SEQUENTIAL ADAPTIVE DESIGNS IN COMPUTER EXPERIMENTS FOR RESPONSE SURFACE MODEL FIT

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Chen Quin Lam, M.S., M.Sc.

* * * * *

The Ohio State University

2008

Dissertation Committee:

Dr. William Notz, Adviser

Dr. Thomas Santner

Dr. Angela Dean

Approved by

____

Adviser
Graduate Program in
Statistics

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 thesis, we consider sequential adaptive designs as an "efficient" alternative to fixed-point 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 shown to perform well.

The issue of robustness for the proposed sequential adaptive designs is also addressed in this thesis. While we find that sequential adaptive designs are potentially more effective and efficient than fixed-point designs, issues such as numerical instability do arise. We address these concerns and also propose a diagnostic tool based on cross validation prediction error to improve the performance of sequential designs.

We are also interested in the design of computer experiments where there are control variables and environmental (noise) variables. We extend the implementation of the proposed sequential designs to achieve a good fit of the unknown integrated response surface (i.e., the averaged response surface taken over the distributions of the environmental variables) using output from the simulations. The goal is to find an optimal choice of the control variables while taking into account the distributions of the noise variables.

Dedicated to my family back home, my wife *Serena* and baby boy *Ernest*.

ACKNOWLEDGMENTS

I would like to begin by saying a big thank you to my advisor, Dr. William Notz, for the guidance, support and inspiration that he has given me throughout the entire process of research experience and completion of this dissertation.

I am very fortunate to have come halfway around the world to this department to experience the warmth of the everyone in the department and the intellectually-stimulating learning environment. My gratitude goes out to the faculty members in the department, in particular, Dr. Tom Santner and Dr. Angela Dean for being in my defence committee and organizing the wonderful sessions of computer experiment journal club discussions over the past years. It has been an eye-opening experience for me to be involved in various research projects and discussion groups, and I would like to thank Drs. Mark Berliner, Kate Calder, Tao Shi and Steve MacEachern. Last but definitely not least, I am grateful to Dr. Douglas Wolfe and Dr. Elizabeth Stasny for their dedication and support to providing a great education for me and all other graduate students in the department!

This entire episode of my life in Columbus would not be possible without the unconditional support and love from my family in Singapore. And, a big thanks to Ben Kan for his friendship. Most importantly, my wife Serena and baby boy Ernest have been life and soul of my world here. I cannot imagine being here without them! Love you always.

VITA

July 5, 1973	Born - Singapore
1997	B.A. Economics and Statistics, National University of Singapore
1998	B.S.S. Statistics, National University of Singapore
2000	M.Sc. Statistics, National University of Singapore
2003	M.S. Statistics, The Ohio State University
2000-present	Graduate Research Assistant and Teaching Assistant, Department of Statistics, The Ohio State University

PUBLICATIONS

Lam, C.Q. and Notz, W.I. (2008), *Sequential Adaptive Designs in Computer Experiments for Response Surface Model Fit.* Submitted to Journal of Statistics and Applications.

Munroe, D.K., Calder, C.A., Shi, T., Xiao, N., Lam, C.Q., Li, D., Wolfinbarger, S.R. (2007), *The relationships between biomass burning, land-cover/use change, and the distribution of carbonaceous aerosols in mainland Southeast Asia: A review and synthesis.* Department of Statistics Preprint No. 793, The Ohio State University. Submitted to Journal of Land Use Science.

Berliner, L.M., Jezek, K., Cressie, N., Kim, Y., Lam, C.Q. and van der Veen, C.J. (2007), *Modeling Dynamic Controls on Ice Streams: A Bayesian Statistical Approach.* In revision for Journal of Glaciology, October.

Berliner, L.M., Cressie, N., Jezek, K., Kim, Y., Lam, C.Q., and van der Veen, C.J. (2007). *Equilibrium Dynamics of Ice Streams: A Bayesian Statistical Analysis.* Statistical Methods and Applications (in press), November.

Berliner, L. M., Cressie, N., Jezek, K., van der Veen, C.J., Kim, Y. and Lam, C.Q. (2005). *Hierarchical Bayesian Modeling of the Movement of Ice Streams*, in Statistical Solutions to Modern Problems: Preceedings of the 20th International Workshop on Statistical Modelling, Sydney, Australia, July 10-15.

