Statistics and Applications
Volume 6, Nos.1 & 2, 2008 (New Series), pp.207-233

Sequential adaptive designs in computer experiments for response surface model fit

Chen Quin Lam and William I. Notz
Department of Statistics, The Ohio State University,
1958 Neil Avenue, 404 Cockins Hall, Columbus, USA

Abstract

Computer simulations have become increasingly popular as a method for studying physical processes that are difficult to study directly. These simulations are based on complex mathematical models that are believed to accurately describe the physical process. We consider the situation where these simulations take a long time to run (several hours or days) and hence can only be conducted a limited number of times. As a result, the inputs (design) at which to run the simulations must be chosen carefully. For the purpose of fitting a response surface to the output from these simulations, a variety of designs based on a fixed number of runs have been proposed.

In this paper, we consider sequential adaptive designs as an "efficient" alternative to fixed designs. We propose new adaptive design criteria based on a cross validation approach and on an expected improvement criterion, the latter inspired by a criterion originally proposed for global optimization. We compare these new designs with others in the literature in an empirical study and they are shown to perform well.

Key words: Cross validation; Gaussian stochastic process model; Kriging; Non-stationary response surfaces; Sequential designs; Adaptive designs.

Preamble: In the past four decades, Professor Aloke Dey has played a major role in the development and teaching of experimental design theory. In line with his interests in designs for fitting response

surfaces, optimality and robustness of designs, it is our pleasure to contribute to this special issue honoring his work. And, we wish him many more years of productive contributions to this field.

1 Introduction

In the last decade or so, computer experiments have become very popular with the advent of affordable computing power. Traditionally, physical experiments have been used to establish a cause-and-effect relationship between input variables and the response output. Given the increasingly complex nature of scientific research, many physical experiments are difficult, if not impossible, to carry out. In their place, computer simulations have been used to provide a representation of the "real" physical system. Put in a simplistic way, these simulations are attempts to represent the complex reality in a computer code (or mathematical model). However, for code that runs slowly, it is not possible to carry out computer simulations at very fine grids in any realistic time frame. Thus, computer experiments are often performed to allow one to determine an approximation to the unknown response surface generated by the code. This has led to the development of statistical methodologies for predicting the unobserved responses of the code at selected input points. The approach taken in this paper assumes that the response can be modeled as a realization from a Gaussian stochastic process (see Sacks *et al.*, 1989, and Santner *et al., 2003*).

Using the Gaussian stochastic process (GASP) model as an approximation to the actual computer code, the focus of this paper is on the selection of input points at which to run the simulations so as to obtain a good overall fit (i.e., predictive accuracy) of the GASP model. We propose several sequential adaptive designs and compare them against one another and also against a fixed-point design.

Experimental designs relevant to computer experiments can be broadly categorized into two classes: space-filling designs and criterion-based designs. Given that the goal is to achieve good predictive accuracy, it is intuitive to consider a space-filling design strategy in order to minimize the overall prediction error of the GASP model across the

entire input space. Examples of space-filling designs include methods based on selecting random samples (e.g., Latin hypercube designs (LHD)), distance based designs (e.g., maximin and minimax designs), uniform designs, and even sequential space-filling designs (e.g., Sobol sequences). See Santner *et al.* (2003), Koehler and Owen (1996) and Bates *et al.* (1996) for thorough discussions of different design strategies. While space-filling designs are good for initial exploratory purposes, they are constructed based on the assumption that interesting features of the true computer model are equally likely across the entire input space. Selection of input points for these designs are not adaptive to what we learn about the response surface as we observe the code and space-filling designs may result in a loss of prediction accuracy and efficiency in many situations. The second class of designs are constructed based on some statistical criteria rather than the geometric criteria used in space-filling designs. Designs based on certain optimality criteria, such as mean squared prediction error and the notion of entropy, have been used to construct designs for computer experiments. However, they are not easily implemented because they depend on the unknown correlation parameters present in the GASP model.

The most popular designs for computer experiments are LHDs, mainly due to availability of software to generate them easily even when the number of inputs is large. A major limitation of the LHD and fixed-point designs in general is that they make no use of information gained about the shape of the response surface as we add observations. While designs based on certain optimality criteria, such as mean squared prediction error and entropy, can be converted into sequential designs, it is not clear if these designs will result in an accurate predictive model, because they also make no use of what one learns about the response surface from the observed responses. This will be investigated further in the empirical study in Section 4.

In general, we are optimistic that sequential designs can be more effective and efficient for prediction of responses at unobserved input points than fixed-point designs if the sequential designs are adaptive (i.e., the GASP model is updated sequentially and design points are added based on the new information/features of the approximated response surface). It is worth emphasizing that some space-filling de-

signs are sequential (e.g., Sobol sequences) but not adaptive. Several sequential as well as adaptive designs, based on cross validation and a modified expected improvement criterion (Schonlau, 1997), are proposed to obtain a GASP model that gives accurate predictions.

The outline of this paper is as follows. In Section 2, we present the stochastic process model assumed in this study and the choice of correlation functions used in the model. Following this in Section 3, we present the various sequential design criteria proposed in this study. Three examples are given in Section 4 to illustrate the effectiveness of the various designs. We conclude with a discussion of the proposed criteria and simulation results.

2 Statistical model

The computer code for simulation can be thought of as a function h with inputs denoted by $\boldsymbol{x} \in \mathcal{X} \subset \Re^p$. The output from the computer code is denoted as $y = h(\boldsymbol{x})$. In this paper, we restrict attention to the case of a univariate output from the computer code or simulator. One can treat the simulator as a black box and model the computer ouput as a stochastic process to be described in Section 2. For our approach, the best linear unbiased predictor is used to predict the response at unobserved $\boldsymbol{x}$, based on the available training data.

2.1 Model and best linear unbiased Predictors

Following the approach of Sacks *et al.* (1989), it is assumed that the deterministic output $y(\boldsymbol{x})$ is a realization of a stochastic process (or random function), $Y(\boldsymbol{x})$. The typical model used in computer experiments is

$$Y(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x}))^T$ is a $k \times 1$ vector of known regression functions, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ is a $k \times 1$ vector of unknown regression parameters. And, $Z(\boldsymbol{x})$ is assumed to be a random process with mean 0, variance σ^2 and a known correlation function $R(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The $Z(\cdot)$ component models the systematic local trend or bias from

the regression part of (1) and the correlation function $R(\cdot)$ essentially controls the smoothness of the process.

