Statistica Sinica **28** (2018), 899-919
doi:https://doi.org/10.5705/ss.202016.0217

COMPUTER EXPERIMENTS: PREDICTION ACCURACY, SAMPLE SIZE AND MODEL COMPLEXITY REVISITED

Ofir Harari[1], Derek Bingham[1], Angela Dean[2] and Dave Higdon[3]

[1]*Simon Fraser University,* [2]*The Ohio State University and* [3]*Biocomplexity Institute of Virginia Tech*

Abstract: We revisit the problem of determining the sample size for a Gaussian process emulator and provide a data analytic tool for exact sample size calculations that goes beyond the $n = 10d$ rule of thumb and is based on an IMSPE-related criterion. This allows us to tie sample size and prediction accuracy to the anticipated roughness of the simulated data, and to propose an experimental process for computer experiments, with extension to a robust scheme.

Key words and phrases: Computer experiments, Gaussian processes, sample size calculation.

1. Introduction

The Gaussian process model was proposed by Sacks et al. (1989) as a statistical emulator for deterministic computer codes, and a large body of literature has subsequently been devoted to the exploration of its performance under various conditions. Experimental design for computer experiments has been extensively investigated, whether space-filling designs (see e.g. Johnson, Moore and Ylvisaker (1990), Joseph, Dasgupta and Wu (2012)) or optimal designs driven by different statistical criteria (e.g. Sacks, Schiller and Welch (1989), Shewry and Wynn (1987), Harari and Steinberg (2014)).

One might think that the design of experiments for deterministic computer model emulation would be a topic with little left to study. Still, Only a small amount of literature has been dedicated to the foundational topic of sample size in computer experiments, and practitioners often make the naive assumption that the sample size should grow linearly with the input dimension, known as the "$n = 10d$" rule of thumb, where d is the number of inputs to the computer model (Chapman et al. (1994), Jones, Schonlau and Welch (1998)). A partial justification for the latter is given by Loeppky, Sacks and Welch (2009), who advocate the $n = 10d$ rule in specific cases. Their message is that, for relatively uncomplicated surfaces and moderate d, good prediction accuracy can be

obtained with $n = 10d$ observations in an initial experiment, and increasing n further can increase accuracy further. However, if $n = 10d$ observations results in poor accuracy, (which tends to happen in complicated or high dimensional codes with input factors having similar complexity), then the improvement in accuracy through adding more runs tends not to be helpful.

In this paper, we take the view that the choice of experimental design ought to take into account the prior belief about the complexity of the response surface, the desired prediction accuracy, and the available resources.

The underlying structural assumptions about the approximated response surface by embracing the $n = 10d$ rule are far-reaching, and are often not fully understood. For $d = 20$, for example, $n = 200$ would not even suffice to estimate a linear model with 20 main effects and all 190 two-factor interactions. Fitting a Gaussian process model using so few data points must then reflect the belief by experimenter that either the change in the response is purely additive in many of the factors or that several factors are entirely inert.

Originally, small sample sizes were a consequence of lengthy simulation runtimes, numerical singularity in the correlation matrix and computational issues, stemming from the need to store and repeatedly invert a large $n \times n$ covariance matrix. With remedies in place (see e.g. Ranjan, Haynes and Karsten (2011), Kaufman et al. (2011)), in conjunction with improving computing capabilities, it is expected that large-scale computer experiments will routinely take place in the near future. It is easy to imagine models with a large number of inputs (≥ 100) where factor sparsity (Box and Meyer (1986)) implies that relatively few (≈ 20) of the inputs are important. In such cases, careful consideration of the model structure and the goals of the experiment are important.

The sample size for an experiment, the complexity of the model, and the prediction goals of the experiment are intimately related. In this paper we attempt to provide methodology for computer experiments to address their interrelationships.

This paper is organized as follows. Section 2 provides a brief introduction to the Gaussian process model commonly used for computer model emulation. In addition, we take high prediction accuracy as an experimental objective and propose different interpretations for that goal. Section 3 ties the prediction accuracy to the sample size and the model hyperparameters that specify the response surface complexity, and discusses consequences for experimental design. In Section 4 we develop a methodical experimental process for computer experiments, and in Section 5 we demonstrate the proposed process on an experiment involving

a piston simulator. We then provide a robust scheme, in Section 6, to handle uncertainty with regard to the hyperparameters. Section 7 includes a discussion and some thoughts for future work.

For the reader's convenience, a web applications was created to accompany this paper and facilitate future analyses. For details, see the Supplementary Materials section.

2. Gaussian Processes Emulators for Deterministic Computer Models

2.1. Gaussian process regression

Building a computer model emulator can be viewed as nonparametric regression for deterministic simulators. The reasons for using the conventional specification for the Gaussian process (Sacks et al. (1989)) lie with its ability to interpolate the model output and to quantify uncertainty at unsampled inputs.

In this setting, the computer model output, $y(\boldsymbol{x})$, is viewed as a realization of a stationary, zero mean, Gaussian process with covariance function $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 R_{\boldsymbol{\theta}}(\boldsymbol{x} - \boldsymbol{x}')$. The *correlation function*, $R_{\boldsymbol{\theta}}(\cdot)$, depends on the vector of hyperparameters $\boldsymbol{\theta}$ that govern the correlation between responses at separate locations. Denote the experimental design $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the experimental region. We will assume, without loss of generality, that $\mathcal{X} = [0,1]^d$. Let $\boldsymbol{y} = [y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_n)]^{\mathsf{T}}$ be a vector of observations at the design points, and denote by $\boldsymbol{R}_{\boldsymbol{\theta}}$ the matrix whose entries are $R_{ij} = R_{\boldsymbol{\theta}}(\boldsymbol{x}_i - \boldsymbol{x}_j)$. Then for any $\boldsymbol{x} \in \mathcal{X}$, choosing the *Kriging* predictor

$$\widehat{y}(\boldsymbol{x}) = \mathbb{E}\left\{y(\boldsymbol{x}) \middle| \boldsymbol{y}\right\} = \boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}, \tag{2.1}$$

to predict $y(\boldsymbol{x})$ would yield the *Mean Squared Prediction Error* (MSPE)

$$\mathbb{E}\left[\{\widehat{y}(\boldsymbol{x}) - y(\boldsymbol{x})\}^2 \middle| \boldsymbol{y}\right] = \mathbb{V}\mathrm{ar}\left\{y(\boldsymbol{x}) \middle| \boldsymbol{y}\right\} = \sigma^2\left\{1 - \boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\}, \tag{2.2}$$

for $\boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{x}) = [R_{\boldsymbol{\theta}}(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, R_{\boldsymbol{\theta}}(\boldsymbol{x} - \boldsymbol{x}_n)]^{\mathsf{T}}$. For the rest of this paper we suppress the $\boldsymbol{\theta}$ subscript, keeping in mind that the correlation between responses at different locations depends heavily on these hyperparameters.