Lam, C.Q. and Stasny, E.(2003), *Handling Undecided Voters: Using Missing Data Methods in Election Forecasting*, Technical Report No. 757, Department of Statistics, The Ohio State University.

FIELDS OF STUDY

Major Field: Statistics

TABLE OF CONTENTS

Appendices:

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

ANALYSIS OF COMPUTER EXPERIMENTS

1.1 Introduction to Computer Experiments

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. Computer simulations have been run to provide a representation of the "real" physical system. Put in a simplistic way, these simulations are attempts to represent the complex reality by means of 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. This has led to the development of statistical methodologies for predicting the unobserved responses at selected input points. The approach taken in this thesis assumes that the response can be modeled as a realization of a Gaussian stochastic process (see Sacks *et al.*, 1989, and Chapter 2 in Santner *et al.*, 2003).

While both computer and physical experiments are attempts to collect data for modeling the relationship between inputs and the response variable, it is important to note that we are now dealing with two distinctive sets of observations and hence an appropriate set of design strategies and analysis methods has to be developed.

While random error is inherent in physical experiments, computer experiments are generally deterministic (i.e., repeated runs of the computer code will give the same output). Design and analysis of physical experiments must account for sources of random variability by taking multiple observations at the same site, blocking, and randomization of assignment. Popular choices of designs include factorial, orthogonal, randomized block and "optimal" designs ("optimal" designs are constructed according to some specific objectives such as minimizing the trace or determinant of the covariance matrix). However, since the output from the computer code is deterministic, the notion of "uncertainty" in computer experiments is fundamentally different. Issues such as replication, blocking and randomization, and other principles for controlling bias and noise in physical experiments are no longer relevant. Underlying the design and analysis of computer experiments is a computer code (also called a simulator) representing the relationship between the inputs and the response variable. Uncertainty arises because the modeler may not have full acesss to the simulator. This may occur either when the simulator is proprietary and not made available to modelers, or the simulator is highly computer intensive and it is not possible to run it at all inputs. As a result, computer experiments are attempts to determine the details of this input-response relationship using an approximate stochastic model with limited observations.

Due to this distinctive feature of computer experiments, the basic idea behind experimental designs is to select input points that will allow us to model and minimize the discrepancy between the output from the the computer code and predictions from the stochastic model. Without the presence of random errors, experimental designs for computer experiment should not take multiple samples at the same input point. And, experimental designs should also be space-filling to enable an extensive exploration of the relatively unknown shape of the response function. Designs based on optimality criteria are also available for computer experiments (for example, the integrated mean squared prediction error) and are generalizations of the corresponding class of designs in physical experiments.

Physical experiments and computer experiments also differ in terms of the types of input variables and how they are dealt with. The three types of variables are the *control* variable, *environmental* variable and *model* variable (as described in Santner *et al.*, 2003, Chapter 2). The *control* and *environmental* variables are present in both physical experiments and computer experiments - *control* variables are settings that can be controlled by the experimenter, and *environmental* variables can be thought of as noise variables. Even though physical experiments may have very few observations, useful inference is still possible when combined with the outputs from computer experiments. In cases where physical observations are available and the the response from the computer code is directly used in the analysis, the third type of *model* variable may arise due to some unknown constants or parameters in the simulator. The objective is then to calibrate these model variables with the physical observations so that the model provides a more accurate representation of reality.

Modeling and parameter estimation are also very different for the two types of experiments. For physical experiments, response surface modeling is a popular approach. Typically, a polynomial model with i.i.d. noise is fitted and least squares estimates are obtained. In computer experiments, the usual practice is to fit a simple constant mean model with correlated errors. Departure from the mean is captured through some parametric class of correlation functions. Estimation of parameters is usually by maximum likelihood, restricted maximum likelihood or cross validation.

Model validation is also very different in both types of experiments. Without taking additonal samples, cross validation is a popular method in computer experiments while other summaries such as F test statistics, R-square values are typically used in the presence of random errors in physical experiments.