Suppose we have n observations from the computer simulator. Let $\boldsymbol{Y}^n = (Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))'$ denote the responses from the computer simulator and suppose the goal is to predict the response $Y(\boldsymbol{x}_0)$ at some untried $\boldsymbol{x}_0$ with a linear unbiased predictor

$$\hat{Y}(\boldsymbol{x}_0) = c^T(\boldsymbol{x}_0)\ \boldsymbol{Y}^n.$$

Cressie (1993) provides more details on linear unbiased predictors in the context of geostatistical kriging.

The best linear unbiased predictor (BLUP) finds the vector $c(\boldsymbol{x}_0)$ that minimizes the *mean squared prediction error* (MSPE)

$$MSPE[\hat{Y}(\boldsymbol{x}_0)] = E[(c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n - Y(\boldsymbol{x}_0))^2] \tag{2}$$

subject to the unbiasedness constraint $E[c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n] = E[Y(\boldsymbol{x}_0)]$. The BLUP can be shown to be

$$\hat{Y}(\boldsymbol{x}_0) = c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n = \boldsymbol{f}^T(\boldsymbol{x}_0)\hat{\boldsymbol{\beta}} + r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}\hat{\boldsymbol{\beta}}), \tag{3}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{Y}^n$ is the generalized least-squares estimate of $\boldsymbol{\beta}$ and $\boldsymbol{F} = [f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)]^T$ is the $n \times k$ matrix of regressors whose $(i,j)th$ element is $f_j(x_i)$ for $1 \le i \le n$, $1 \le j \le k$. The MSPE of the BLUP is then given by

$$\begin{aligned} MSPE[\hat{Y}(\boldsymbol{x_0})] &= \sigma^2[1 - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x_0}) + \\ (f^T(\boldsymbol{x_0}) &- r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}(f^T(\boldsymbol{x_0}) - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})^T]. \end{aligned} \tag{4}$$

where $r(\boldsymbol{x}_0) = (R(\boldsymbol{x}_1, \boldsymbol{x}_0), ..., R(\boldsymbol{x}_n, \boldsymbol{x}_0))^T$ is the $n \times 1$ vector of correlations between observations at the previously sampled points, $\boldsymbol{Y}^n$, and $Y(\boldsymbol{x}_0)$. Usually, $f^T(\boldsymbol{x})\boldsymbol{\beta}$ in (1) is simply assumed to be a constant mean term, β, unless there is strong evidence that a more complex function (e.g., a polynomial function or even a crude version of the "simulator") is needed to capture a global trend. In practice, use of only a constant mean term has been found to work well. The stochastic process $Z(\boldsymbol{x})$ captures the local trend which usually suffices to produce excellent fit.

Given that the correlation function $\boldsymbol{R}(\cdot)$ is known, the BLUP can be easily calculated using (3). Typically, the correlation parameters have to be estimated (for example, by maximum likelihood estimation) and the resulting predictor is termed as the *empirical best linear unbiased predictor* (EBLUP).

2.2 Parametric correlation functions

As seen from the equations (3) and (4) above, the correlation function $\boldsymbol{R}(\cdot)$ plays an important role and has to be specified by the user. This section presents a review of some of the neccessary restrictions imposed on $\boldsymbol{R}(\cdot)$. A valid correlation function must possess certain properties such as (i) $R(0) = 1$, (ii) $\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j R(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \forall n, \forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and all real $w_1, \ldots, w_n$; (iii) $R(h) = R(-h)$ and does not depend on the location. We consider correlation functions for $x_1, x_2 \in S$

$$R(\boldsymbol{x}_1, \boldsymbol{x}_2) = R(|\boldsymbol{x}_1 - \boldsymbol{x}_2|)$$

so that $Z(\cdot)$ in (1) is stationary. In higher dimensions (i.e. 2 or higher), taking the products of correlation across each dimension is a common practice for computational convenience,

$$R(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{j=1}^{p} R(|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}|).$$

These are sometimes called separable correlation functions. Two popular choices are the cubic and power exponential correlation functions and their one-dimensional forms are given below.

Cubic correlation. The non-negative cubic correlation function takes the form of

$$\begin{aligned} R(d) &= 1 - 6\left(\tfrac{d}{2}\right)^2 + 6\left(\tfrac{|d|}{\theta}\right)^3, && |d| < \tfrac{\theta}{2} \\ &= 2\left(1 - \tfrac{|d|}{\theta}\right)^3, && \tfrac{\theta}{2} \leq |d| < \theta \\ &= 0, && |d| \geq \theta \end{aligned}$$

where $\theta > 0$ (see Currin *et al.*, 1991, and Mitchell *et al.*, 1990). This correlation function permits a very local correlation structure since the range parameter θ can be made very small. Another appealing feature of this correlation function is that beyond distance θ, the correlation between two points drops to zero, thus providing some intuition concerning the interpretation of θ. The prediction function is a piecewise cubic spline interpolating predictor in the context of computer experiments.

Power exponential correlation. Another very popular correlation function takes the form of

$$R(d) = exp(-\theta|d|^p), \tag{5}$$

where $0 < p \leq 2$ and $\theta \in (0, \infty)$. For the special case of $p = 2$, this corresponds to the Gaussian correlation function which gives an EBLUP that is infinitely differentiable. Taking $p = 1$ gives the exponential correlation function. For $0 < p < 2$, the EBLUP is continuous but not differentiable. As θ increases, the dependence between two sites decreases but does not go to zero. See Sacks *et al.* (1989) for an application with this correlation function. If one knows that the physical process being modeled by the simulator is smooth, $p = 2$ should be used.

Both the cubic and power exponential with $p = 2$ correlation functions will be used for the examples in Section 4. We will also use the product correlation structure, $R(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{j=1}^{p} R(|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}| \mid \theta_j)$, and let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$.

3 Sampling design criteria

The choice of design is crucial to the success of building an efficient and accurate GASP model. This section begins with an introduction to LHDs and some statistical design criteria, such as those based on the mean squared prediction error and entropy, that have been used to construct designs for computer experiments. Unfortunately, these statistical design criteria are not implementable because they require knowledge of the unknown parameters in the correlation function. To overcome this problem, we consider a sequential implementation of these criteria in Subsection 3.2 where previous observations are used to estimate the parameters. We also consider several alternative design criteria based on the cross validation approach in subsection 3.3. Finally, we consider an expected improvement (EI) criterion, inspired by a criterion studied in Schonlau (1997) for our objective of obtaining a good global model fit, specifically targeting subregions with interesting features.