2.2. A measure for prediction accuracy

Using (2.2), the Integrated MSPE (IMSPE) of the Kriging predictor (2.1) is given by

$$\mathcal{J}(\widehat{y}, \mathcal{D}; \boldsymbol{\theta}) = \int_{[0,1]^d} \mathbb{E}\left[\{\widehat{y}(\boldsymbol{x}) - y(\boldsymbol{x})\}^2 \middle| \boldsymbol{y}\right] \mathrm{d}\boldsymbol{x}$$

$$= \sigma^2 - \sigma^2 \mathrm{tr}\left\{ \boldsymbol{R}^{-1} \int_{[0,1]^d} \boldsymbol{r}(\boldsymbol{x}) \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \mathrm{d}\boldsymbol{x} \right\}. \tag{2.3}$$

Weighted versions of (2.3) have been proposed to emphasize prediction in certain areas of the design region (see e.g. Sacks et al. (1989)).

As $y(\boldsymbol{x})$ is taken to be stationary, $\mathbb{V}\mathrm{ar}\{y(\boldsymbol{x})\} = \sigma^2$, we may then consider the normalized quantity

$$\frac{\mathcal{J}(\widehat{y}, \mathcal{D}; \boldsymbol{\theta})}{\sigma^2} = \int_{[0,1]^d} \frac{\mathbb{V}\mathrm{ar}\left\{y(\boldsymbol{x}) \middle| \mathcal{D}\right\}}{\mathbb{V}\mathrm{ar}\{y(\boldsymbol{x})\}} \mathrm{d}\boldsymbol{x} \tag{2.4}$$

as the average proportion of the variability of $y(\boldsymbol{x})$ that remains unexplained by design $\mathcal{D}$. This is reminiscent of the proportion of unexplained variability in linear regression models, typically calculated as the ratio of the sum of squares for error and the total sum of squares. In this case. however, (2.4) is for out of sample observations. Following this analogy, we can form the counterpart of the squared multiple correlation coefficient in regression.

Proposition 1. *Let* $y(\boldsymbol{x}) \sim \mathrm{GP}\left(0, \sigma^2 \mathrm{R}\right)$ *and let* $\widehat{y}(\boldsymbol{x}) = \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{y}$ *be the Kriging predictor of* $y(\boldsymbol{x})$*. Then*

$$1 - \frac{\mathcal{J}(\widehat{y}, \mathcal{D}; \boldsymbol{\theta})}{\sigma^2} = \overline{\rho}^2(y, \widehat{y}) := \int_{[0,1]^d} \rho^2\left(y(\boldsymbol{x}), \widehat{y}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}, \tag{2.5}$$

where $\rho(\cdot, \cdot)$ *denotes the correlation coefficient.*

Proof. First note that, from (2.1)

$$\mathrm{cov}\left(y(\boldsymbol{x}), \widehat{y}(\boldsymbol{x})\right) = \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \mathrm{cov}\left(y(\boldsymbol{x}), \boldsymbol{y}\right) = \sigma^2 \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{r}(\boldsymbol{x}) = \mathbb{V}\mathrm{ar}\{\widehat{y}(\boldsymbol{x})\}$$

and, since $\mathbb{V}\mathrm{ar}\{y(\boldsymbol{x})\} = \sigma^2$, we have

$$\rho^2\left(y(\boldsymbol{x}), \widehat{y}(\boldsymbol{x})\right) = \frac{\left\{\sigma^2 \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{r}(\boldsymbol{x})\right\}^2}{\sigma^2 \cdot \sigma^2 \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{r}(\boldsymbol{x})} = \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{r}(\boldsymbol{x}).$$

The result follows from (2.2) and (2.4).

We can interpret $\overline{\rho}^2(y, \widehat{y})$ in (2.5) as the average squared correlation among the simulator responses and predicted responses at unsampled inputs. Proposition 1 sheds new light on the interpretation of IMSPE-optimal designs (see e.g. Sacks et al. (1989)). The proposition demonstrates that minimizing the IMSPE is equivalent to maximizing the average, squared, out-of-sample correlation between $y(\boldsymbol{x})$ and $\widehat{y}(\boldsymbol{x})$. By minimizing the IMSPE we can expect to improve the predictive ability of the Kriging predictor.

Definition 1. *The Root Average Unexplained Variability (RAUV) of predictor* $\widehat{y}$*, evaluated at design* $\mathcal{D} \subset [0,1]^d$ *is*

$$\mathrm{RAUV}\left(\widehat{y};\mathcal{D},\boldsymbol{\theta}\right)=\left(\frac{\mathcal{J}\left(\widehat{y},\mathcal{D};\boldsymbol{\theta}\right)}{\sigma^2}\right)^{1/2}=\left(\int_{[0,1]^d}\frac{\mathbb{V}\mathrm{ar}\left\{y\left(\boldsymbol{x}\right)\middle|\mathcal{D}\right\}}{\mathbb{V}\mathrm{ar}\left\{y\left(\boldsymbol{x}\right)\right\}}\mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

We propose RAUV as a measure of expected prediction error on designing a computer experiment. Its indicated magnitude is relative to the prior standard deviation, and the choice of the square root scale is in line with similar measures used by Loeppky, Sacks and Welch (2009) and Chen et al. (2016). It is common practice to measure model performance by its Empirical Root Mean Square Error (ERMSE) computed on a holdout set, normalized by some measure of the variation in the data measured in the original units, such as the range or the empirical standard deviation. We have found sample sizes that warrant low a priori RAUV to be consistent with good empirical prediction accuracy, much more so than the average unexplained variability without the square root. It can also be justified equivalently in terms of uncertainty quantification: requiring RAUV ≤ 0.05 means that we want the square root of the average squared length of our prediction intervals to shrink by 95% once data is observed. From Proposition 1

$$\mathrm{RAUV}\left(\widehat{y};\mathcal{D},\boldsymbol{\theta}\right)=\sqrt{1-\overline{\rho}^2\left(y,\widehat{y}\right)}.$$

Thus, ensuring $\mathrm{RAUV}\left(\widehat{y}\right)\leq\varepsilon$, explaining at least $100\left(1-\varepsilon^2\right)\%$ of the variability in $y\left(\boldsymbol{x}\right)$ by the model and achieving $\left|\overline{\rho}\left(\widehat{y},y\right)\right|\geq\sqrt{1-\varepsilon^2}$ are equivalent for Gaussian process model fitting in the context of computer model emulation.

3. Model Complexity, Sample Size and Prediction Accuracy

We invoke the result of Micchelli and Wahba (1981) and Harari and Steinberg (2014) to shed some light on the link between the complexity of the model being estimated, the prediction accuracy and sample size.