To date, there have been applications of computer experiments in many fields, for instance:

1. Global response surface model fit and global optimization using the Gaussian stochastic process model : For global model fit, Sacks *et al.* (1989) was among the first to propose the Gaussian process model, while Currin *et al.* (1991) looked at applications involving predictions at unobserved inputs for electrical circuits and thermal storage systems. More recent work by Dirgnei (2006) studied an approximation to a complex ocean model and proposed a two-stage method for multivariate outputs. For optimization problems, Jones *et al.* (1998) looked at finding the gobal optimum of response surfaces. Ranjan (2007) looked at contour estimation, while Williams *et al.* (2000) proposed an algorithm for global optimization in the presence of both *control* and *environmental* variables.

2. Sensitivity analysis: Oakley and O'Hagan (2004) explored how inputs may affect the uncertainty in the response and how to screen out input variables that have little effect on the response.

3. Calibration and predictions: Kennedy and O'Hagan (2001) first proposed the idea of fusing output from the computer code and physical observations using a Bayesian framework, and Higdon *et al.* (2003) developed a fully Bayesian model, using Markov Chain Monte Carlo methods, to characterize the uncertainty in predictions made from computer experiments.

1.2 Outline of Thesis

This thesis provides a review of the design and analysis of computer experiments for constructing (parsimonious) surrogate models to replace the actual complex computer code. These surrogate models may be used for several purposes, such as global response surface model fit, global optimization, contour estimation and integration etc.

The focus of this thesis 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 Gaussian stochastic process model (GASP) model which is used as an approximation to the actual computer code. We propose several sequential adaptive designs and compare them against one another and also against a fixed-point design.

An overview of the stochastic process model and estimation methods used will be given in the next two sections. Chapter 2 provides an overview of the experimental designs commonly used in computer experiments. These include various space-filling designs and (statistical) criterion-based designs. Other designs for global model fit and global optimization will also be introduced.

Following this in Chapter 3, we present several sequential adaptive designs proposed in this thesis and carry out an empirical study to examine the performances of these designs in terms of how well the selected input points lead to an accurate predictive GASP model.

Chapter 4 gives a brief review of two studies to highlight some recent work in the area of design and analysis of computer experiments for non-stationary looking response surfaces. A new sequential adaptive design criterion is proposed to address this issue of fitting a single stationary GASP model across the entire input space of a non-stationary looking response surface.

Chapter 5 highlights potential problems with sequential designs in this thesis and presents the use of sequential diagnostic tools within the proposed sequential design algorithms to improve their predictive performances using the GASP model.

Chapter 6 considers the situation where both control and environmental variables are present and explores the use of sequential adaptive designs for integrated response surfaces model fit.

We conclude, in Chapter 7, with a discussion of the proposed design criteria and future research.

1.3 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 thesis, 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 1.3. For our approach, the best linear unbiased predictor is used to predict the response at unobserved $\boldsymbol{x}$, based on the available training data.

1.3.1 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

$$Y(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{1.1}$$

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 (1.1) 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)

$$MSPE[\hat{Y}(\boldsymbol{x}_0)] = E[(c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n - Y(\boldsymbol{x}_0))^2] \tag{1.2}$$

subject to the unbiasedness constraint $E[c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n] = E[Y(\boldsymbol{x}_0)]$ which can be re-expressed as

$$c^T(\boldsymbol{x}_0)\boldsymbol{F}\boldsymbol{\beta} = \boldsymbol{f}^T(\boldsymbol{x}_0)\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \Re^k \quad \text{or} \quad \boldsymbol{F}^T c(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0) \tag{1.3}$$

where $\boldsymbol{F} = [f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)]^T$ is the $n \times k$ matrix of regressors whose $(i,j)th$ element is $f_j(x_i)$ for $1 \leq i \leq n$, $1 \leq j \leq k$. Minimizing (1.2) is then equivalent to minimizing

$$Var(c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n - Y(\boldsymbol{x}_0)). \tag{1.4}$$

and leads to minimizing

$$MSPE[\hat{Y}(\boldsymbol{x}_0)] = \sigma^2[1 + c^T(\boldsymbol{x}_0)\boldsymbol{R}c(\boldsymbol{x}_0) - 2c^T(\boldsymbol{x}_0)\boldsymbol{r}(\boldsymbol{x}_0)]. \tag{1.5}$$

