Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: www.tandfonline.com/journals/lqen20

Comparing Computer Experiments for the Gaussian Process Model Using Integrated Prediction Variance

Rachel T. Silvestrini, Douglas C. Montgomery & Bradley Jones

To cite this article: Rachel T. Silvestrini, Douglas C. Montgomery & Bradley Jones (2013) Comparing Computer Experiments for the Gaussian Process Model Using Integrated Prediction Variance, Quality Engineering, 25:2, 164-174, DOI: 10.1080/08982112.2012.758284

To link to this article: https://doi.org/10.1080/08982112.2012.758284

Quality Engineering, 25:164–174, 2013

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2012.758284

Taylor & Francis
Taylor & Francis Group

Comparing Computer Experiments for the Gaussian Process Model Using Integrated Prediction Variance

Rachel T. Silvestrini[1], Douglas C. Montgomery[2], Bradley Jones[3]

[1]Naval Postgraduate School, Monterey, California
[2]Arizona State University, Tempe, Arizona
[3]SAS Institute, Cary, North Carolina

ABSTRACT Space-filling designs are a common choice of experimental design strategy for computer experiments. This article compares space-filling design types based on their theoretical prediction variance properties with respect to the Gaussian process model. An analytical solution for calculating the integrated prediction variance (*IV*) of the Gaussian process model is given. Using the analytical calculation of *IV* as a response variable, this article presents a study of the effects of dimension; sample size; value of parameter vector, $\boldsymbol{\theta}$; and experimental design type using a factorial design and regression analysis.

KEYWORDS computer simulation, gaussian process models, integrated variance, space-filling designs

INTRODUCTION

Computer simulation experiments are a potentially beneficial alternative to physical experimentation for early stage product design and process development activities and the study of transactional systems where full-scale experiments are impractical. Unlike physical experiments, which have been developed and studied for close to 80 years, computer experiments are a relatively new application area.

Currently there are few references comparing experimental designs for computer models. Allen et al. (2003) compared combinations of experimental design classes with respect to second-order response surfaces and kriging models. They pointed out that the utility of a given modeling method was highly dependent on the choice of the experimental design. Hussain et al. (2002) presented seven two-dimensional test functions that they used to compare two surrogate models and two design types. They concluded that the factorial design had better performance with respect to the polynomial model and the Latin hypercube design (LHD) had better performance with respect to the radial basis functions. Bursztyn and Steinberg (2006) developed a new method of design comparison based on a Bayesian interpretation of an alias matrix. They compared Latin hypercube designs, uniform designs, lattice designs, rotation designs, and fractional factorial

Address correspondence to Rachel T. Silvestrini, Naval Postgraduate School, 1411 Cunningham Rd., Monterey, CA 93943. E-mail: rtsilves@nps.edu

designs. They found that the alias sum of squares criterion tended to favor the rotation designs. Fractional factorial designs performed best in terms of the entropy and minimum distance criteria, whereas the integrated mean squared error (IMSE) criterion favored space-filling designs. R. T. Johnson et al. (2010) compared the prediction variance of designs for fitting high-order polynomial models. Their work demonstrated that the space-filling designs perform poorly compared to optimal designs. Of the space-filling designs, the sphere-packing designs generally exhibited the best performance in terms of prediction variance with respect to polynomial models.

In this article we compare the maximin Latin hypercube design (Mm LHD), sphere packing (SP) design, uniform (U) design, maximum entropy (ME) design, and the Gaussian process integrated mean square error (GP IMSE or *I*-optimal) design with respect to the prediction variance of the Gaussian process (GP) model. We provide comparative plots of predictive variance with respect to the GP model as well as comparisons of the integrated variance.

We start with a brief description of the GP model used in this analysis and a motivating example. This is followed by a section that describes classes of experimental designs, and then we introduce our comparison metrics and the graphical comparison technique, followed by the results of our study. Following the results we present a sensitivity analysis for the GP IMSE design with respect to the parameters of the assumed model. Then we revisit the motivating example and finally present our conclusions.

GAUSSIAN PROCESS MODEL

Computer simulation outputs can result in nonlinear response surfaces that traditional regression methods cannot adequately model. The GP model fitting technique can capture such complex behavior. As a result, GP models are a standard for fitting deterministic computer output. See Jones and Johnson (2009), Bayarri et al. (2007), Linkletter et al. (2006), Fang et al. (2006), Santner et al. (2003), Welch et al. (1992), Currin et al. (1991), Sacks, Schiller, and Welch (1989), and Sacks, Welch, et al. (1989) for examples of the GP model and its application to deterministic computer simulation outputs.

The GP model fits a response, $y(\mathbf{x})$, using a stochastic process; specifically, the multivariate normal distribution. The GP model is an attractive model to use with a deterministic response because it acts as an exact interpolator, but it can also be used with a stochastic response. The output response is represented as an $n \times 1$ data vector $\mathbf{y}(x)$, where $\mathbf{y}(x) \sim \mathrm{N}(\mu\mathbf{1}_n, \sigma^2\mathbf{R}(\mathbf{X}, \boldsymbol{\theta}))$. $\mathbf{R}(\mathbf{X}, \boldsymbol{\theta})$ is an $n \times n$ correlation matrix that can be represented by one of a variety of forms (see Sacks, Welch, et al. 1989). We use the Gaussian correlation function below:

$$R_{ij}(\mathbf{X}, \boldsymbol{\theta}) = \exp\left(-\sum_{k=1}^{d} \theta_k (x_{ik} - x_{jk})^2 \right)$$

where $\theta_k \geq 0$ and d is the number of factors in the experiment. If $\theta_k = 0$, then the correlation is 1.0 across the range of the kth factor and the fitted surface will be flat in that direction. Large θ_k corresponds to low correlation in the kth factor and the fitted surface will exhibit strong curvature in the direction of the kth variable.