Theorem 1. *If*

$$R\left(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta}\right)=\sum_{k=1}^{\infty}\lambda_k\varphi_k\left(\boldsymbol{x}\right)\varphi_k\left(\boldsymbol{x}'\right)$$

is the Mercer expansion of $R(\cdot;\boldsymbol{\theta})$ *on* $[0,1]^d$*, with eigenfunctions* $\{\varphi_k\left(\boldsymbol{x}\right)\}$ *and eigenvalues* $\{\lambda_k\}$*, then*

$$\inf_{\mathcal{D}}\left\{\frac{\mathcal{J}\left(\widehat{y};\mathcal{D},\boldsymbol{\theta}\right)}{\sigma^2}\right\}\geq\sum_{k\geq n+1}\lambda_k. \tag{3.1}$$

Over $[0,1]^d$, the eigenfunctions and eigenvalues can be solved numerically; in particular, for separable correlation functions, the problem reduces to a series

of univariate eigendecompositions. Details are provided in Harari and Steinberg (2014).

The inequality (3.1) encapsulates the complex relationship between sample size, model complexity (in the form of the Gaussian process hyperparameters) and prediction accuracy. One immediate result is the following.

Corollary 1. *Let* $\{\lambda_k\}$ *be the set of eigenvalues of* $R(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$. *Let* n^c *be the critical sample size required to achieve* RAUV $\leq \varepsilon$ *for some* $\varepsilon > 0$. *Then*

$$n^c \geq \min\left\{ n : \sqrt{\sum_{k \geq n+1} \lambda_k} \leq \varepsilon \right\} = \min\left\{ n : \sum_{k=1}^{n} \lambda_k \geq 1 - \varepsilon^2 \right\}. \tag{3.2}$$

Given some idea about the way in which the different inputs act, one can then, (at least approximately) derive analytically the required sample size for a given average level of prediction accuracy. Thus, to achieve an acceptable amount of unexplained variability, one has a choice of ε, and this depends on the particular application and its goals.

For the piston example of Section 5, Figure 7 there displays actual vs. predicted values for a large holdout set, where the predicted values are obtained from (2.1). The four panels show sample sizes of 30, 50, 70, and 120, respectively, which, for the hyperparameters $\boldsymbol{\theta}$ selected for the piston example, correspond to ε values of 0.191, 0.120, 0.086 and 0.045. In the top left panel of the figure, $n = 30$ and $\varepsilon = 0.191$, the quality of the fit is far from perfect, but this corresponds to 96% of the average explained variability. In our experience, agreement such as that indicated in top right panel of Figure 7 (with $n = 50$ and $\varepsilon = 0.1$) reflects, for deterministic simulators, an acceptable fit with at least 99% of the variability (on average) explained. An even better fit, in the bottom right corner of Figure 7, corresponds to $\varepsilon = 0.05$ (and $n = 120$).

Remark 1. Ideally, one would like an upper bound of the form

$$\inf_{\mathcal{D}} \left\{ \frac{\mathcal{J}(\widehat{y}; \mathcal{D}, \boldsymbol{\theta})}{\sigma^2} \right\} \leq \sum_{k \geq n+1} \lambda_k + \alpha^2(n, \boldsymbol{\theta})$$

to bound the critical sample size $n_{_L} \leq n^c \leq n_{_U}$, with $n_{_U}$ guaranteeing an RAUV below the desired threshold. Without an upper bound we treat inequality (3.1) roughly as an equality throughout this paper, and remember that we need to choose a slightly larger sample size than the one recommended by (3.2). This approach is supported by Figure 2 and the findings of the simulation study in Section 5.

We now proceed to use these ideas in a practical setting by fixing any two of

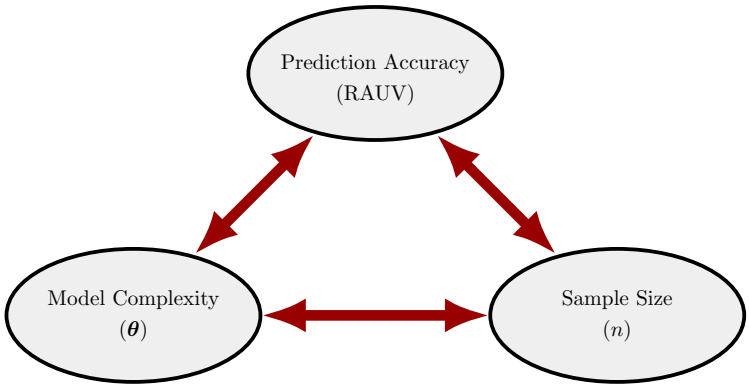

Figure 1. The three interacting elements of a (computer) experiment.

the three vertices of the triangle appearing in Figure 1 and observing the impact on the third vertex.

3.1. Sample size for a desired level of prediction accuracy and fixed $\boldsymbol{\theta}$

For fixed Gaussian process hyperparameters (model complexity) and a given level of prediction accuracy (RAUV), determination of the required sample size follows a simple application of Corollary 1.

Example 1. Consider the computer code used by Yi et al. (2005) to simulate ligand activation of G-protein in yeast. Loeppky, Sacks and Welch (2009) fixed five of nine factors and used a 4-dimensional Gaussian process emulator for the response, using the *squared-exponential* correlation function

$$R\left(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}\right) = \exp\left\{-\sum_{i=1}^{4} \frac{|x_i - x_i'|^2}{\theta_i}\right\}, \tag{3.3}$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ a vector of *correlation length* parameters.

Based on a design with $n = 80$ runs, they found the maximum likelihood estimates for the emulator to be $\widehat{\boldsymbol{\theta}} = (9.09, 1.59, 1.79, 0.56)$. Treating these for the moment as the actual parameters for the data-generating process, we can find the eigenvalue. Figure 2 shows the lower bound $\sqrt{\sum_{k \geq n+1} \lambda_k}$ for the RAUV versus n. If we set a threshold $\varepsilon = 0.05$, the smallest sample size for which the threshold is crossed is $n = 21$. Also plotted are the RAUV values for IMSPE-optimal designs of various sample sizes and the given $\boldsymbol{\theta}$. While the theoretical lower bound is closely approached by the empirical values, caution needs to be taken and a few more runs (in this example 5 or 6) may be needed to guarantee that the desired precision level is achieved. More conservatism is called for when the design to be

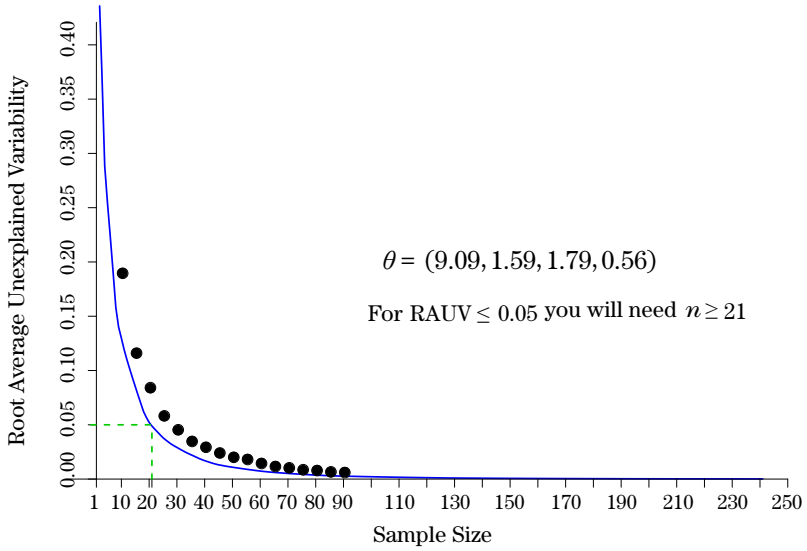

Figure 2. RAUV lower bound curve for the squared exponential correlation function, with the estimated correlation length parameters for the G-protein example from Loeppky, Sacks and Welch (2009). The dots in the figure denote RAUV values calculated for IMSPE-optimal designs.