Next, a $k \times 1$ vector of Langrange multipliers ($\boldsymbol{\lambda}$) is introduced and taking the derivative of

$$\frac{MSPE[\hat{Y}(\boldsymbol{x}_0)]}{\sigma^2} - 2\boldsymbol{\lambda}^T(\boldsymbol{x}_0)[f(\boldsymbol{x}_0) - \boldsymbol{F}^T c(\boldsymbol{x}_0)]$$

with respect to $\boldsymbol{\lambda}$ and $c(\boldsymbol{x}_0)$, yields (1.3) and $\sigma^2 Rc(\boldsymbol{x}_0) - \sigma^2 r(\boldsymbol{x}_0) - F\boldsymbol{\lambda} = 0$. This system of equations can be expressed in matrix form

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{F}^T \\ \boldsymbol{F} & \boldsymbol{R} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}_0) \\ c(\boldsymbol{x}_0) \end{pmatrix} = \begin{pmatrix} f(\boldsymbol{x}_0) \\ r(\boldsymbol{x}_0) \end{pmatrix}. \tag{1.6}$$

Assuming that $\boldsymbol{F}$ and $\boldsymbol{R}$ are of full column rank, the solution for $c(\boldsymbol{x}_0)$ in (1.6) is substituted into (1.3) to give the associated BLUP

$$\hat{Y}(\boldsymbol{x}_0) = c^T(\boldsymbol{x}_0)\boldsymbol{Y}^n = \boldsymbol{f}^T(\boldsymbol{x}_0)\hat{\boldsymbol{\beta}} + r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}\hat{\boldsymbol{\beta}}), \tag{1.7}$$

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 $\boldsymbol{\beta}$. Similarly, the MSPE of the BLUP (1.5) is then given by

$$\begin{aligned} MSPE[\hat{Y}(\boldsymbol{x_0})] &= \sigma^2[1 - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x_0}) + \qquad (1.8)\\ (f^T(\boldsymbol{x_0}) &- r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}(f^T(\boldsymbol{x_0}) - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})^T], \end{aligned}$$

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 (1.1) 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 if the response surface is not too highly non-stationary. The stochastic process $Z(\boldsymbol{x})$ captures the local trend which usually suffices to produce excellent fit.

Given that the correlation function $R(\cdot)$ is known, the BLUP can be easily calculated using (1.7). Typically, the correlation parameters have to be estimated (for example, by maximum likelihood estimation) and the resulting predictor is termed the *empirical best linear unbiased predictor* (EBLUP).

It is noted that the BLUP in (1.7) is an interpolating predictor as follows. Using the fact that $\boldsymbol{R}^{-1}\boldsymbol{R} = \boldsymbol{I}_n$ (identity matrix) and supposing $\boldsymbol{x}_0$ is one of the training points (say $\boldsymbol{x}_i$), then $\boldsymbol{R}^{-1}r(\boldsymbol{x}_0)$ is a unit vector with 1 in the i^{th} position and 0 elsewhere. As a result, (1.7) reduces to $\hat{Y}(\boldsymbol{x}_0) = f^T(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}} + (Y(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}}) = Y(\boldsymbol{x}_i)$.

1.3.2 Bayesian Approach.

Currin *et al.* (1991) and Koehler and Owen (1996) presented examples using the alternative Bayesian approach to Subsection 1.3. The simplest case is when the parameters $R(.)$, $\boldsymbol{\beta}$ and σ^2 are known and fixed. From the joint distribution,

$$\begin{pmatrix} \boldsymbol{Y}^n \\ Y(\boldsymbol{x}_0) \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{F} \\ f^T(\boldsymbol{x}_0) \end{pmatrix} \boldsymbol{\beta}, \sigma^2 \begin{pmatrix} \boldsymbol{R} & r(\boldsymbol{x}_0) \\ r(\boldsymbol{x}_0) & \mathbf{1} \end{pmatrix} \right], \tag{1.9}$$

we obtain the conditional posterior distribution as

$$[Y(\boldsymbol{x}_0)|\boldsymbol{Y}^n] \sim N\Big(f^T(x_0)\boldsymbol{\beta}+r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}(\boldsymbol{Y}^n-\boldsymbol{F}\boldsymbol{\beta}),\ \sigma^2[1-r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}r(\boldsymbol{x}_0)]\Big). \tag{1.10}$$

Notice that the point predictor (the posterior mean) is the same as the BLUP in (1.7), while the posterior predictive variance is different from (1.8) because the estimation of $\boldsymbol{\beta}$ is ignored. Currin *et al.* (1991) adopted an empirical Bayesian approach and estimated the parameters (and hyperparameters) via maximum likelihood.