3.1 Latin hypercube designs

Latin hypercube (LH) sampling was first introduced by McKay *et al.* (1979) as an alternative to simple random sampling and stratified sampling. LH sampling is a way to ensure that the input points are spread evenly over the range of each input separately. Despite their marginal space-filling properties, not all LHDs are space-filling across the entire input space. Attempts are made to improve on this by incorporating distance-based designs, such as maximin distance, and other criteria-based designs within the class of LHD (see Santner *et al.*, 2003, and Koehler and Owen, 1996). There exist a number of space-filling design criteria as mentioned in Section 1 but studies (e.g., Marin, 2005) suggest they perform similarly. Thus, in Section 4 we use a maximin LHD as representative of a space-filling design.

3.2 Sequential criterion-based optimal designs

Mean squared prediction error designs. The *mean squared prediction error* equation in (4) can be used as a design criterion. It can be implemented sequentially by selecting a new input point, $\boldsymbol{x}_0$, with the largest MSPE based on the constant mean GASP model that is fitted using the existing input points,

$$\max_{\boldsymbol{x}_0} MSPE(\boldsymbol{x}_0) \;=\; \max_{\boldsymbol{x}_0}\left(\sigma^2\left[1 - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}r(\boldsymbol{x_0}) + \frac{(1 - 1'\boldsymbol{R}^{-1}r(\boldsymbol{x}_0))^2}{1'\boldsymbol{R}^{-1}1}\right]\right). \tag{6}$$

Given that the correlation $r(\boldsymbol{x}_0)$ decreases with increasing distance between two input points (as is the case for the cubic and power exponential correlation functions), this *maximum MSPE* design tends to spread points out and often initially on the boundaries of the input space. Unless important features of the true response surface are on or near the boundary, the fitted surface can be poor unless the total number of observations is large enough to guarantee the interior of the design region is adequately sampled.

Sacks *et al.* (1989) considered the *integrated mean squared prediction error* (IMSPE) criterion

$$\int_{\mathcal{X}} \frac{\mathrm{MSPE}[\hat{Y}(\boldsymbol{x})]}{\sigma^2}\, w(\boldsymbol{x}) d\boldsymbol{x} \tag{7}$$

where $w(.)$ is a non-negative function satisfying $\int_{\mathcal{X}} w(\boldsymbol{x})d\boldsymbol{x} = 1$. Typically, one might consider a uniform weighting and simply take the average of the MSPE across all $\boldsymbol{x}$. An n-point design is said to be IMSPE-optimal if it minimizes (7) over the set of candidate points $\mathcal{X}$.

Direct implementation of IMSPE as a sequential criterion has been found to lead to clumping of new input points around existing points (see Sacks *et al.*, 1989). In our attempt to introduce new sequential sampling designs, we propose a slight modification to the IMSPE criterion by taking into account the distance of candidate points to the existing input points and imposing a penalty to prevent the additional design points from clustering together. The new criterion is to select $\boldsymbol{x}_0$ that

$$\min_{\boldsymbol{x}_0 \in \mathcal{X}} \{IMSPE(\boldsymbol{x}_0)/\min(d(\boldsymbol{x}_i, \boldsymbol{x}_0))\} \tag{8}$$

where $\boldsymbol{x}_i$ denotes an existing input point that is closest to $\boldsymbol{x}_0$. The distance penalty is incorporated to push subsequent points away from existing input points and hence prevent the clumping problem.

Entropy designs. The amount of information provided by an experiment can also be used as a design criterion. Shewry and Wynn (1987) introduced the notion of sampling by maximum entropy when the design space is discrete. They showed that the expected change in information provided by an experiment is maximized by the design D that maximizes the entropy of the observed responses.

Recall in Subsection (2.1) that the training data have the following conditional distribution

$$\boldsymbol{Y}^n|\boldsymbol{\beta}, \boldsymbol{\theta} \sim N(\boldsymbol{F}\boldsymbol{\beta}, \sigma_z^2 \boldsymbol{R}).$$

Using a Bayesian approach, one can specify a prior distribution for the $\boldsymbol{\beta}$ coefficients, say, $\boldsymbol{\beta} \sim N_p(\boldsymbol{b}_0, \tau^2 \boldsymbol{V}_0)$. Then, the marginal covariance matrix of the observations $\boldsymbol{Y}^n|\boldsymbol{\theta}$ can be expressed as

$$\sigma_z^2 \boldsymbol{R} \; + \; \tau^2 \boldsymbol{F} \boldsymbol{V}_0 \boldsymbol{F}^T. \tag{9}$$

One can show (see Koehler and Owen, 1996) that the *maximum entropy* design maximizes the determinant of the observation variance in (9).

The choice of prior distributions for the $\boldsymbol{\beta}$ coefficients will affect the quantity that the criterion is maximizing (see Koehler and Owen,

1996). We consider two simple cases discussed in Koehler and Owen (1996):
(i) If the $\boldsymbol{\beta}$ are treated as fixed (i.e. $\tau^2 = 0$), the maximum entropy criterion reduces to

$$det(\sigma_z^2 \boldsymbol{R}). \tag{10}$$

(ii) If the $\boldsymbol{\beta}$ are diffuse (i.e. $\tau^2 \to \infty$), one can show the maximum entropy criterion becomes

$$\det(\sigma_z^2 \boldsymbol{R}) \det(\boldsymbol{F}^T (\sigma_z^2 \boldsymbol{R})^{-1} \boldsymbol{F}). \tag{11}$$

The *maximum entropy* design criterion, either (10) or (11), can also be modified for use as a sequential algorithm. $\boldsymbol{R}$ (which now includes the candidate point as well) can be partitioned into

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_n & r_n(\boldsymbol{x}_0) \\ r_n(\boldsymbol{x}_0)' & 1 \end{pmatrix}, \tag{12}$$

where $\boldsymbol{R}_n$ is the correlation matrix based on the existing n design points only. The cross correlation between the observation at a new candidate point $\boldsymbol{x}_0$ and observations at the existing design points is denoted by the vector $r_n(\boldsymbol{x}_0)$. Hence, one can show the sequential maximum entropy criterion based on (10) reduces to selecting a new point that maximizes

$$1 - r^T(\boldsymbol{x}_0) \boldsymbol{R}_n^{-1} r(\boldsymbol{x}_0). \tag{13}$$

Notice that (13) is very similar to (6) except for the last term. For (11) where the $\boldsymbol{\beta}$ coefficients have a diffuse prior distribution, the sequential maximum entropy criterion is equivalent to the sequential MSPE criterion in (6).