used is not IMSPE-optimal. In this example both the lower bound curve and the empirical IMSPE values indicate that a sample size of $n = 10d = 40$ should be more than enough for an adequate fit (if parameter estimates are to be trusted), which is consistent with the findings of Loeppky, Sacks and Welch (2009).

Example 2. We look to find an n that will give an RAUV of $\varepsilon = 0.05$ for a Gaussian process model with the product Matérn correlation function

$$R\left(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\phi}, \boldsymbol{\nu}\right) = \prod_{i=1}^{d} \frac{1}{\Gamma(\nu_i) 2^{\nu_i - 1}} \left(\frac{2\sqrt{\nu_i}\,|x_i - x_i'|}{\phi_i} \right)^{\nu_i} \mathcal{K}_{\nu_i} \left(\frac{2\sqrt{\nu_i}\,|x_i - x_i'|}{\phi_i} \right),$$

where ϕ_i and ν_i are the correlation length and smoothness parameters along the ith direction, respectively, and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of order ν. Here we focus on an isotropic 4-dimensional process with $\phi_1 = \cdots = \phi_4 = 1$ and $\nu_1 = \cdots = \nu_4 = 5/2$ (to guarantee twice differentiable realizations). In this case the lower bound curve (see Figure 3) indicates that a sample size of $n \geq 112$ is required for the precision target we set for ourselves, and the $n = 10d$ rule with $d = 4$ is inadequate; a process whose realizations are harder to predict compared to a process that is based on the squared exponential correlation function, requires more observations for the same level of prediction accuracy.

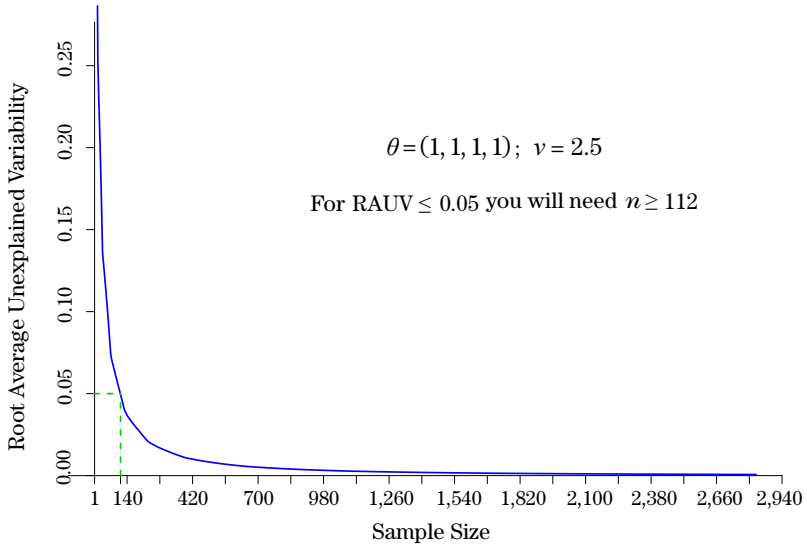

Figure 3. RAUV lower bound curve for the product Matérn correlation function of an isotropic process with $\nu = 5/2$ and $\theta = 1$.

3.2. Assessing prediction accuracy for fixed n and $\boldsymbol{\theta}$

With a limited computation budget, the experimenter is interested in anticipated quality of the predictions that can be achieved for the emulator. For instance, solving for eigenvalues and taking advantage of (3.1), the product Matérn correlation function specified in Example 2, $n = 40$ runs, leads to RAUV ≥ 0.128. If this level of inaccuracy is excessive, the experimenter may consider instead exploring only a sub-set of the inputs or increasing the computational budget.

3.3. Maximum model complexity for fixed n and prediction accuracy

Inequality (3.1) does not define a one-to-one relationship between the Gaussian process hyperparameters and the sample size required for a desired level of prediction accuracy. If, however, one only considers isotropic models, (3.2) can be inverted. Suppose, for example, that one wishes the most complicated isotropic Gaussian process model that can be investigated with prediction accuracy RAUV ≤ 0.05 and the product Matérn correlation function with $\nu = 5/2$ and $n = 40$, correlation length of $\phi = 1.46$ is the minimum value that results in $\sqrt{\sum_{k \geq 41} \lambda_k} \leq 0.05$. With this parameter determined, the experimenter can then produce Functional Analysis of Variance (FANOVA) plots (see Saltelli et al. (2008)) of realizations from an isotropic Gaussian process with $\phi = 1.46$.

Figure 4 displays 20 realizations from a univariate Gaussian process based

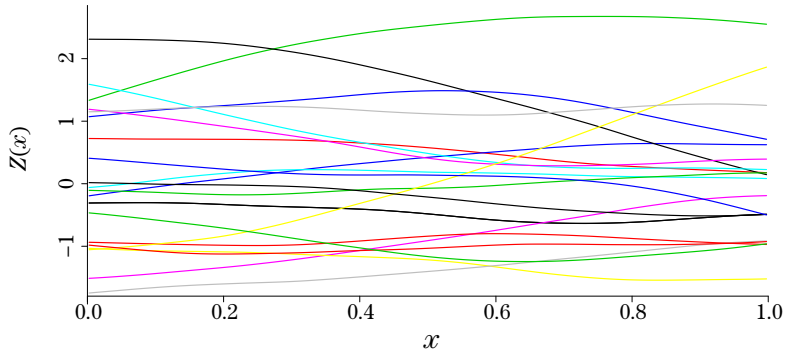

Figure 4. 20 realizations drawn from a univariate Gaussian process based on the product Matérn correlation function with $\phi = 1.46$ and $\nu = 5/2$.

on the product Matérn correlation function with $\phi = 1.46$ and $\nu = 5/2$. If the realizations demonstrate a complexity that is less than the experimenter's belief in the computer model then this ϕ is unlikely to achieve the desired level of precision. In that case, one may either find a way to increase the sample size or lower the a priori expectations with regard to prediction accuracy.

4. Stepping Through Design Process

Ultimately, the design of computer experiments shares many of the same features as the design of physical experiments. The experimenter is usually faced with having to propose 2 of the 3 values in Figure 1. In general, we recommend roughly following the proposed scheme, below, when designing a computer experiment.