A fully Bayesian approach is to assign prior distributions to the parameters and integrate out these parameters to obtain the posterior distribution (1.10). An example of an informative prior is a normal prior, such as $\boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \tau^2\boldsymbol{V})$. Alternatively, one choice of non-informative prior is $[\boldsymbol{\beta}] \propto 1$ with $R(.)$ and σ^2 known, and the posterior distribution $[Y(\boldsymbol{x}_0)|\boldsymbol{Y}^n]$ is Gaussian with mean and variance given by the familiar form in (1.7) and (1.8) respectively.

An extensive study of assigning different priors and the corresponding analytical form of the posterior distribution can be found in Chapter 4 of Santner *et al.* (2003) and Berger *et al.* (2001).

1.3.3 Parametric Correlation Functions.

As seen from the equations (1.7) and (1.8) above, the correlation function $R(\cdot)$ 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 $R(\cdot)$. We consider correlation functions for $x_1, x_2 \in S$

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

so that $Z(\cdot)$ in (1.1) is stationary. A valid stationary correlation function must satisfy the following conditions: (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 $w_1, \ldots, w_n$, and (iii) $R(d) = R(-d)$ and does not depend on the location. In higher dimensions (i.e. two or higher), taking the products of correlation across each dimension $j = 1, \cdots, m$ is a common practice for computational convenience,

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

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. A third choice, the Matern correlation function, is sometimes used too but requires more computation time to estimate.

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

$$\begin{aligned} R(d) &= 1 - 6\left(\tfrac{d}{2}\right)^2 + 6\left(\tfrac{|d|}{\theta}\right)^3, && |d| < \tfrac{\theta}{2} \\ &= 2\left(1 - \tfrac{|d|}{\theta}\right)^3, && \tfrac{\theta}{2} \leq |d| < \theta \\ &= 0, && |d| \geq \theta \end{aligned} \tag{1.11}$$

where $\theta > 0$ and d denotes the distance between two points (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 (1.7) is a piecewise cubic spline interpolating predictor in the context of computer experiment.

Power Exponential Correlation. Another very popular correlation function takes the form of

$$R(d) = exp(-\theta|d|^p), \tag{1.12}$$

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 (and BLUP) that is infinitely differentiable. Taking $p = 1$ gives the exponential correlation function. For $0 < p < 2$, the BLUP and EBLUP are continuous but not differentiable. As θ increases, the dependence between the response at two input points 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, then $p = 2$ should be used.

Both the cubic and power exponential (with $p = 2$) correlation functions will be used for the examples in later chapters. We will also use the product correlation structure, $R(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod\limits_{j=1}^{m} R(|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}| \,\big|\theta_j)$, and let $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$.

Matérn Correlation. The Matérn class correlation was proposed in Matérn (1960).

$$R(d) \;=\; \frac{1}{\Gamma(\nu)2^{\nu-1}}\Big(\frac{2\sqrt{\nu}|d|^{\nu}}{\theta}\Big)\Big(K_{\nu}\frac{2\sqrt{\nu}|d|}{\theta}\Big) \tag{1.13}$$

where $\nu > 0$ is the smoothness parameter (controls the amount of differentibility), and $\theta > 0$ is the spatial scale parameter (controls the range of the correlation). The function K_ν is the modified Bessel function of the third kind of order ν (see Stein, 1999 Chapter 6). The parameter ν controls the smoothness of the process, while θ controls the range of correlation in each dimension. $R(0)$ is defined to be 1. The effect of a change in ν can be clearly seen in a change in the smoothness of the random function generated with the Matérn correlation function (see pages 41-45 in

Santner *et al.*, 2003, for plots of random functions generated using various correlation functions). For computing purposes when $d = 0$, d is set to a very small value (e.g., $d = 1e - 10$).