The fitted GP prediction equation is

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{r}'(\mathbf{x}, \hat{\boldsymbol{\theta}})\mathbf{R}^{-1}(\mathbf{X}, \hat{\boldsymbol{\theta}})(\mathbf{y} - \hat{\mu}\mathbf{1}_n)$$

where the fitted mean and the θ_js are represented by $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$. These parameters are usually estimated via maximum likelihood. In the fitted equation, $\mathbf{r}'(\mathbf{x}, \hat{\boldsymbol{\theta}})$ is an $n \times 1$ vector of estimated correlations of the unobserved $y(\boldsymbol{x})$ at a new value of the explanatory variables with the observations in the data, $y(\boldsymbol{x})$:

$$r_i(\mathbf{x}, \boldsymbol{\theta}) = \exp\left\{ -\sum_{k=1}^{d} \theta_k (x_k - x_{ik})^2 \right\},$$

$\hat{y}(\mathbf{x})$ interpolates the data. Using this model, the relative prediction variance is

$$\frac{Var[\hat{y}(\mathbf{x})]}{\sigma^2} = 1 - \mathbf{r}'(\mathbf{x}, \boldsymbol{\theta})\mathbf{R}^{-1}(\mathbf{X}, \boldsymbol{\theta})\mathbf{r}(\mathbf{x}, \boldsymbol{\theta}). \qquad [1]$$

The variance of the predicted response at a new point depends on the design, $\mathbf{X}$, and the unknown parameter vector, $\boldsymbol{\theta}$. It also depends implicitly on the sample size (number of rows in $\mathbf{X}$) and the number of design or experimental factors in the simulation model d.

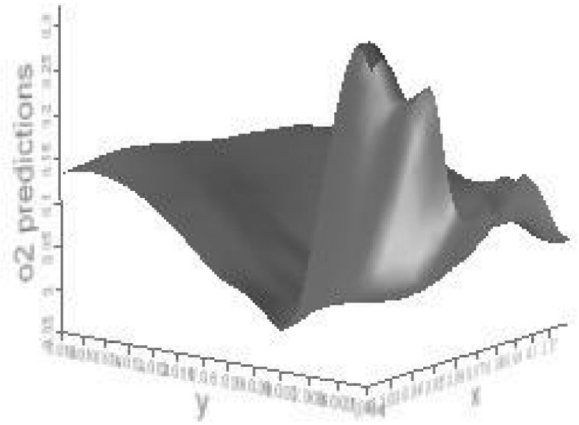

FIGURE 1 **CFD response surface. (Color figure available online.)**

As an example of a computer experiment, consider the output from a computational fluid dynamics (CFD) model of combustion in two input variables, the *x*-axis and *y*-axis. This model is used by the National Aeronautics and Space Administration (NASA). A graph of the response surface of residual oxygen as a function of *x*-axis and *y*-axis location is in Figure 1.

It is clear that simple models such as first- or second-order polynomials would not provide an adequate fit to this surface. The computer model in this study ran quickly enough to allow for the simulation of approximately 5,000 design points. The development of the next CFD model for the larger scale combustion experiment is expected to result in a much more complex set of differential equations, and it may take hours to simulate a single design point. Past applications using CFD have shown that GP models perform well as surrogates for the complex simulation code. Assuming that we would fit the data using a GP model, we were interested in finding out what design is best and how many design points are required to get acceptable prediction performance. This example serves as motivation to determine which design strategies have the best predictive capabilities when the expected form of the surrogate model is a GP model. We assume that the true surface is either a GP model or is closely approximated by this model. Though it is unlikely that many deterministic simulation models produce an output that is actually a GP model (indeed, there are potentially a very large number of nonlinear functions that may describe the output), experience has shown that the GP model is an excellent approximation in many situations. Therefore, comparing different design strategies under this assumption can provide useful insight to experimenters.

SPACE-FILLING DESIGNS

Inherent differences between computer and physical experiments have led to the development of space-filling experimental designs for use solely in deterministic computer simulations. Many space-filling design alternatives have been proposed. In our study we use the sphere-packing design, the Latin hypercube design, the uniform design, the maximum entropy design, and the GP IMSE design. We chose these designs because of their popularity in the literature and the ability to create them using commercially available software. Table 1 provides information about the paper(s) where the designs are introduced, the goal or criterion of the design, and paper(s) containing examples and applications of the designs. Two dimensional plots of these five space-filling designs can be found in Jones and Johnson (2009) and R. T. Johnson et al. (2010). Note that the LHD used in this article is the maximin LHD, meaning that it is a LHD with an added criterion that the optimization used to create the design maximizes the minimum distance between points with in the constraints of the Latin hypercube. Also note that another name for the sphere packing design is a maximin design.

