passing through the data point $\{x^{(i)}, y^{(i)}\}$. When the model is given by (3), the ZPE manifolds are hyperplanes.

The formulation below seeks RPMs that maximize the likelihood of the ZPE manifolds. Evaluating the likelihood on the parameter space p rather than on the input-output space $X \times Y$ eliminates the need for simulating the random process at each solution candidate considered in the search for the optimal RPM (i.e., reconstructing the random process and evaluating the likelihood at each of the N data points). Therefore, this technique is more computationally efficient than such an alternative. In contrast to the other techniques proposed thus far, and because the likelihood is not uniquely dependent on θ, this approach requires assuming a structure for the random variables describing p up front. Hence, this is a distribution-fixed approach. In this section we will assume the parameters in p are staircase variables having a fixed number of bins n_b. Any other family of random variables can be used instead. However, we expect the versatility of the staircase variables to render a greater likelihood of the data. The formulation for generating a maximum log likelihood RPM is as follows.

Assume that the parameters p in (3) are the independent staircase random variables $p_j \sim S\left(\theta_{p_j}, n_b, J\right)$,

$j = 1, \ldots n_p$, where n_b is fixed. A maximum log likelihood RPM is given by

$$\left\langle \hat{\theta}_{p_1}, \ldots \hat{\theta}_{p_{n_p}} \right\rangle = \underset{\theta_{p_1}, \ldots \theta_{p_{n_p}}}{\arg\max} \left\{ \frac{1}{N} \sum_{i=1}^{N} \log\left(L_i\right) : \theta_{p_j} \in \mathbb{S}(n_b),\ \Omega_p \cap Z_i \neq \emptyset;\ i = 1, \ldots N,\ j = 1, \ldots n_p \right\}, \tag{14}$$

where the likelihood corresponding to the i-the data point, L_i, is given by the surface integral

$$L_i = \int_{\Omega_p \cap Z_i} f_{p(\theta_{p_1}, \ldots \theta_{p_{n_p}})} p \, ds, \tag{15}$$

and $f_{p(\theta_{p_1}, \ldots \theta_{p_{n_p}})}$ is the joint density PDF of p.

Hence, this RPM maximizes the average log likelihood of the ZPE manifolds over the set[2] Ω_p. The cost in (15) can be readily evaluated using sampling. When the uncertainty set Ω_p is fixed in advance, which is the case considered hereafter, the corresponding samples are independent of the remaining optimization variables. Consequently, the samples falling into $\Omega_p \cap Z_i$ will only have to be generated once, substantially mitigating the computational demands of the algorithm. Furthermore, by choosing the set in Equation (12) the data containment constraint will be satisfied by design. The case in which the parameters in p are not independent can be considered by augmenting the set of design variables with the hyper-parameters of a copula with staircase marginals. This practice, however, will not be carried out here.

Example 2: A maximum likelihood RPM based on the same data, IPM and Ω_p of Example 1 was generated using (14). The staircase structure $S_p(\theta, 500, E)$ was assumed for all $n_p = 10$ parameters. Figure 4 shows the PDFs of the RPM in the input-output space. The RPM allocates most of the probability near the data as intended. Notice the multimodal predictions near $x = 0.36$ and $x = 0.55$. The solution to (14) maximizes the likelihood of reproducing the data without any consideration on the resulting shape of the RPM, so there is no basis to expect convergence to a prediction assuming any particular form. The average log likelihood for this RPM is 2.4318. For comparison sake, the cost corresponding to an RPM having a uniform distribution over Ω_p is -0.6016.

The comparison of Figures 2 and 4 indicates that the maximum-likelihood RPM allocates more probability near the data than the minimum-dispersion RPM. However, the spurious multimodal prediction and the zero likelihood subdomains occurring at $x = 0.33$ and $x = 0.55$ are undesirable. These anomalies can be avoided by further restricting the feasible design space Θ, e.g., forcing the parameters to by simply supported unimodal variables, at the expense of a reduction in the value of the optimal likelihood [8].

Figure 5 shows the corresponding staircase variables. Note that the PDFs of p_1, p_4, p_6, p_9 and p_{10} are multimodal with p_6 exhibiting a large spike. The multimodal nature of the parameters yields the multimodal prediction seen at some input values. Note however that desirable features on the PDFs of the parameters might not necessarily transfer to the predicted output, e.g., a linear combination of unimodal PDFs for all p's does not necessarily yield a unimodal $y(x)$.

[2] One could alternatively evaluate the moments of y given those of p using Equations (4-7), calculate a staircase variable at each observed input $x^{(i)}$ and then evaluate the likelihood of $y^{(i)}$ for all $i = 1, \ldots N$. This practice requires calculating N staircase variables for each optimization point found in the search for the optimum, instead of the considerably less n_p variables required by (14). Having to solve n_p optimization programs is computationally viable thanks to the convex structure of the staircase variables.