As special cases, the power exponential correlation function (1.12) in p dimensions with $\alpha_1 = ... = \alpha_p = 2$ and $1/\theta^2$ in place of θ is the limiting case of the Matern correlation function (1.13) as $\nu \rightarrow \infty$.

1.4 Parameter Estimation

The previous sections presented a few approaches to the prediction problem and so far the correlation parameters in the models have been assumed to be known. We consider two estimation methods: maximum likelihood and cross validation.

1.4.1 Maximum Likelihood Estimation

Assuming the stochastic model in (1.1) where $Z(\cdot)$ has a Gaussian distribution, the log likelihood for up to an additive constant is

$$l(\boldsymbol{\beta}, \sigma_z^2, \boldsymbol{\theta}|\boldsymbol{Y}^n) \ = \ -\frac{1}{2}[n\log\sigma_z^2 + log|\boldsymbol{R}| \ + \ (\boldsymbol{Y}^n - \boldsymbol{F\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F\beta})/\sigma_z^2]. \quad (1.14)$$

Given $\boldsymbol{\theta}$, the maximum likelihood estimates (MLE) of $\boldsymbol{\beta}$ and σ_z^2 are given by

$$\hat{\boldsymbol{\beta}} \ = \ (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{Y}^n \quad (1.15)$$

and

$$\hat{\sigma}_z^2 \ = \ \frac{1}{n}(\boldsymbol{Y}^n - \boldsymbol{F}\hat{\boldsymbol{\beta}})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}\hat{\boldsymbol{\beta}}). \quad (1.16)$$

After substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_z^2$ back into (1.14) and some cancellation of terms, we obtain the MLEs of $\boldsymbol{\theta}$ by maximizing $-\frac{1}{2}$ [$n\log\hat{\sigma}_z^2 \ + \ \log|\boldsymbol{R}|$] numerically.

1.4.2 Cross validation

Cross validation is an alternative method (prediction oriented) for estimating model parameters in parametric model settings. Here, we consider the leave-one-out approach in estimating the correlation parameters in Subsection 1.3.3. The basic idea is based on leaving the i^{th} observation out and predicting the value of the i^{th} observation (using the BLUP as shown in (1.7)) based on the remaining $(n-1)$ observations. Let $\boldsymbol{\psi}$ denotes the vector of unknown correlation parameters. The cross validation estimator of $\boldsymbol{\psi}$ is found by minimizing the cross validation prediction error $(XVPE_c)$,

$$XVPE_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} (\hat{Y}^{(-i)}(\boldsymbol{\psi}, \boldsymbol{x}) - y(\boldsymbol{x}_i))^2, \tag{1.17}$$

where $\hat{Y}^{(-i)}(\boldsymbol{\psi}, \boldsymbol{x})$ denotes the BLUP of $y(\boldsymbol{x})$ based on all observations except $\{\boldsymbol{x}_i, y(\boldsymbol{x}_i)\}$ where $i = 1, ..., n$ sampled points, and $y(\boldsymbol{x}_i)$ is the computer output at $\boldsymbol{x}_i$.

1.5 Cross Validation for Model Validation

Cross validation is also a popular method to assess the fit of the estimate (BLUP/ EBLUP) of the stochastic model in Section 1.3. This is an easy and practical approach since additional observations are not needed (see Morris and Mitchell, 1995). Cross-validated predictions and corresponding prediction errors are computed at each of the training sites using a subset of the existing dataset of n observations. Using the leave-one-out cross validation approach in Chapter 5, we define the *cross validated prediction error* $(XVPE_f)$ as

$$XVPE_f(\boldsymbol{x}_i) \;=\; [\hat{Y}^{(-i)}(\boldsymbol{x}) - y(\boldsymbol{x}_i)]^2, \tag{1.18}$$

where $\hat{Y}^{(-i)}(\boldsymbol{x})$ denotes the BLUP of $y(\boldsymbol{x})$ based on all observations except $\{\boldsymbol{x}_i, y(\boldsymbol{x}_i)\}$ where $i = 1, ..., n$ are the sampled points, and $y(\boldsymbol{x}_i)$ is the computer output at $\boldsymbol{x}_i$.