COMPARISON TECHNIQUE

Our purpose is to evaluate the prediction performance of design strategies with respect to the GP model. We use the integrated prediction variance as the basis of comparison. Santner et al. (2003) provided a general expression for the integrated prediction variance, which for our situation reduces to

$$IV = tr(\mathbf{R}^{-1}\mathbf{M}) \qquad [2]$$

where $\mathbf{M}$ and $\mathbf{R}$ are $n \times n$ matrices. Now the elements of $\mathbf{M}$ are

$$m_{ij} = \int r(\mathbf{x}_i, \mathbf{X}) r(\mathbf{x}_j, \mathbf{X}) d\mathbf{x},$$

TABLE 1 Description of Space-Filling Designs Used in this Article

Design	Developed by:	What the design does	Applications/examples
Sphere packing	M. E. Johnson et al. (1990)	Maximizes the minimum distance between pairs of design points	Jank and Shmueli (2007), Liefvendahl and Stocki (2006), Chen et al. (2006), Roux et al. (2006), Bursztyn and Steinberg (2006)
Latin hypercube	McKay et al. (1979)	A permutation of points in each column	Welch et al. (1992), Mease and Bingham (2006), Tyre et al. (2007), Storlie and Helton (2007)
Uniform	Fang (1980)	A set of design points uniformly scattered in the design space	Wang and Fang (1981), Fang et al. (2006), Bursztyn and Steinberg (2006)
Maximum entropy	Shewry and Wynn (1987)	Maximizes the amount of information contained in the distribution of a data set	Ko et al. (1995)
GP IMSE	Sacks, Welch, et al. (1989)	Minimizes the integrated mean squared error of the GP model	Sacks, Schiller, and Welch (1989)

which reduces to

$$m_{ii} = c\prod_{k=1}^{d}\left[\Phi(2\sqrt{\theta_k}(1-X_{ik})) - \Phi(2\sqrt{\theta_k}(-1-X_{ik}))\right]$$

and

$$m_{ij} = c\prod_{k=1}^{d}\exp\left(\mu - \frac{\theta_k}{2}(X_{ik} - X_{jk})^2\left[\Phi(2\sqrt{\theta_k}(1-x^*)) - \Phi(2\sqrt{\theta_k}(-1-x^*))\right]\right.$$

where

$$c = \sqrt{\exp(d\ln(\pi/2)/\prod_{k=1}^{d}\theta_k}$$

and

$$x^* = \frac{X_{ik}+X_{jk}}{2}$$

A computer program was written to evaluate Eq. [2]. The results are provided in the following section.

DESIGN COMPARISON STUDY

In this section we compare the performance of five space-filling designs and other factors with respect to integrated prediction variance of the GP model. The five space-filling designs used in the comparison are the Mm LHD, SP, ME, U, and GP IMSE (*I*-optimal). The integrated prediction variance (*IV*) of computer experiments involving two, three, four, and five factors was studied for each of the design types across a range of sample sizes and values in the $\boldsymbol{\theta}$ vector. Recall that for a GP model, the length of the $\boldsymbol{\theta}$ vector is equal to the number of factors in the model, thus equal to the dimension of the design. For each case, two to five design factors, we set the values of the sample size and θs using a factorial design. We used a separate factorialdesign for each of the two to five design factor scenarios. Each design has the following factors:

- n: sample size (three levels: $(5d)$, $10d$, $(15d)$, where $d = 2, 3, 4, 5$ and d =the number of factors in the computer model being studied)
- τ: sum of $\boldsymbol{\theta}$ vector elements (three levels: 5, 10, 15)
- R: ln ratio of values in $\boldsymbol{\theta}$ vector (two levels: 0, 1; Note: there will be $d-1$ ratios)
- D: design type (five levels: Mm LHD, SP, ME, U, and GP IMSE)

The values for the elements of $\boldsymbol{\theta}$ vector are determined by τ and R. The range of the ratio of θs in the $\boldsymbol{\theta}$ vector is 1 and 10; thus, the log of the ratios is 0 and 1. Each row in the resulting design matrix requires the creation of a design with specified sample size, $\boldsymbol{\theta}$ vector, and design type. Note that the ME design and the GP IMSE design require the specification of the unknown θ_j parameters. In an actual experiment, the θ_j are unknown in advance. Therefore, we create ME and GP IMSE designs

by using equal scaling of the theta constants, where the thetas are set equal to the mean value of the $\boldsymbol{\theta}$ vector.

Our response variable, y, is the *IV* that is calculated numerically from Eq. [2]. The design matrix and results were analyzed by stepwise regression using a full quadratic model as the initial model to determine what effects have significant impact on *IV*. The summary results for the two to five factor cases are presented in Table 2. In this table we have shown only the three most important factors along with some regression model summary statistics.

Table 2 shows that two of the main effects (n = sample size and τ = sum of the θ vector elements) are important in all cases. In addition, three of the two-factor interactions ($n \times \tau$, $n \times D$, and $\tau \times D$) were significant in each model, although their effect magnitudes were much smaller than the main effects. The R^2 and adjusted R^2 values (not reported) were also all greater than 0.95.