4.1. Eliciting a prior

Choose a suitable correlation family and hyperparameters. This amounts to choosing the response surface model and related complexity. We recommend presenting the experimenter with several plots of realizations similar to Figure 4 as guidance, and recommend choosing a prior process that is slightly less smooth than the anticipated response as a matter of robustness.

4.2. Calculating the required sample size/expected accuracy

Obtain a lower bound for the required number of runs for a desired level of prediction accuracy (as in Subsection 3.1). Alternatively, on a fixed budget of runs, the expected prediction accuracy of the Gaussian process model (see Subsection 3.2) can be assessed. As another alternative, examine the complexity

level of the response that the budget and prediction accuracy allow, as in Section 3.3.

4.3. Making operational decisions

If the calculated required number of runs is operationally feasible, one can proceed that number of runs, and preferably more in view of (3.2). If for a feasible number of runs, the calculated RAUV appears to be greater than a tolerable level, a mitigating measure could be considered.

1. **Reducing dimensionality by eliminating inputs:** Obviously, every input coded into the simulator likely matters to some degree. Some inputs, however, may be thought to be less influential than others. If an expert can identify inputs whose absence from the model may have little bearing on the response, omitting those from the model (while holding them fixed at, say, their midpoint during computer runs) can significantly decrease the required number of runs.

2. **Altering the emulation model:** Excessive RAUV is an indication that lack of training data can lead to predicting a constant everywhere (except for spikes at the observed simulator outputs), in which case one might consider sacrificing interpolation and retreating to "traditional", parametric statistical models.

5. An Illustration

Consider the piston simulation appearing in Kenett and Zacks (1998). Here, a piston's linear motion is transformed into circular motion of a rod connected to a disk. The measured response is the time it takes to complete one cycle,

$$C\left(M,S,V_0,k,P_0,T_a,T_0\right)=2\pi\left(\frac{M}{k+S^2(P_0V_0/T_0)(T_a/V^2)}\right)^{1/2} \tag{5.1}$$

for $V=(S/2k)\{(A^2+4k(P_0V_0/T_0)T_a)^{1/2}-A\}$ and $A=P_0S+19.62M-(kV_0/S)$, where $M\in[30,60]$ is the piston weight (kg), $S\in[0.005,0.020]$ is the piston surface area (m^2), $V_0\in[0.002,0.010]$ is the initial gas volume (m^3), $k\in$ [1,000, 5,000] is the spring coefficient (N/m), $P_0\in[9\times10^4,11\times10^4]$ is the atmospheric pressure (N/m^2), $T_a\in[290,296]$ is the ambient temperature (K), and $T_0\in[340-360]$ is the filling gas temperature (K). More documentation (and code) for this model can be found at `http://www.sfu.ca/~ssurjano/emulat.html`.

Suppose that the underlying model is unknown, but that an expert believes that with all other factors being held fixed, letting M increase will increase the

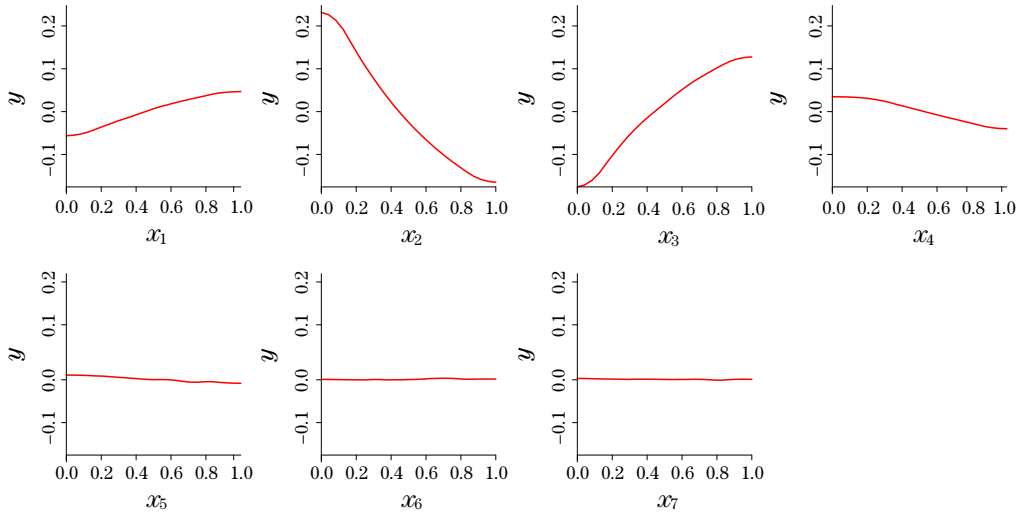

Figure 5. Main effect plots for the piston simulator, provided by the `tgp` **R** package (Gramacy and Taddy (2010)), where $x_1, \ldots, x_7$ represent the input variables M, S, V_0, k, P_0, T_a, T_0, respectively.

cycle time moderately in a nonlinear fashion and that the same is true for k, although a reverse trend is expected; likewise, that increasing S will result in a sharp, nonlinear decrease in cycle time, while the opposite will happen when V_0 alone varies; that within the experimental region the average effect of P_0 is very limited; and that varying T_a has an unnoticeable effect on the cycle time, the same being true for T_0.

In practice, we might use realizations drawn from univariate Gaussian processes, with different values of the correlation lengths, to help specify $\boldsymbol{\theta}$. In the absence of a domain expert our assessments would then be based on the main effect plots from the FANOVA of the cycle time $C(\boldsymbol{x})$, see Figure 5. In light of the sensitivity plots, a Gaussian process prior with a squared exponential correlation function (3.3) would be selected with hyperparameters $\boldsymbol{\theta} = (1, 0.4, 0.4, 1, 3, 10, 10)$, leading to a sample size of $n \geq 210$ when aiming at RAUV ≤ 0.05 in (3.2). Instead, ignoring T_0 and T_a and treating the model as 5-dimensional leads to a critical sample size of $n \geq 111$. We to compare the performance of the different sample sizes.