An assessment of the fit of the model can be performed by comparing the values of the predicted $\hat{Y}^{(-i)}(\boldsymbol{x})$ against $y(\boldsymbol{x}_i)$. An application of using the $XVPE_f$ is given in Section 5.1.

1.6 Cross Validation for Prediction Error Assessment

Besides using cross validation for estimating model parameters (Subsection 1.4.2) and for model validation (Section 1.5), one can also 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 (1.8). In turn, this prediction error will be used as a design criterion to select additional input points.

The main idea of the cross validation approach is as follows. Based on n observations, first estimate the correlation parameters $\boldsymbol{\theta}$ by MLE or any other method. Suppose $\boldsymbol{x}$ is one of the candidate points to be considered. For each subset of the $n-1$ observations, predict y at $\boldsymbol{x}$ using the EBLUP version of (1.7). The n predictions at $\boldsymbol{x}$ are used to provide an estimate of the prediction error at $\boldsymbol{x}$ instead of the MSPE (1.8). Subsection 3.2.2 describes how the prediction error is quantified.

CHAPTER 2

EXPERIMENTAL DESIGNS FOR COMPUTER EXPERIMENTS

Experimental designs for computer experiments have received a great deal of attention given the increased use of computer simulation models in scientific research. 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.

It is important to note that design strategies for computer experiments differ from traditional physical experiments in that: (i) Designs should not take more than one observation at any input point due to the fact that the computer output is deterministic. The same input yields the same response in repeated runs, (ii) Designs should be flexible enough so that the responses at the sampled points provide sufficient information on the functional form of the response surface across the entire input space. This is usually taken to mean designs should be space-filling. The rationale is that the functional relationship between the inputs and output is typically unknown, and hence it is not clear where to search for features of interest, such as local/global optimum, flat contours, subregions where the responses vary significantly etc., in the input space.

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 is not adaptive to what we learn about the response surface as we observe the code, and space-filling designs may result in poor prediction accuracy and efficiency in many situations. An overview of these designs is provided in Section 2.1. 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. More details are given in Section 2.2.

Designs for computer experiments have almost been exclusively restricted to 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 whether these designs will result in an accurate predictive model, because they also make no direct use of what one learns from the observed responses about the form of the response surface. We describe these sequential optimality-based designs in Section 2.3. This will be investigated further in the empirical study in Section 3.4.

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 designs are sequential (e.g., Sobol' sequences) but not adaptive. Several sequential, as well as adaptive, designs based on cross validation will be reviewed in Section 2.4. Other designs for global optimization will also be introduced in Section 2.5.

2.1 Space-Filling Designs

Space-filling designs are intuitively appealing in that observations are spread out over the entire range of the input space to minimize the prediction error of the GASP model. LHD, distance-based designs (such as maximin, minimax, etc.) and uniform designs are examples of such space-filling designs. There exists a number of space-filling design criteria as mentioned in the introduction of this chapter but studies (e.g., Marin, 2005) suggest they perform similarly in terms of prediction accuracy.

2.1.1 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 out over the range of each input separately.

Suppose we want to generate an n-point LHD based on input points from $\boldsymbol{X} = (X_1, ..., X_d)$ where d denotes the number of dimensions. Assuming independence of each component (i.e. $X_k \overset{iid}{\sim} F_k$ where $k = 1, \cdots, d$), the range of each X_k is divided into n equal-probablity strata which are labeled $\{1, \cdots, n\}$. This creates a total of n^d equal-space partitions in the d-dimension input space. A Latin hypercube sample of size n cells is sampled from the entire set of n^d cells.

The actual construction of a LHD proceeds as follows. Assuming the input region of $\boldsymbol{X}$ is distributed over $[0, 1]^d$, let $\boldsymbol{\Pi} = \{\Pi_{jk}\}$ be an $n \times d$ matrix ($j = 1, \cdots, n$ and $k = 1, \cdots, d$) with columns of d different randomly selected permutations of $\{1, 2, ..., n\}$. The sampled values are

$$X_{jk} \;=\; F_k^{-1}\Big(\frac{1}{n}(\Pi_{jk} - 1 + U_{jk})\Big)$$

where j denotes the j^{th} sample value, k denotes the k^{th} component (or dimension), and U_{jk} are i.i.d. uniform(0,1) random variables. The d elements in the j^{th} row of Π identify the partition (in each dimension) that X_{jk} is selected from, while the corresponding U_{jk} determine the location of X_{jk}. Alternatively, the design point may be placed in the middle of the selected cell (i.e., $U_{jk} = 0.5$, a fixed constant).