Though three two-factor interactions were statistically significant in all models, in only one case—that with two design factors—did any of these interactions have a relatively important effect. Figure 2 presents the $n \times \tau$ (interaction plot for the two-factor case. Note that when the sample size was at the low level ($5d = 10$) the effect of τ was relatively large, so that as the magnitude of the θs increased there was a large increase in the theoretical integrated prediction variance. However, when sample size was at the high level ($15d = 30$), τ had very little effect on the theoretical integrated prediction variance.

Figure 3 presents the $n \times D$ interaction plot for the two-variable case. This plot shows that if n is at the high level the design type has very little effect on the theoretical integrated prediction variance. When the sample size is at the low level there is some difference between designs with the *I*-optimal, Latin hypercube, and uniform designs slightly outperforming the maximum entropy and sphere-packing designs.

TABLE 2 **Summary Results for Significant Effects in Two- to Five-Factor Cases**

Two factors	Three factors	Four factors	Five factors
n	n	n	n
t	t	t	t
n^2	n^2	D	D
RMSE = 0.029	RMSE = 0.039	RMSE = 0.05	RMSE = 0.047
Mean = 0.186	Mean = 0.343	Mean = 0.480	Mean = 0.563

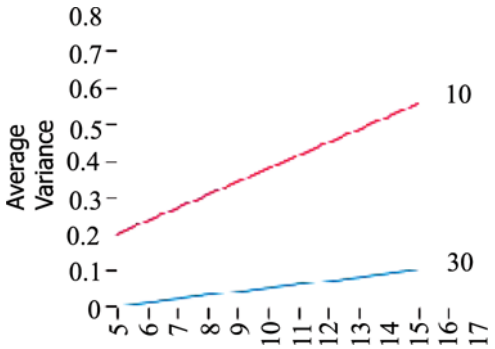

FIGURE 2 $n \times \tau$ **interaction plot. *Y*-axis is *IV*, *x*-axis is τ, and the two lines represent n at the low and high levels (10 and 30, respectively). (Color figure available online.)**

Figures 4–7 illustrate the main effects profilers for each of the regression models on the experiments for the two-, three-, four-, and five-factor cases, respectively.

The profilers provide the average *IV* (response variable) values for chosen values of the main effects. Figures 4–7 are set on the middle values for each of the main effects *N*, *R*, and τ. These figures show that the GP IMSE (listed in the figures as *I*-optimal), Mm LHD, and U designs all had similar *IV* results and clearly outperformed the ME and SP designs. The magnitude of the performance of the designs as the number of factors increased is also apparent. Notice that the *IV* for the *I*-optimal (GP IMSE) design was approximately 0.098, 0.235, 0.423, and 0.477 for the two-, three-, four-, and five-factor cases, respectively. For a fixed set of values and ratios in the $\boldsymbol{\theta}$ vector and for a $10d$ sample size, the *IV* increased nonlinearly.

Based on the response surface model for the *IV* we observed the following:

1. For a fixed sample size (N), increasing the value of any element of $\boldsymbol{\theta}$ increased the integrated variance *IV* for all design types.
2. Generally, increasing the number of runs reduced the integrated variance for all design types given a fixed $\boldsymbol{\theta}$ vector. However, in most cases there was minimal benefit beyond $N = 10d$.

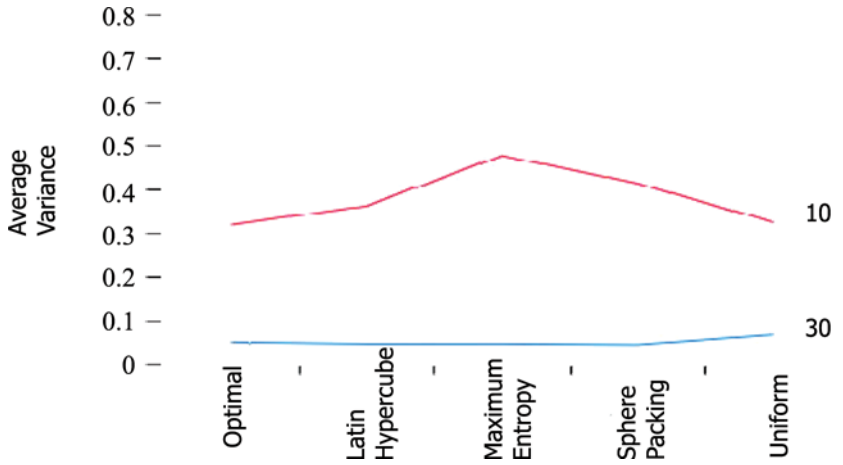

FIGURE 3 *n* × *D* **interaction plot. *Y*-axis is *IV*, *x*-axis is design type (*D*), and the two lines represent *n* at the low and high levels (10 and 30, respectively). (Color figure available online.)**