Since randomness in the response can only be incorporated via randomness in the design, we generated 50 random 5-dimensional Latin hypercube samples of various sample sizes. We first took a conservative approach to the sample size and chose $n = 120$ instead of $n = 111$ suggested by the inequality (3.2). In order

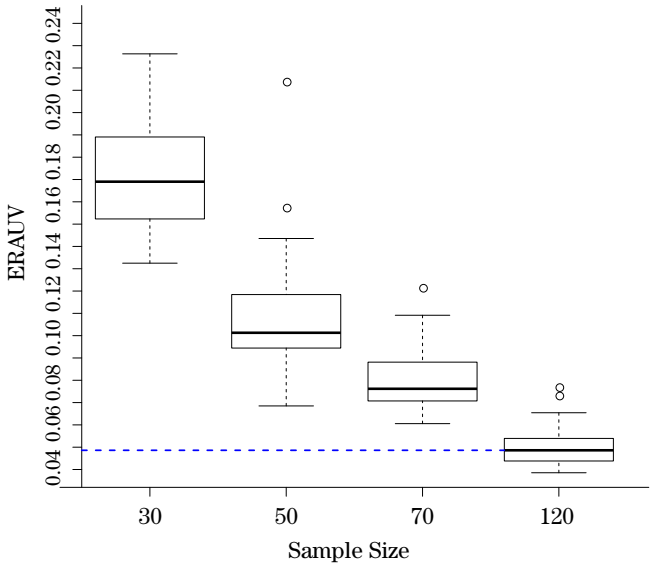

Figure 6. Empirical RAUV values over 50 repetitions for the piston example. Each repetition stands for a randomly chosen Latin hypercube (for each sample size).

to compare the $10d$ rule of thumb for $d = 5$ or $d = 7$, we also considered $n = 50$ and $n = 70$. The corresponding RAUV is shown in column 2 of Table 1.

To measure performance using simulated data, we evaluated the *Empirical RAUV*

$$\text{ERAUV}\left(\widehat{y};\mathcal{D}\right) = \left(\frac{\sum_{i=1}^{n_{\text{ho}}}\left(\widehat{y}_i^{\text{ho}} - y_i^{\text{ho}}\right)^2}{n_{\text{ho}}\,\widehat{\sigma}^2}\right)^{1/2}$$

at a size $n_{\text{ho}} = 100{,}000$ holdout set in the 7-dimensional space, where the estimate $\widehat{\sigma}^2 = 0.022$ was obtained by fitting a one time 5-dimensional model to a size 1,000 dataset that remained fixed throughout the simulation study.

Results of the simulation study appear in Table 1 and Figure 6. The variability in the results shows how the randomness of the choice of the design (Latin hypercube in this case) propagates into the model. With an additional design optimality criterion (maximin distance for example) one should expect to see clear separation between the different sample sizes.

The maximum likelihood estimates, $\widehat{\boldsymbol{\theta}} = (1.60, 0.30, 0.50, 0.44, 1.99)$, turned out to be quite different from our early assessment. In spite of this, our procedure seems to have captured the overall complexity of the model to a good degree, judging by the proximity of the theoretical RAUV values (for the specified parameter values) to the empirical ones in Table 1. Section 6 discusses a

Table 1. Summary of simulation results for the piston cycle time model. Average ERAUV results appear with ± one standard deviation.

Sample Size	Theoretical RAUV	ERAUV
30	≥ 0.191	$0.172\,(\pm 0.025)$
50	≥ 0.120	$0.107\,(\pm 0.024)$
70	≥ 0.086	$0.080\,(\pm 0.013)$
120	≥ 0.045	$0.050\,(\pm 0.008)$

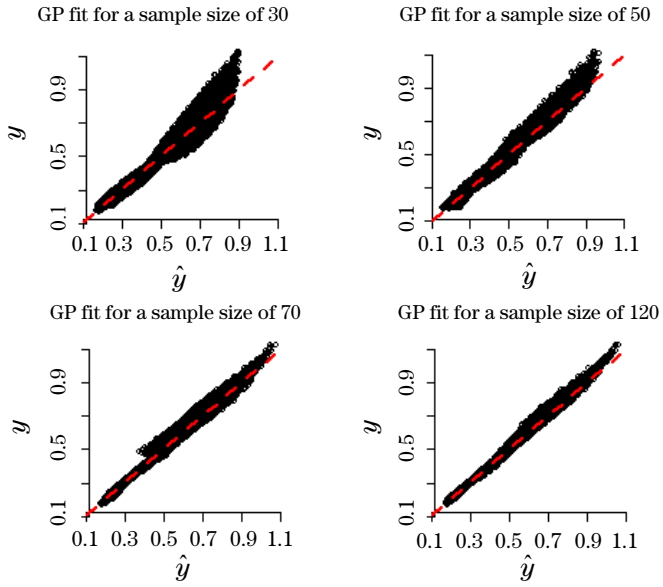

Figure 7. True response vs. fitted plots for the piston example, based on sample sizes of 30, 50, 70 and 120, respectively, and randomly constructed Latin hypercube designs.

robust procedure that allows the experimenter to provide a range of values for each correlation length parameter, rather than giving a single guess.

Finally, Figure 7 provides a visualization of the improvement of the RAUV (for randomly chosen designs) from 0.167 to 0.106, 0.079 and finally 0.048 as the sample size increases from 30 to 50, 70 and 120, respectively (for a single fit, each).

6. Robust Sample Size Calculations

The task of choosing values for the correlation parameters is challenging, and specifying a range of values may be easier in practice. It is therefore natural to consider incorporating some uncertainty with respect to these chosen values to

Table 2. Robust sample size calculations for the piston simulation, based on different levels of uncertainty.

Very high uncertainty		High uncertainty		Moderate uncertainty	
Range	n (95%)	Range	n (95%)	Range	n (95%)
$\theta_1 : [0.001, 5]$		$\theta_1 : [0.5, 1.5]$		$\theta_1 : [0.8, 1.2]$	
$\theta_2 : [0.001, 5]$		$\theta_2 : [0.1, 0.7]$		$\theta_2 : [0.25, 0.55]$	
$\theta_3 : [0.001, 5]$	≥ 157	$\theta_3 : [0.1, 0.7]$	≥ 214	$\theta_3 : [0.25, 0.55]$	≥ 140
$\theta_4 : [0.001, 5]$		$\theta_4 : [0.5, 1.5]$		$\theta_4 : [0.8, 1.2]$	
$\theta_5 : [0.001, 5]$		$\theta_5 : [1, 5]$		$\theta_5 : [2, 4]$	

enhance robustness.

Denote by $g : \mathbb{R}^d \to \mathbb{N}^+$ the function that maps each vector of correlation parameters $\boldsymbol{\theta}$ to a critical sample size n^c through (3.2). If we now assign a distribution $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, g will induce a probability measure on n^c. Drawing a random sample $\{\boldsymbol{\theta}_i\}$ from $\pi(\boldsymbol{\theta})$ will then result in a Monte Carlo sample $\{g(\boldsymbol{\theta}_i)\}$ from a distribution $\pi(n^c)$ of sample sizes.

As an example, consider the piston simulation of Section 5 and assign $\theta_1, \ldots, \theta_5$ independent uniform priors on $[0.8, 1.2]$, $[0.25, 0.55]$, $[0.25, 0.55]$, $[0.8, 1.2]$ and $[2, 4]$, respectively (see the "Moderate uncertainty" scenario in Table 2). We drew a random sample of size 10,000 from $\pi(\boldsymbol{\theta})$ and produced the corresponding random sample from $\pi(n^c)$ through solving for eigenvalues and calculating the critical sample size (3.2) for each drawn vector. Figure 8 shows the histogram for the sample sizes. Given the uncertainty in the response surface specification, a choice must be made in order to run the experiment. One could, of course, choose the maximum sample size from those observed. Doing so would be extreme in our view (and also would require more random samples to appropriately estimate the maximum sample size). Looking at the plot, a line at $n = 140$, marking the 95th percentile of the sample sizes, is added. We view this as representing a safe choice for a sample size that accounts for uncertainty in $\boldsymbol{\theta}$. Although the eventual recommended sample size is somewhat larger than that from Section 5, it is still rather economical compared to the worst case scenario sample of 187.