3.3 Cross validation prediction error criteria

Cross validation is a popular method for estimating model parameters and for model validation. We use cross validation to come up with a semi-parametric (prediction-oriented) measure of the prediction error as an alternative to the MSPE for the stochastic model specified in (4). In turn, this prediction error will be used as a design criterion to

select additional input points. This approach is motivated by noting that the MSPE of the model depends only on the distance between sampled input points, $\boldsymbol{x}$, and the correlation function $\boldsymbol{R}(\cdot)$, but not the response values observed at these input points or the predicted values given by the fitted surface. By considering criteria based on cross validation, we use the observed and predicted responses.

Let $\boldsymbol{x}$ denote a candidate point and $\hat{Y}^{(-j)}(\boldsymbol{x})$ denote the EBLUP of $y(\boldsymbol{x})$ based on all the data except $\{\boldsymbol{x}_j, y(\boldsymbol{x}_j)\}$ where $j = 1, \ldots, n$, while $\hat{Y}_n(\boldsymbol{x})$ denotes the EBLUP of $y(\boldsymbol{x})$ using all the data. To reduce computational burden, the correlation parameters for the EBLUP are estimated based on all n observations. The cross validation prediction error (XVPE) criterion is then to pick the point, $\boldsymbol{x}$, that has the largest "mean" prediction error, in senses we now define.

We first consider the *arithmetic mean* in

$$XVPE_{\mathrm{A}}(\boldsymbol{x}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2 \times \min_j(d(\boldsymbol{x}_j, \boldsymbol{x}))} \quad (14)$$

which was also considered in Jin *et al.* (2002) in the context of radial basis function modeling. A penalty term, $d(\cdot)$, based on Euclidean distance is incorporated to penalize candidate points that are closer to existing sampled points to prevent the next point picked from being close to one of the existing design points.

We propose three new criteria for the cross validation approach that avoid the "distance" penalty by using the *geometric mean, harmonic mean,* and the *maximin* error as alternative summaries of the cross validation prediction variability. These three error summaries also avoid selecting points very close to existing design points. Unlike the maximum MSPE, maximum entropy, and the cross validation method in (14), they do not make explicit use of the correlation matrix $R(\cdot)$ or distance from existing points to penalize candidate points.

The *geometric* mean is appealing because, unlike the arithmetic mean, candidate points close to existing design points will not likely be selected as the product term penalizes the small $(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))$

prediction error components.

$$XVPE_{\mathrm{G}}(\boldsymbol{x}) = \sqrt[n]{\prod_{j=1}^{n}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2} \tag{15}$$

The *harmonic* mean tends to be more affected by small values than large values. Since the harmonic mean of the set of n cross validation prediction errors tends strongly toward the smallest elements of the set, it tends (compared to the arithmetic mean) to mitigate the impact of larger values and aggravate the impact of small ones. As a result, it prevents subsequent design points from clumping together.

$$XVPE_{\mathrm{H}}(\boldsymbol{x}) = \frac{n}{\sum_{j=1}^{n}\frac{1}{(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2}} \tag{16}$$

The third summary is to compute the minimum cross validation prediction error for every candidate point and choosing the next point that has the largest error. We shall call it the *maximin* criterion given by

$$XVPE_{\mathrm{M}}(\boldsymbol{x}) = \min_{j}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2. \tag{17}$$

3.4 Expected improvement for global fit criterion

The *expected improvement* (EI) criterion proposed by Schonlau(1997) was originally developed as a global optimization design criterion. Instead of locating the global optimum or optima, we consider a modification of the criterion to obtain a good global model fit of the GASP model. The objective is to search for "informative" regions in the domain that will help improve the global fit of the model. By informative we mean regions with significant variation in the response values.

Suppose we have the computer outputs $y(\boldsymbol{x}_j)$ at sampled points $\boldsymbol{x}_j$, $j = 1, ..., n$. For each potential input point $\boldsymbol{x}$, its improvement is defined as

$$I(\boldsymbol{x}) = (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 \tag{18}$$

where $y(\boldsymbol{x}_{j^*})$ refers to the observed output at the sampled point, $\boldsymbol{x}_{j^*}$, that is closest (in distance) to the candidate point $\boldsymbol{x}$. We shall determine this nearest sampled design point using Euclidean distance. The *expected improvement for global fit* (EIGF) criterion is to choose the next input point that maximizes the expected improvement

$$E(I(\boldsymbol{x})) = (\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 + var(\hat{Y}(\boldsymbol{x})). \tag{19}$$

The expected improvement in (19) consists of two search components-local and global. The first (local) component of the expected improvement will tend to be large at a point where it has the largest (response) increase over its nearest sampled point. The second (global) component is large for points with the largest prediction error as defined in (4), i.e., points about which there is large uncertainty and, as mentioned in Section 3.2, these tend to be far from existing sampled points.

4 Examples

The following examples illustrate the implementation and prediction performance of the various sequential designs and the fixed-point maximin LHD. Various functions are used as "true" functions to compare the prediction performances of these designs using a small number of sampled points. The design strategies to be compared are:

- Sequential maximum mean squared prediction error (m)
- Sequential maximum entropy (e)
- Cross validation approaches: arithmetic mean penalized by distance (xa), geometric mean (xg), harmonic mean (xh), maximin criteria (xm)
- Sequential integrated mean squared prediction error, penalized by distance (id)
- Expected improvement for global fit (ei)

- Fixed-point or fixed sample size maximin Latin hypercube design (s)

(the abbrevations in parentheses will be used to denote these methods later in the figures).

The total number of design points is fixed. For this number, N, we consider the rule of thumb suggested in Jones et al. (1998) for selecting a fixed-point design, namely to use $N = 10 \times p$ points, where p is the number of dimension of the input space. Due to the complexity of the response surfaces used in our examples, the final number of input points are at least 30, because this rule of thumb did not always provide enough points for any method to perform well.