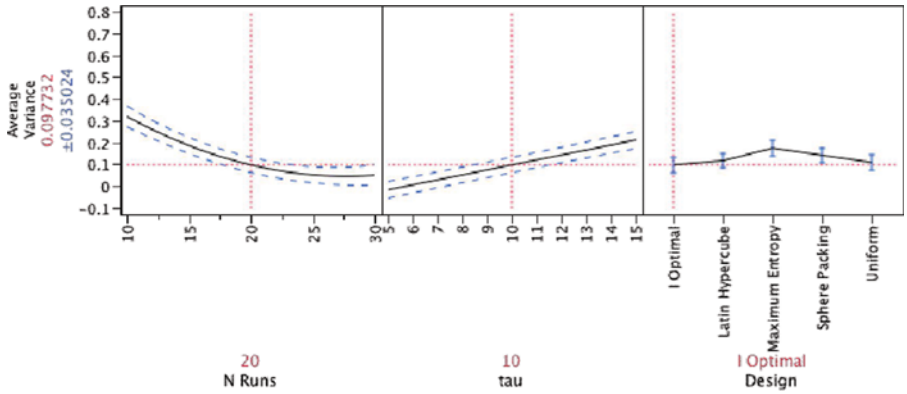

FIGURE 4 **Profile plot for the two-factor case. (Color figure available online.)**

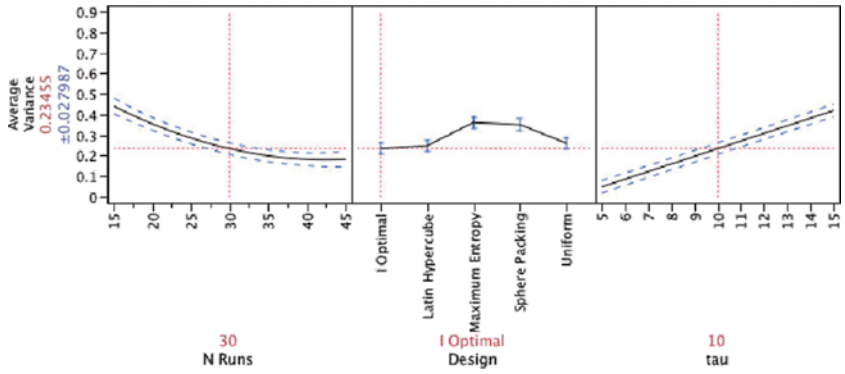

FIGURE 5 **Profile plot for the three-factor case. (Color figure available online.)**

3. Overall there was not much difference in performance between the GP IMSE, LHD, and U design, especially in cases where the sample size was at the high level (15*d*).
4. The ME design performance greatly improved as the sample size increased.

Sensitivity of the GP IMSE Design to the Elements of θ

In the previous section we demonstrated that the GP IMSE design performed competitively with respect to the *IV* criterion. Of course, this is not

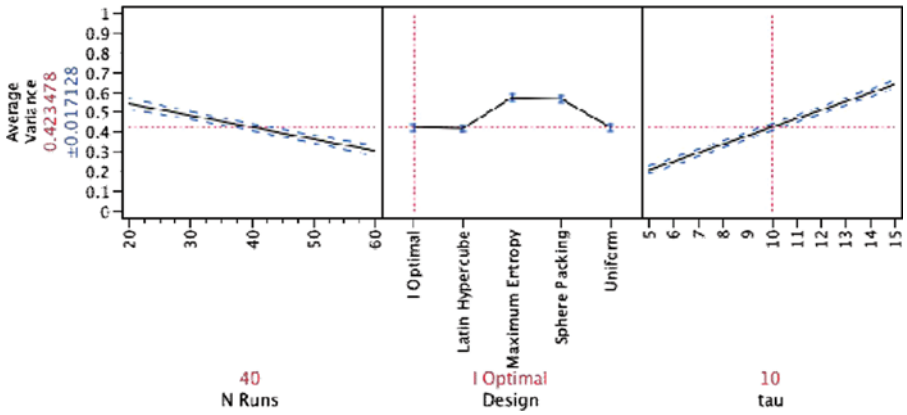

FIGURE 6 **Profile plot for the four-factor case. (Color figure available online.)**

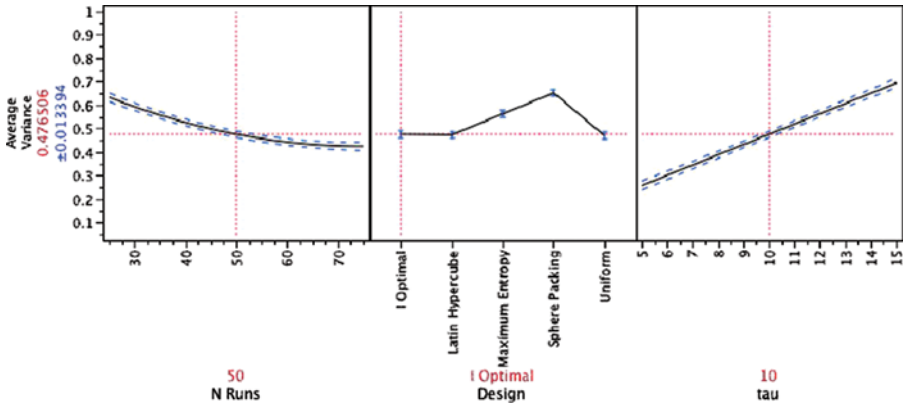

FIGURE 7 **Profile plot for the five-factor case. (Color figure available online.)**

surprising because the GP IMSE design criterion minimizes the integrated mean square error of a design for a specified parameter vector, $\boldsymbol{\theta}$, which is a necessary input to the algorithm that generates the design. In the experimental study described in the previous section the parameters in the (vector were always specified correctly for the GP IMSE design. In general this will not be the case in a real application. Therefore, it is of interest to explore the effect of misspecification of the unknown parameters on the prediction properties of the design.