The choice of prior distribution for the correlation parameters impacts the sample size. One might be tempted to think that more uncertainty in the complexity requires more samples, but this is not necessarily the case. To study how sample size calculations are impacted by the choice of $\pi(\boldsymbol{\theta})$, we assigned $\theta_1, \ldots, \theta_5$ independent uniform priors on respective intervals about their conjectured values. Table 2 summarizes the results of a small scale simulation. We considered three scenarios: "Very high uncertainty", "High uncertainty" and "Moderate

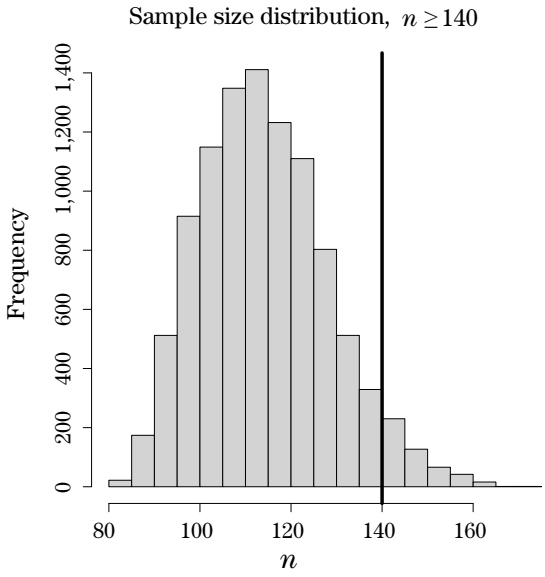

Figure 8. A histogram of a Monte Carlo sample from the sample size distribution $\pi(n^c)$, induced by assigning a distribution $\pi(\boldsymbol{\theta})$. The cutoff at $n \geq 140$ marks the 95th percentile.

uncertainty", pertaining to long, fairly long, and medium length intervals, respectively. For each scenario we drew a random sample of size 10,000 from $\pi(\boldsymbol{\theta})$ and produced the corresponding random sample from $\pi(n^c)$ through solving for eigenvalues and calculating the critical sample size (3.2) for each sampled vector.

Looking at Table 2, increased uncertainty does not automatically translate into increased sample size. The wide intervals in the leftmost column resulted in some inactive dimensions for many of the randomly drawn vectors, and in turn to a smaller sample size than the one recommended for the more focused "High uncertainty" scenario.

7. Discussion

The machine learning community refers to the IMSPE curve versus n as the "learning curve" (see e.g. Williams and Vivarelli (2000)) and, over the years, tighter lower bounds than (3.1) for the case of noisy data – as well as upper bounds – have been derived (see e.g. Sollich (1999)). However, in the interpolation setting we are considering, these bounds do not apply. Thus the bound in (3.1) is used in this paper. It has been found to be fast to calculate and, as Figure 2 implies, fairly tight in practice.

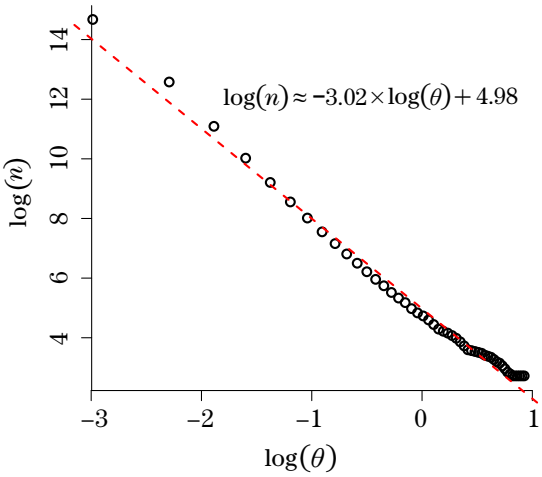

Figure 9. The minimum sample size (3.2) required to achieve RAUV ≤ 0.05 vs. correlation length for a 4 dimensional isotropic Gaussian process based on the product Matérn correlation function with smoothness $\nu = 2.5$.

As a possible topic for future research, one could look to establish an explicit expression governing the trade-off between the correlation length $\boldsymbol{\theta}$ and the required sample size for a given ε. We performed a simulation study for the isotropic 4-dimensional product Matérn kernel with smoothness parameter $\nu = 2.5$ by calculating n^c for $\theta = 0.05, 0.1, 0.15, \ldots, 2.5$ and $\varepsilon = 0.05$. Figure 9 shows the results in the log-log scale, along with the fitted linear regression line. The estimated slope is very close to -3, suggesting the critical sample size decays at a rate of $\mathcal{O}\left(\theta^{-3}\right)$ for this example.

We are in agreement with Loeppky, Sacks and Welch (2009) in noting that in high dimensions, the number of samples may be onerous. As an illustration, we considered a setting with $d = 26$ inputs and varied the number of active factors from 1 to 26. For each number of active inputs, and sample sizes of $n =$ 50, 100, 200, 400, and 500, we generated 50 realizations of a Gaussian process with the product power-exponential covariance and $\theta = 1$ for the active factors, using a randomly drawn Latin hypercube design along with a hold-out set of 100 randomly chosen trials. The Kriging model, with the product power-exponential covariance, was fit to each simulated dataset and the predictive performance was evaluated on the corresponding hold-out set using the empirical value of

$$\int_{[0,1]^d} \frac{\mathbb{V}\mathrm{ar}\left\{y\left(\boldsymbol{x}\right)\middle|\mathcal{D}\right\}}{\mathbb{V}\mathrm{ar}\left\{y\left(\boldsymbol{x}\right)\right\}}\,\mathrm{d}\boldsymbol{x}.$$