There is not a unique maximin LHD. In addition, the software that we used does not neccessarily produce a maximin LHD but rather one that is "nearly" a maximin LHD, and thus adds additional variation to the choice of designs. Our comparisons are based on 30 runs of our software for generating maximin LHDs. For sequential designs, this means 30 different starting designs all approximately maximin LHDs. In addition, different numbers of starting design points (denoted as N_0) are also considered for the sequential designs. The initial starting designs are thus generated using an N_0-point (nearly) maximin LHD. N_0 is chosen to be 5, 10, 15 or 20 depending on the example. For this study, the smallest number of starting design points is taken to be 5 for the two-dimensional functions. We would suggest starting the initial design with at least N_0 = number of dimensions + 2 (i.e., 4 in the two-dimensional functions) so as to capture the non-linearity of the surface at the start of the algorithm. However in our study, we chose to start with 5 points since the maximin LHD criterion is used to generate the starting design points and it was found that starting with 4 points put all points near or on the boundaries and did not work well. Starting with 5 points tended to ensure at least one point is in the interior region.

The values of the correlation parameters are estimated by maximum likelihood in this study and they are updated at every stage when a new input point is added. For the cross validation methods, we choose not to re-estimate the correlation parameters, $\boldsymbol{\theta}$, for each of the j^{th} observation deletions. The $\boldsymbol{\theta}$ are estimated using the entire

n observations at each stage.

Prediction accuracy of each of the designs is evaluated using the empirical *root mean squared prediction error* (ERMSPE),

$$ERMSPE = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}(\boldsymbol{x}_i) - y(\boldsymbol{x}_i))^2}{m}} \tag{20}$$

where $\boldsymbol{x}_i$, $i = 1, ...m$ ($m >> N_0$) are a grid of points used for evaluating the prediction accuracy and m is the total number of grid points; $\hat{y}(\boldsymbol{x}_i)$ is the predicted value at the $\boldsymbol{x}_i$; $y(\boldsymbol{x}_i)$ are the true values. We used a regular grid, but some other method (e.g., maximin LHD) of choosing the m points could be used provided the points are evenly spread over $\mathcal{X}$. Boxplots are used for each of the test functions to show the distribution of the ERMSPE for the 30 runs.

4.1 Test functions and features

In this study, three examples are used to evaluate the predictive performace of the GASP model with the input points chosen by the various design criteria. Details about the functions are given below and a plot of the true response surfaces is shown in Figure 1.

Function 1: Six-hump camel-back function
This surface has features both at the boundaries and interior region. The function for the six-hump camel-back surface proposed in Branin (1972) is

$$f(x_1, x_2) = (4 - 2.1x_1^2 + \frac{x_1^4}{3})x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2,$$

where $x_1 \in [-2, 2],\ x_2 \in [-1, 1]$. The true surface is plotted in Figure 1 on $m = 30 \times 30 = 900$ points which coincide with the m points used to evaluate the designs in (20). Due to the complexity of the surface, the final number of input points, N, is taken to be 40.

Function 2: Simulated surface
Next, we have a surface where most of the features of the surface lie in the middle of the domain, where $x_1 \in [-5, 10],\ x_2 \in [0, 15]$.

This surface is generated by combining four bivariate Gaussian density functions each centered at different locations of the input space. The input domain is finely divided into $m = 40 \times 40 = 1{,}600$ points. This surface is constructed to examine the performances of the design strategies in a setting where the boundaries are "flat". The final number of input points, N, is 40.

Function 3: Two-dimensional exponential function
We consider the two-dimensional exponential function as an example of a non-stationary looking response function (also used in Gramacy, 2005) given by

$$y = x_1 \exp(-x_1^2 - x_2^2) \tag{21}$$

for $x_1, x_2 \in [-2, 6]$. This surface has two distinctively different regions (i.e. non-stationary) but the transition across the regions is smooth. The features lie mainly in the region where x_1 and x_2 are both negative. The input domain is finely divided into $m = 40 \times 40 = 1{,}600$ points. The final number of input points, N, is 30. The motivation for comparing the various designs in a non-stationary setting arose from studies by Gramacy (2005) and Farhang-Mehr and Azarm (2005). One might expect that procedures based on a model that assumes a stationary process would not perform well here but, as we will see in Section 4.2, this is not neccessarily the case. It suggests that a good design can lead to good fit with a GASP model even for non-stationary looking functions.

4.2 Results: Comparison of design criteria

Results from our simulation study show that there are significant differences in the final designs and predictive accuracy depending on whether the cubic or Gaussian correlation function is used. For function 1, the use of the cubic correlation for all the designs results in more accurate prediction of the response surfaces compared to the Gaussian correlation (see Figure 2). As for function 2 (see Figure 3), the Gaussian correlation function seems to perform better (with less outlying ERMSPE values) at least for smaller starting designs (i.e. $N_0 = 5$ and 10). With $N_0 = 20$, predictions using the two types of correlation are comparable. The two correlation functions are comparable for function 3 (Figure 4). In all three cases, for the "best"

designs, the cubic performs as well as the Gaussian. Based on various examples that we have examined (including the three in this section), the use of the cubic correlation function is found to be a more robust option for designs that perform well. For subsequent discussions, predictive performances of the design criteria will be based on the cubic correlation function.

Function 1: The maximin LHD design (s) is the worst performer among all the designs (see Figure 2). Both the maximum MSPE (m) and maximum entropy (e) criteria outperform other sequential criteria but the differences are not large. The other sequential procedures are roughly comparable although the cross validation with the harmonic mean and the maximin criteria (xh and xm) look the worst given their larger ERMSPE median and spread. The fact that the maximum MSPE (m) and maximum entropy (e) criteria select more points on the boundaries seems to be an advantage in this example. Although there are some larger values of ERMSPE for some of the criteria, the predicted surfaces based on these designs reproduce the true surface rather accurately except for the top-right corner of the surface.

Function 2: Among the sequential designs, the cross validation, with the geometric mean (xg) (except for $N_0 = 5$) and the arithmetic mean (xa), and EIGF (ei) designs are the better performers in this example where most of the features are in the interior of the input domain. This example again highlights the tendency of the maximum MSPE (m) and maximum entropy (e) criteria to place relatively more input points on the boundaries and thus results in a poorer fit for this function (compared to function 1). Except for the EIGF criterion (ei), the other sequential designs generally perform better with a larger starting design. The fixed-point maximin LHD (s) and integrated mean squared prediction error (id) designs do not perform too badly in this example.