We attacked this problem with two different studies because the effect of incorrect specification is different depending of whether all of the elements of the $\boldsymbol{\theta}$ vector are the same or not. If all of the elements of the $\boldsymbol{\theta}$ vector are the same, then the GP IMSE designs look qualitatively similar. Figure 8 shows side-by-side views of three designs.

The left panel of Figure 8 shows the GP IMSE design for two factors generated under the assumption that both values of the $\boldsymbol{\theta}$ vector are equal to one. For the middle panel both values are 5 and for the right panel both values are 10. We see few qualitative differences across the three panels; however, we note that as $\boldsymbol{\theta}$ increased, the design points tended to pull in toward the middle design point value, which in this case was zero. In addition, there was a noticeable gap in the center of the region for the design where both values of the $\boldsymbol{\theta}$ vector are equal to one.

Table 3 shows the resulting *IV* for the true value of (given in the first column for the design generated under the assumption that $\boldsymbol{\theta}$ is as given in the second column. When the true value of $\boldsymbol{\theta}$ equals the specified value (i.e., for rows 1, 5, and 9) the *IV* is the smallest for the given true value. The effect of the changing τ on the *IV* is overwhelmingly more substantial than the effect of incorrect specification of $\boldsymbol{\theta}$ in generating the design.

We now consider the effect of incorrect specification of the $\boldsymbol{\theta}$ vector on two-factor GP IMSE designs where the

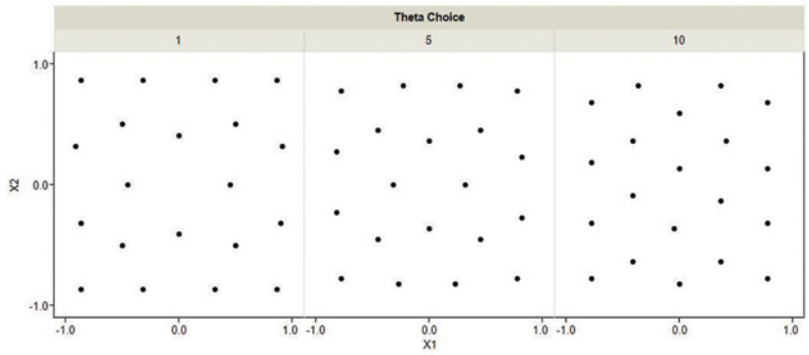

FIGURE 8 Three different GP IMSE designs for two factors varying in the assumed equal values of the θ vector. (Color figure available online.)

TABLE 3 Effects of Incorrect Specification Assuming the Values of the θ Vector are Equal

True value of θ	Specified for design	Integrated variance
1	1	0.000743
1	5	0.00107
1	10	0.0015
5	1	0.139
5	5	0.118
5	10	0.124
10	1	0.3901
10	5	0.347
10	10	0.345

TABLE 4 *IV* for Increasingly Incorrect Specification of the θ Vector

True ratio	Ratio specified for design	Integrated variance
1:2	2:1	0.0125
1:2	5:1	0.0206
1:2	10:1	0.0247
1:5	2:1	0.0778
1:5	5:1	0.0996
1:5	10:1	0.1179
1:10	2:1	0.1965
1:10	5:1	0.2597
1:10	10:1	0.2787

ratio of the elements of the $\boldsymbol{\theta}$ vector varies from 2:1 to 10:1. We again generated three designs. We specified the three $\boldsymbol{\theta}$ vectors for the designs as (2, 1), (5, 1), and (10, 1). Figure 9 shows the resulting GP IMSE designs.

Here the variation in the look of the designs as one scans from the left to the right panel of the plot is more pronounced. Assuming that the $\boldsymbol{\theta}$ vector is (2, 1), there appear to be roughly four levels of X2 and the distribution of X1 values is more uniform. For ratios of 5 to 1 and 10 to 1, the plots both show roughly three levels of X2 again and the levels of X1 are more uniform.

In our choice of $\boldsymbol{\theta}$ vectors we are assuming that the response surface has more curvature in the X1 direction. Suppose that we are wrong and that there is actually more curvature in the X2 direction.

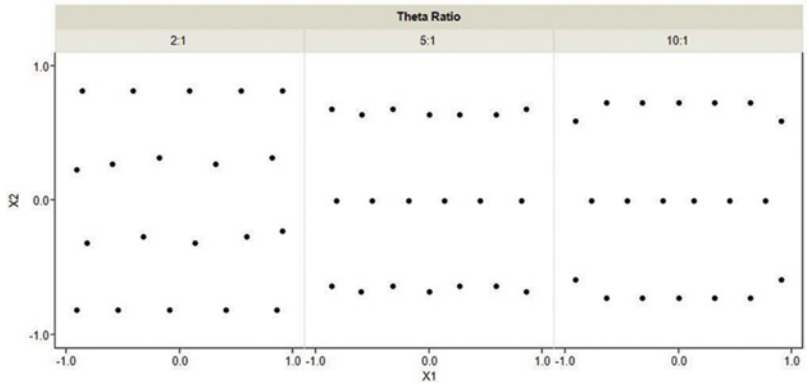

FIGURE 9 GP IMSE designs varying the ratio of the elements of the θ vector. (Color figure available online.)