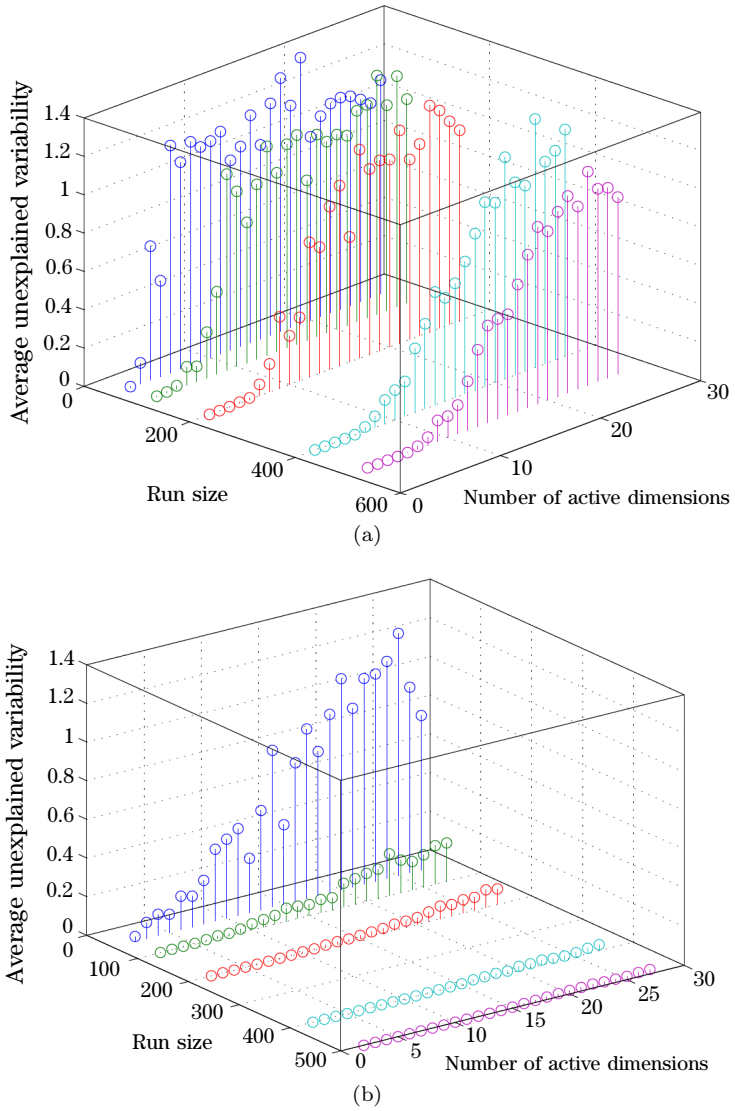

Figure 10. Empirical average unexplained variability values for data generated by multiplicative (10a) and additive (10b) univariate Gaussian processes vs. the number of active inputs and sample size.

There, for a fixed sample size, the average unexplained variability grows fairly rapidly as the number of active factors increases. Indeed, even when $n = 500$, we see that, when the number of active factors is about 15, the Gaussian process emulator has a difficult time predicting the response surface accurately. Overall, the more complex the response surface and the more active inputs influencing

the response, the larger the sample size required.

We repeated the same procedure, but generated data for the same scenarios using a sum of 1-d independent Gaussian processes (one for each active dimension) with $\theta = 1$ for the active factors. However, the data analysis was performed using the same Kriging model as before with the product squared-exponential covariance. The results are summarized in Figure 10b. There, for the same sample sizes and numbers of active dimensions, the standard GP explains almost all of the variability on average, except for the relatively small setting of $n = 50$. The take-away message here is that, when the model is simple, the Gaussian process does an admirable job at computer model emulation.

Our belief is that the complexity of many computer models lies somewhere between these two extremes, and thus the methodology proposed in this paper is a conservative approach. The illustrations point to a need for a class of simpler random functions that represent the space in which the computer model response surface lie for both design and analysis.

Supplementary Materials

A web application that performs all the analyzes presented in this paper is available at `https://harario.shinyapps.io/Sample_Size_Shiny`. In addition, **R** code for the study of Section 5 is available online as supplementary material.

Acknowledgment

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute. The authors would like to thank the referees for their insightful comments.

References

Box, G. and Meyer, R. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–18.

Chapman, W. L., Welch, W. J., Bowman, K. P., Sacks, J. and Walsh, J. E. (1994). Arctic sea ice variability: model sensitivities and a multidecadal simulation. *Journal of Geophysical Research: Oceans* **99**, 919–935.

Chen, H., Loeppky, J. L., Sacks, J. and Welch, W. J. (2016). Analysis methods for computer experiments: how to assess and what counts? *Statistical Science*, to appear.

Gramacy, R. B. and Taddy, M. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed Gaussian process

models. *Journal of Statistical Software* **33**, 1–48.

Harari, O. and Steinberg, D. (2014). Optimal designs for Gaussian process models via spectral decomposition. *Journal of Statistical Planning and Inference* **154**, 87–101.

Johnson, M., Moore, L. and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**, 131–148.

Jones, D. R., Schonlau, M. and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13**, 455–492.

Joseph, V., Dasgupta, T. and Wu, C. (2012). Minimum energy designs: from nanostructure synthesis to sequential optimization. *under revision*.

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K. and Frieman, J. A. (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics* **5**, 2470–2492.

Kenett, R. S. and Zacks, S. (1998). *Modern Industrial Statistics: Design and Control of Quality and Reliability*. Pacific Grove: Duxbury Press.

Loeppky, J., Sacks, J. and Welch, W. (2009). Choosing the sample size of a computer experiment: a practical guide. *Technometrics* **51**, 366–376.

Micchelli, C. and Wahba, G. (1981). Design problems for optimal surface interpolation. In *Approximation Theory and Applications*, Z. Ziegler, ed. New York: Academic Press, 329–347.

Ranjan, P., Haynes, R. and Karsten, R. (2011). A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics* **53**, 366–378.

Sacks, J., Schiller, S. and Welch, W. (1989). Designs for computer experiments. *Technometrics* **31**, 41–47.

Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science* **4**, 409–423.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Trantola, S. (2008). *Global Sensitivity Analysis: the Primer*. John Wiley.

Shewry, M. and Wynn, H. (1987). Maximum entropy sampling. *Journal of Applied Statistics* **14**, 165–170.

Sollich, P. (1999). Learning curves for Gaussian processes. In *Advances in Neural Information Processing Systems 11*, M. Kearns, S. Solla and D. Cohn, eds. MIT Press, 344–350.

Williams, C. K. and Vivarelli, F. (2000). Upper and lower bounds on the learning curve for Gaussian processes. *Machine Learning* **40**, 77–102.

Yi, T.-m., Fazel, M., Liu, X., Otitoju, T., Goncalves, J., Papachristodoulou, A., Prajna, S. and Doyle, J. (2005). Application of robust model validation using sostools to the study of g-protein signaling in yeast. In *Proceedings of the First Conference on Foundations of Systems Biology in Engineering*. Santa Barbara, CA.

MTEK Sciences Inc., 777 West Broadway Vancouver, BC, Canada V5Z 1J5.

E-mail: oharari@mteksciences.com

(Work was done while Ofir was a postdoctoral fellow at Simon Fraser University)

Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive Burnaby, BC, Canada V5A 1S6.

E-mail: dbingham@stat.sfu.ca

Department of Statistics, The Ohio State University 1958 Neil Avenue, Columbus, OH 43210-1247, USA.

E-mail: amd@stat.ohio-state.edu

Social Decision Analytics Laboratory Biocomplexity Institute, Virginia Tech Virginia Tech, National Capital Region Arlington VA 22203 USA.

E-mail: dhigson@vbi.vt.edu

(Received April 2016; accepted September 2016)