Function 3: Here, for both correlation functions, the EIGF criterion (ei) with $N_0 = 5$ stands out as the best (see Figure 4). The closest competitor is the cross validation with the arithmetic mean (xa) criterion with $N_0 = 20$, while the GASP model based on the other criteria fails to approximate the response surface well in most of the 30 runs. Interestingly, the fixed-point maximin LHD (s) outperforms many of the sequential designs. Due to its "non-adaptive" space-filling crite-

rion, it manages to detect some of the non-stationary features in the bottom left region but too much sampling effort is wasted in the flat region of the surface.

The narrow spread of the ERMSPE (in Figure 4) using the EIGF criterion (ei) shows that it is not too sensitive to the variation in the starting design that seem to negatively affect the other sequential designs.

Figure 5 shows comparative plots of predicted surfaces of function 3 using the EIGF criterion (ei), cross validation with the arithmetic mean (xa), and the fixed-point maximin LHD (s). In the second and third row, it is very encouraging to see that the *worst* case prediction using the EIGF criterion (ei) with $N_0 = 5$ does not perform too badly compared to the *best* case prediction using the cross validation with the arithmetic mean (xa) criterion with $N_0 = 5$. The EIGF criterion (ei) manages to identify the irregular region very quickly and focuses most of the sampling effort there. Again, the EIGF criterion (ei) performs better with a smaller initial design of $N_0 = 5$. It is noted that starting the EIGF criterion (ei) with a larger design ($N_0 = 15$) leaves fewer points (15) to add and can sometimes result in most of the added sampling effort being concentrated on only one of the two "peaks", as shown in the worst predicted surface in the fourth row of Figure 5.

As an informal comparison, we note that the predicted surface with $N = 30$ input points using the EIGF criterion (ei) is more accurate (graphically) compared to the Bayesian Treed approach (see example in Gramacy, 2005). This suggests that stationary GASP models with good designs can fit non-stationary looking surfaces as well as methods that attempt to account for nonstationarity. This needs to be investigated further to understand the roles design and model play, as it may be that the Bayesian Treed model with a suitable design is very efficient. Overall, the EIGF criterion (ei) with a small starting design and the cross validation with the arithmetic mean criterion (xa) with a larger starting design perform well in all examples with the cubic correlation function.

5 Discussion and conclusion

For the objective of achieving good global model fit, it has been demonstrated that sequential adaptive designs typically outperform fixed-point designs such as the maximin LHD used in our examples. Studies in Marin (2005) suggests similar results will occur if other fixed-point designs are used. Among the sequential adaptive designs, the cross validation prediction error criterion with the arithmetic mean and larger starting designs, and EIGF criterion with smaller starting designs are very competitive in terms of prediction accuracy using the GASP model with the cubic correlation. Both criteria with a cubic correlation function are found to perform well in a variety of examples and are never significantly outperformed by any of the other designs. The adaptive property of these design criteria in this study enable the GASP model to identify interesting features in the input space and result in a more accurate statistical predictor. Also, sequential algorithms have the desirable property that additional observations are naturally accommodated if an increased budget or the need to improve the accuracy of the GASP model allows or requires additional observation. Similar results have been noted for higher dimensional examples not presented in this paper.

Other issues do arise during the implementation of these sequential designs. For instance, one of the key issues in sequential design is the number of starting design points. This is crucial to their success in surface predictions. Being the two top performing criteria in the examples, the EIGF criterion seems to perform better with smaller starting designs ($N_0 = 5$) while the cross validation with the arithmetic mean is superior with a larger starting design (e.g., N_0 being half of the final number of points, N). A decision also has to be made on the final number of input points. Although this number has been fixed in our examples, it is generally not clear exactly what this number should be but one can make use of the usual cross validation approach for assessing model fit to decide on a stopping criterion.

In applying the various designs to a non-stationary looking response function, we have also shown that the naive approach of specifying a single stationary GASP model across the entire input space of a clearly non-stationary surface need not suffer in terms of prediction

if the design criterion is able to target regions with high variation in the response. Further refinements can be made to the design and/or model approach taken in this paper. For example, one might combine sequential adaptive designs with more complicated stochastic models, such as the Bayesian Treed approach by Gramacy (2005). However, this is the topic of further research. Here, we have seen that with an appropriate adaptive design, GASP models can give good fit to even non-stationary looking surfaces.

References

Bates, R. A., Buck, R. J., Riccomagno, E., and Wynn, H. P. (1996). Experimental design and observation for large systems. (Disc: p.95-111), *Journal of the Royal Statistical Society* **B 58**, 77-94.

Branin, F. H. (1972). Widely convergent methods for finding multiple solutions of simultaneous nonlinear equations. *IBM Journal of Research Developments* **16**, 50-522.

Cressie, N. C. (1993), *Statistics for spatial data*. Wiley, New York.

Currin, C., Mitchell, T., Morris, M., and Vlvisaker, 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.

Farhang-Mehr, A. and Azarm, S. (2005). Bayesian meta-modeling of engineering design simulations: A sequential approach with adaptation to irregularities in the response behavior. *International Journal for Numerical Methods in Engineering* **62**, 2104-2126.

Gramacy, R. B. (2005). *Bayesian Treed Gaussian Process Models*. PhD thesis, University of California, Santa Cruz, CA 95064. Department of Applied Math. & Statistics.

Jin, R., Chen, W., and Sudjianto, A. (2002). On sequential sampling for global metamodeling in engineering desing. In Proceedings

of DETC 2002. ASME 2002 Design Engineering Technical Conferences and computers and information in Engineering Conference.

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

Koehler, J. R. and Owen, A. B. (1996). Computer experiments. In Ghosh, S. and Rao, C. R., editors. *Handbook of Statistics* Volume **13** pages 261-308. Elsevier Science, New York.

Marin, O. (2005). *Designing Computer Experiments to Estimate Integrated Response functions*. PhD thesis, The Ohio State University.

McKay, M. D., Beckman, R. J., and Conover, W. 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.

Mitchell, T., Morris, M., and Ylvisaker, D. (1990). Existence of smoothed stationary processes on an interval. *Stochastic processes and their applications* **35**, 109-119.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments (with comments, p.423-435). *Statistical Science* **4**, 409-423.