Table 4 shows the resulting *IV* for increasingly incorrect specification of the $\boldsymbol{\theta}$ vector. As for Table 3, the magnitude of the true $\boldsymbol{\theta}$ vector had the largest effect on the *IV*. The effect of incorrect specification is less pronounced.

CASE STUDY RESULTS

Recall the NASA-sponsored air breathing propulsion experiment described in the Gaussian Process Model section. This section presents empirical prediction variance results for this case study. The simulation results are based on a CFD model built to mimic the flow field parameters within an open jet flame. Because the CFD is not available for commercial use, we created a mathematical model that predicts the CFD accurately to four decimal places. The CFD is based on a physical experiment described in R. T. Johnson et al. (2009). We are interested in modeling the response, oxygen, as a function of two input factors: *x*-axis and *y*-axis location. The output response is fit using the GP model. Because the theoretical *IV* study that we performed indicated that the performance of both the Mm LHD and the GP IMSE design were anticipated to be very good in this case, we elected to use both of these designs with $n = 30$ runs.

We evaluated the Mm LHD and the GP IMSE design by comparing 5,000 points generated based on the GP fit. For each point we computed the root mean squared error (RMSE) from each model from the equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{5000} [y(x_i - \hat{y}(x_i))]^2}{5000}} \qquad [3]$$

Table 5 summarizes the results showing the RMSE for each design.

Here we see little difference in actual prediction performance between the Mm LHD and the GP IMSE design, with the Mm LHD slightly outperforming the GP IMSE design. This is in general agreement with the results of our analytical study of design performance. More work comparing theoretical performance with respect to the GP model prediction variance and actual performance using both test functions and actual computer models would be of interest to practitioners.

TABLE 5 Root Mean Squared Prediction Error for Each Design Over 5,000 CFD Simulation Runs

Mm LHD	0.0007
GP IMSE	0.0008

CONCLUSIONS

This article compares theoretical prediction performance of five space-filling designs with respect to the theoretical integrated prediction variance (*IV*) of the GP model for experiments involving two to five factors and a range of sample sizes and magnitudes of the elements in the $\boldsymbol{\theta}$ vector. Based on our study, we can draw several conclusions.

All of the designs that we studied (Mm LHD, SP, U, ME, and GP IMSE or *I*-optimal) performed similarly with respect to *IV* when the sample sizes (τ) were at least $10d$ (that is, 10 times the number of factors in the experiment). Design performance did not change dramatically between $10d$ and $15d$, although we would recommend $15d$ as a generally safe rule of thumb for sample size. This is in general agreement with results reported by other researchers (see, for example, Loeppky et al. 2009). In general, as the complexity of the response surface increased, sample size requirements increased. This is an intuitive result.

What may not be as intuitive is that we found that the design performance for *IV* (averaged across all designs) depended on both the size of the elements in the $\boldsymbol{\theta}$ vector and the sample size. That is, there was a statistically significant interaction between these two factors. When the sample size was small ($5d$), *IV* performance deteriorated quickly as the average size of the elements in the $\boldsymbol{\theta}$ vector increased. If the sample size was as large as $15d$ the effect of the $\boldsymbol{\theta}$ vector elements was much smaller. There was some evidence that the SP and ME designs did not perform as well as the other three space-filling designs used in this study. The effect was not large but was observed for all cases of number of factors and size of the elements in the $\boldsymbol{\theta}$ vector.

The GP IMSE design performed as well as the Mn LHD and U designs, but to construct these designs the experimenter must assume values for the elements of the $\boldsymbol{\theta}$ vector. Consequently, we also evaluated the performance of the GP IMSE design with respect to how the elements of the $\boldsymbol{\theta}$ vector are specified in order to construct the design. We found that

assuming that all elements of the $\boldsymbol{\theta}$ vector are equal produces a GP IMSE design that performed well with respect to *IV* when the actual elements of the $\boldsymbol{\theta}$ vector ere equal, even if the magnitude of the estimates was incorrect. When the elements of the $\boldsymbol{\theta}$ vector differed considerably from each other and we had assumed that they were equal for design purposes, there was some degradation in performance. The worst case situations occurred when we assumed that there was more nonlinearity in one direction than in others and then found that the actual surface was fairly smooth in that direction but very nonlinear in others.

Finally, we reported the results of a two-factor case study involving an air breathing propulsion experiment. We fit the GP model to data from this computer model using both the Mm LHD and the GP IMSE design with 30 runs. We used these two designs because they are good performers and were anticipated to perform approximately the same based on our theoretical *IV* study. We were able to evaluate actual prediction variance performance by using a set of 5,000 runs that had been conducted but not used in model fitting. The RMSE of both models was approximately the same, implying that both designs performed similarly. This was only a single confirmation study and we encourage other researchers to conduct additional studies to evaluate our conclusions.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the NASA Langley Research Center. Specifically, thanks to Peter Parker, Phil Drummond, Andrew Cutler, Paul Danehy, Sara Tedder, and Daniel Bivolaru from NASA Langley Research Center and Gaetano Magnotti from the George Washington University for their interest, support, and application of statistical methods in the air breathing propulsion research efforts. We also thank the anonymous referees for many useful comments on earlier drafts of this article. The comments and suggestions greatly improved the presentation.