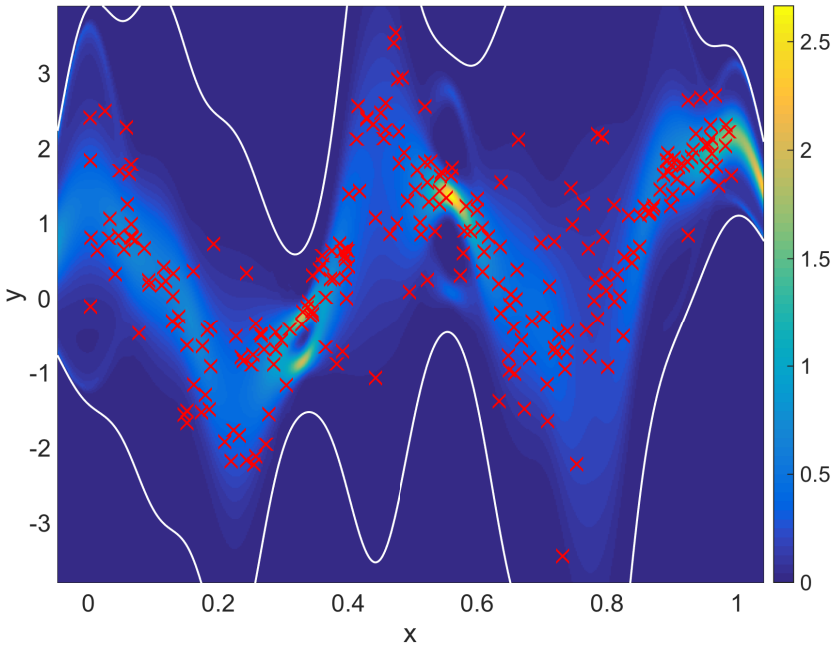

Figure 4: Maximal-likelihood RPM.

4. DISCUSSION

The minimal-dispersion approach is well suited for DGMs having a unimodal distribution concentrated about its mean. The predicted third- and fourth-order moments of the resulting RPM solely depend on the manner by which the high probability region I_α is parameterized. The computational demands of this approach, which separates distribution-free and distribution-fixed steps, are kept low by not requiring the simulation of RPM candidates during the search for the optimal RPM. Conversely, the maximum likelihood approach requires assuming a class of distributions for the model parameters up front. Furthermore, it requires calculating such distributions repeatedly making the method more computationally expensive. The versatility of a predictor is intrinsically linked to that of the assumed distribution structure for its parameters. As such, staircase variables are particularly suitable for describing complex DGMs. However, all this freedom might render predictors having undesirable spikes and spurious multimodal distributions. This can be avoided by further restricting the feasible space Θ or by relaxing some of the staircase constraints [8].

The strategies proposed can be naturally extended to the case in which the parametric model depends arbitrarily on p, and to the case in which the staircase variables are dependent. These situations, however, will require simulating the predictor candidates repeatedly, practice that will substantially increase the computational cost.

References

[1] T. Simpson, J. Peplinski, P. Koch, and J. Allen, "Metamodels for computer-based engineering design: survey and recommendations," *Engineering with Computers*, vol. 17, no. 1, pp. 129–150, 2001.

[2] M. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal*

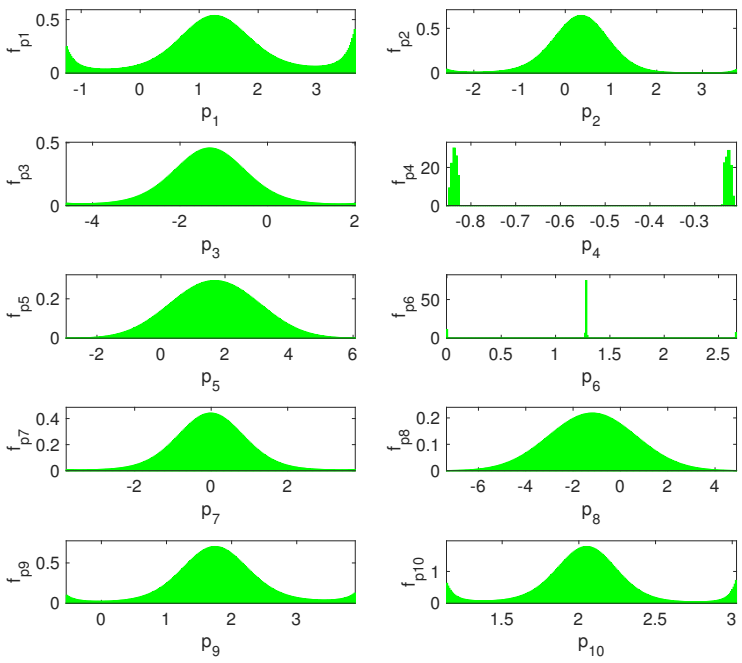

Figure 5: PDFs of the parameters of the maximum-likelihood RPM.

Statistical Society B, vol. 63, no. 3, pp. 425–464, 2001.

[3] M. Allen and K. Maute, "Reliability-based design optimization of aeroelastic structures," *Structural and Multidisciplinary Optimization*, vol. 27, pp. 228–242, 2004.

[4] M. S. Eldred, H. Agarwal, V. Perez, S. Wojtkiewicz, and J. Renaud, "Investigation of reliability method formulations in DAKOTA/UQ," *Structure and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and Performance*, vol. 3, no. 3, pp. 199–213, 2007.

[5] L. P. Swiler, B. Adams, and M. Eldred, "Model calibration under uncertainty: Matching distribution information, AIAA-2008-5944," in *AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, Canada*, September 2008.

[6] C. C. McAndrew, *Compact Modeling: Principles, Techniques, and Applications: Chapter 16 in Statistical Modeling using Backward Propagation of Variance*. New York: Springer, 2010.

[7] L. G. Crespo, D. P. Giesy, and S. P. Kenny, "Moment-matching predictors with a linear parameter dependency," in *Probabilistic Safety Assessment and Management conference, PSAM 14, Los Angeles, CA, USA*, September 2018.

[8] L. G. Crespo, D. P. Giesy, and S. P. Kenny, "On the calculation and shaping of staircase random variables," in *ESREL 2017, Portoroz, Slovenia*, June 2017.

[9] L. G. Crespo, S. P. Kenny, and D. P. Giesy, "Interval predictor models with a linear parameter dependency," *ASME Journal of verification, validation and uncertainty quantification*, vol. 1, no. 2, pp. 1–10, 2016.