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

Schonlau, M. (1997). *Computer Experiments and Global Optimization*. PhD thesis, University of Waterloo.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *Journal of applied statistics* **14,** 165-170.

Chen Quin Lam
Department of Statistics
The Ohio State University
1958 Neil Avenue
404 Cockins Hall
Columbus, OH 43210-1247

William I. Notz
Department of Statistics
The Ohio State University
1958 Neil Avenue
404 Cockins Hall
Columbus, OH 43210-1247
Email:win@stat.osu.edu

Appendix

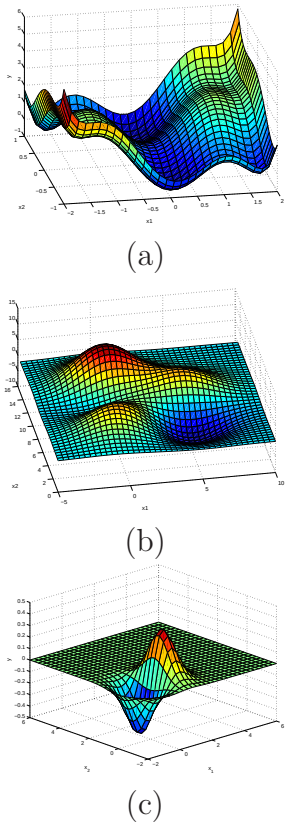
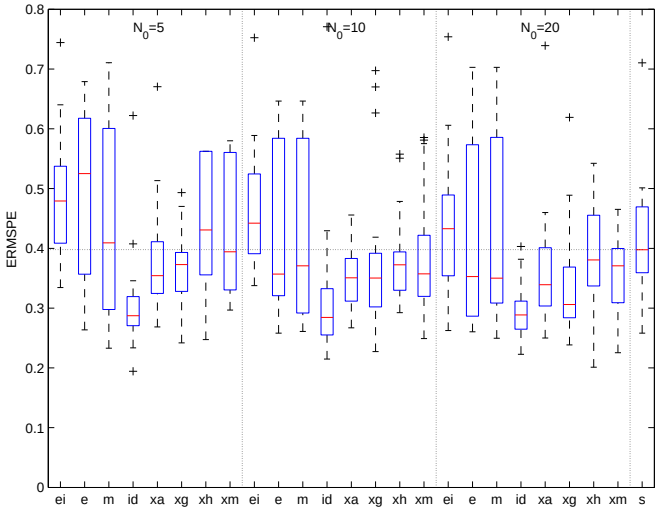
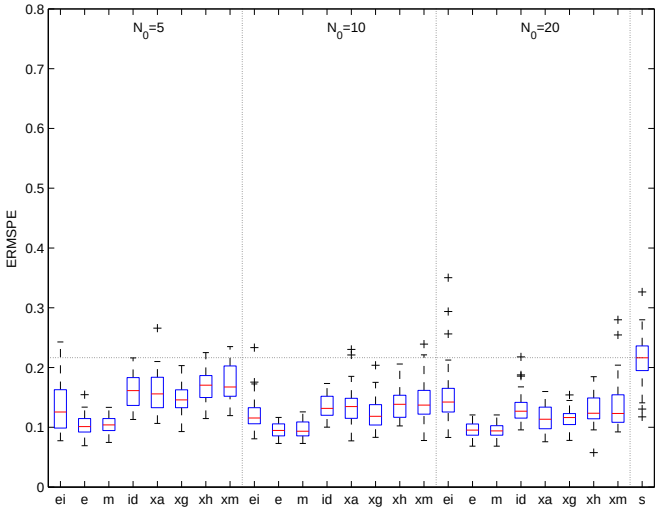


Figure 1: Surface plots of the true surfaces. (a) Function 1: Six-hump camel-back function (b) Function 2: Simulated surface, (c) Function 3: Two-dimensional exponential function

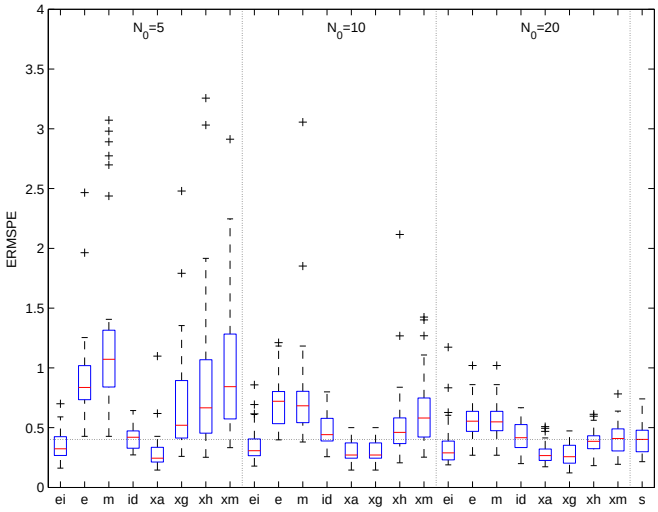


Gaussian Correlation

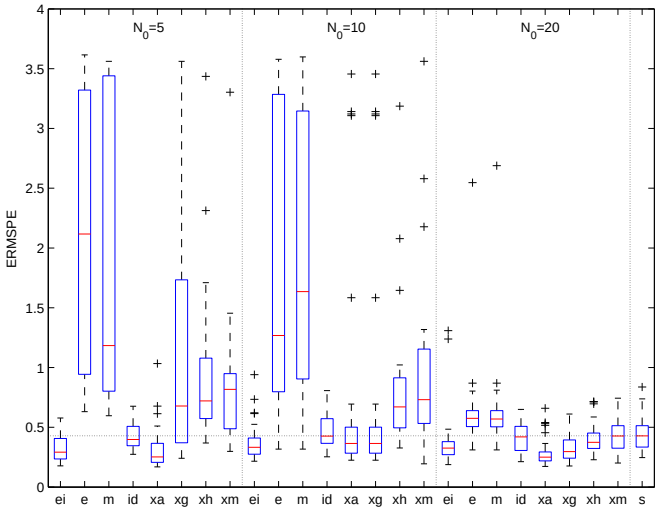


Cubic Correlation

Figure 2: (Function 1, Six-hump camel-back function) Boxplots of the empirical root mean squared prediction error (ERMSPE) for 30 different starting designs and Gaussian/cubic correlation functions - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross validation prediction error (using arithmetic mean (xa), geometric mean (xg), harmonic mean (xh), maximin criteria (xm)), fixed-point maximin LHD (s). N_0 denotes the number of starting input points for the sequential methods. A total of $N = 40$ design points are selected in all the cases.