ABOUT THE AUTHORS

Dr. Rachel T. Silvestrini is an assistant professor in the Operations Research Department at the Naval Postgraduate School. She received her Ph.D. from Arizona State University in industrial engineering. Dr. Silvestrini's research interests include design and analysis of experiments and response surface methods.

Dr. Douglas C. Montgomery is Regents' Professor of Industrial Engineering and Statistics, ASU Foundation Professor of Engineering, and Co-Director of the Graduate Program in Statistics at Arizona State University. He received the Ph.D. in engineering from Virginia Tech. His professional interests are in statistical methodology for problems in engineering and science. He is a recipient of the Shewhart Medal, the George Box Medal, the Brumbaugh Award, the Lloyd S. Nelson award, the William G. Hunter award, and the Ellis Ott Award. He is one of the current chief editors of *Quality & Reliability Engineering International*.

Dr. Bradley Jones is the principal research fellow at the SAS Institute and a guest professor at the University of Antwerp. He is the inventor of the prediction profile plot, an interactive graph for exploring multidimensional response surfaces. At the SAS Institute he is responsible for the design of experiments capabilities in the JMP software package. He is a fellow of the American Statistical Association and a winner of the Brumbaugh Award for 2009.

REFERENCES

Allen, T. T., Bernshteyn, M. A., Kabiri-Bamoradian, K. (2003). Constructing meta-models for computer experiments. *Journal of Quality Technology*, 35(3):264–274.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., and Tu, J. (2007). A Framework for Validation of Computer Models. *Technometrics*, 49(2), 138–154.

Bursztyn, D., Steinberg, D. M. (2006). Comparison of designs for computer experiments. *Journal of Statistical Planning and Inference*, 136:1103–1119.

Chen, V., Tsui, K.-L., Barton, R., Meckensheime, M. (2006). A review on design, modeling and applications of computer experiments. *IEE Transactions*, 38:273–291.

Currin, C., Mitchell, T. J., Morris, M. D., Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86:953–963.

Fang, K. T. (1980). The uniform design: Application of number-theoretic methods in experimental design. *Acta Mathematicae Applicatae Sinica*, 3:363–372.

Fang, K. T., Li, R., Sudjianto, A. (2006). *Design and Modeling for Computer Experiments* Boca Raton, FL: Taylor & Francis.

Hussain, M. F., Barton, R. R., Joshi, S. B. (2002). Metamodeling: Radial basis functions, versus polynomials. *European Journal of Operational Research*, 138:142–154.

Jank, W., Shmueli, G. (2007). Modelling concurrency of events in on-line auctions via spatiotemporal semiparametric models. *Applied Statistics*, 56:1–27.

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

Johnson, R. T., Montgomery, D. C., Jones, B., Parker, P. A. (2010). Comparing computer experiments for fitting high order polynomial metamodels. *Journal of Quality Technology*, 42(1):86–102.

Johnson, R. T., Parker, P. A., Montgomery, D. C., Cutler, A. D., Danehy, P. M., Rhew, R. D. (2009). Design strategies for response surface models for the study of supersonic combustion. *Quality and Reliability Engineering International*, 25:365–377.

Jones, B., Johnson, R. T. (2009). The design and analysis of the Gaussian process model. *Quality and Reliability Engineering International*, 25:515–524.

Ko, C.-W., Lee, J., Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Operations Research*, 43:684–691.

Liefvendahl, M., Stocki, R. (2006). A study on algorithms for optimization of Latin hypercubes. *Journal of Statistical Planning and Inference*, 136:3231–3247.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., Ye, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics*, 48:478–490.

Loeppky, J. L., Sacks, J., Welch, W. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4):366–376.

McKay, N. D., Conover, W. J., Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245.

Mease, D., Bingham, D. (2006). Latin hyperrectangle sampling for computer experiments. *Technometrics*, 48:467–477.

Roux, W., Stander, N., Gunther, F., Mullerschon, H. (2006). Stochastic analysis of highly non-linear structures. *International Journal for Numerical Methods in Engineering*, 65:1221–1242.

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

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

Santner, T. J., Williams, B. J., Notz, W. I. (2003). *The Design and Analysis of Computer Experiments Springer Series in Statistics*. New York: Springer-Verlag.

Shewry, M. C., Wynn, H. P. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, 14:898–914.

Storlie, C. B., Helton, J. C. (2007). Multiple predictor smoothing methods for sensitivity analysis: Example results. *Reliability Engineering & System Safety*, 93:55–77.

Tyre, A., Kerr, G. D., Tenhumberg, B., Bull, M. (2007). Identifying mechanistic models of spatial behaviour using pattern-based modelling: An example from lizard home ranges. *Ecological Modeling*, 208: 307–316.

Wang, Y., Fang, K. T. (1981). A note on uniform distribution and experimental design. *KeXue TongBao*, 26:485–489.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34:15–25.