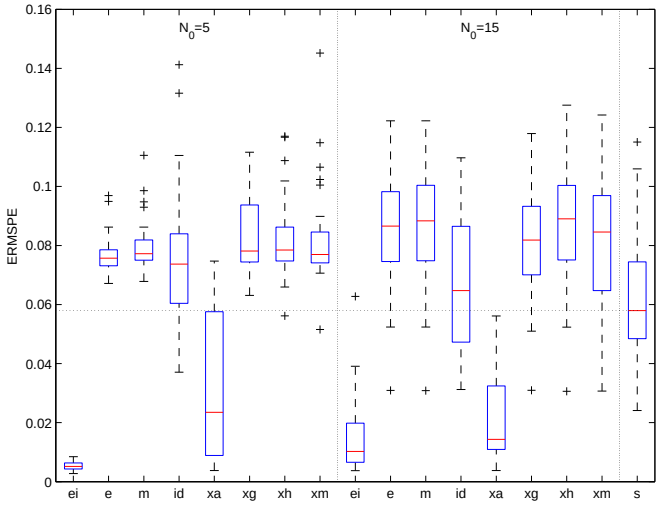


Gaussian Correlation

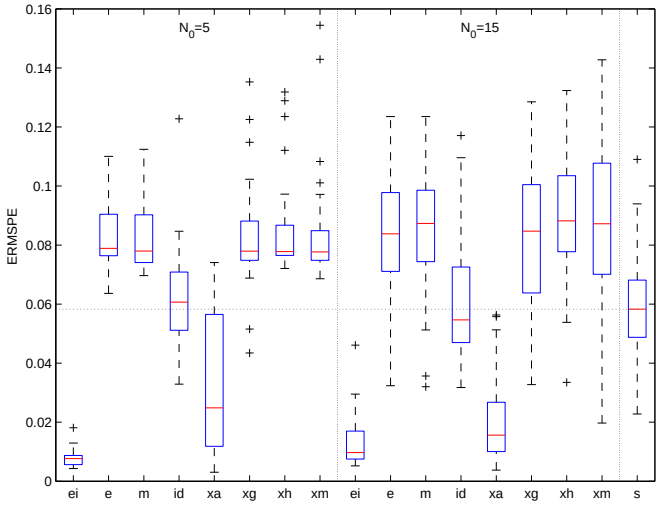


Cubic Correlation

Figure 3: (Function 2, Simulated surface) Boxplots of the empirical root mean squared prediction error (ERMSPE) for 30 different starting designs and Gaussian/cubic correlation functions - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross validation prediction error (using arithmetic mean (xa), geometric mean (xg), harmonic mean (xh), maximin criteria (xm)), fixed-point maximin LHD (s). N_0 denotes the number of starting input points for the sequential methods. A total of $N = 40$ design points are selected in all the cases.



Gaussian Correlation



Cubic Correlation

Figure 4: (Function 3, Two-dimensional exponential function) Boxplots of the empirical root mean squared prediction error (ERMSPE) for 30 different starting designs and Gaussian/cubic correlation functions - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross validation prediction error (using arithmetic mean (xa), geometric mean (xg), harmonic mean (xh), maximin criteria (xm)), fixed-point maximin LHD (s). N_0 denotes the number of starting input points for the sequential methods. A total of $N = 30$ design points are selected in all the cases.

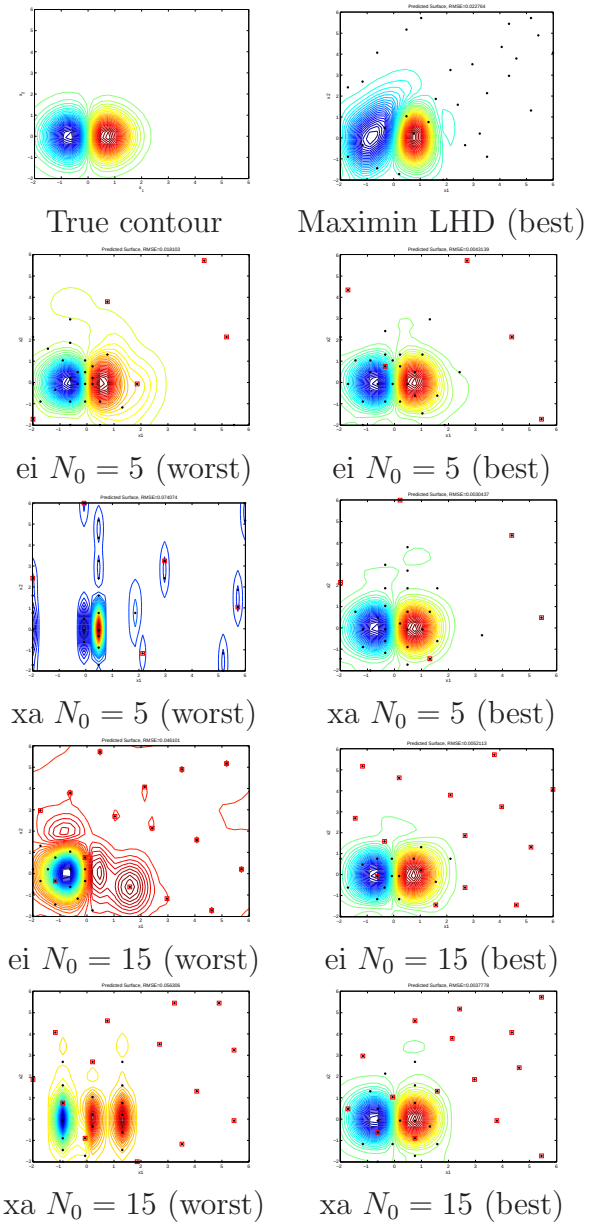


Figure 5: (Function 3, Two-dimensional exponential function) Contour plots of the true surface and predicted surfaces using the fixed-point maximin LHD (s), EIGF (ei) and cross validation with the arithmetic mean (xa) criteria. The plots show the best and worst predicted surfaces based on ERMSPE among the 30 runs. N_0 denotes the number of starting input points for the sequential methods. The red squares denote the location of the inital starting design points and the black dots denote the remaining added points.