In the context of computer experiments, the distribution F_k is often taken to be uniform, i.e., $F_k^{-1}(x) = x, \quad 0 < x < 1$, since the response surface is taken to be unknown and the goal is to explore the surface over the entire region evenly.

Despite the claim of their marginal space-filling properties, not all LHDs are space-filling across the entire input space. For example, in a two-dimensional case, a non space-filling LH sample might have all its points lined up along one of the main diagonals across the 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 Chapter 5 in Santner *et al.*, 2003, Koehler and Owen, 1996, and Subsection 2.1.3). Stein (1987) discussed the case of LHD for dependent X components as an extension.

2.1.2 Distance-based Designs

Johnson, Moore and Ylvisaker (1990) proposed design criteria based on maximin and minimax distances between input points. The intuition behind these designs is to consider explicitly the distance between all pairs of points and to specify a criterion that seeks to spread points out across the input space.

Let the input space be $\mathcal{X} \subset \Re^m$ and define the distance measure for a pair of points by

$$d_p(x_1, x_2) \;=\; \left[\sum_{j=1}^{m} |x_{1j} - x_{2j}|^p\right]^{1/p}$$

where the case of $p = 2$ gives the Euclidean distance between x_1 and x_2. Defining $P_n \subset \mathcal{X}$ to be a potential n-point design, a design P_n^0 is said to be a *maximin distance design* if

$$\max_{P_n} \min_{\boldsymbol{x}, \boldsymbol{x}^T \in P_n} d(\boldsymbol{x}, \boldsymbol{x}^T) = \min_{\boldsymbol{x}, \boldsymbol{x}^T \in P_n^0} d(\boldsymbol{x}, \boldsymbol{x}^T),$$

which ensures the points are located as far apart as possible. Alternatively, P_n^0 is a *minimax distance design* if

$$\min_{P_n} \max_{\boldsymbol{x} \in \mathcal{X}} d(\boldsymbol{x}, P_n) = \max_{\boldsymbol{x} \in \mathcal{X}} d(\boldsymbol{x}, P_n^0),$$

where $d(\boldsymbol{x}, P_n) = \min_{\boldsymbol{x}_0 \in P_n} d(\boldsymbol{x}, \boldsymbol{x}_0)$. The goal is to ensure all points are not too far from one another.

2.1.3 Hybrid Latin Hypercube Designs

Attempts are made to improve LHDs by incorporating criteria, such as maximin distance, and other criteria-based designs within the class of LHDs (see Chapter 5 in Santner *et al.*, 2003, and Koehler and Owen, 1996). Noting that distance-based maximin designs (Subsection 2.1.2) tend to put points out in the boundaries of the input space, Morris and Mitchell (1995) proposed maximizing the minimum Euclidean distance between two points in the input space as a criterion within the class of LHDs, and called this a *maximin* LHD. Park (1994) studied the use of optimality criteria, such as the integrated mean squared error criterion, within the LHD class of designs. Various software can be used to generate maximin LHDs, for example, MATLAB®, ACED (algorithms for the construction of experimental designs) software by Welch (1985) and JMP® by SAS Institute.

Handcock (1991) introduced the *cascading* LHD for exploring both the local trend (i.e., the scale and smoothess parameters in the correlation function) and overall global trend in the model. The construction of this two-stage design is to, first, generate a LHD over the entire input space. A second LHD is then generated in a small region around each of the points of the first LHD. In the context of the GASP model, the global coverage of the points (in the first stage) helps to estimate the global trend component, i.e. the $\boldsymbol{\beta}$ term in (1.1), while the "clustered" sites (second stage) help to estimate the parameters of the correlation function.

2.1.4 Uniform designs

Uniform designs, that seek to spread points uniformly across the input space, were first introduced in Fang (2000). Popular measures of uniformity include the L_p or the star discrepancy measure (L_∞). The goal is to choose a set of points $\mathcal{P}_n = \{x_1, x_2, \ldots, x_n\}$ to minimize a measure of discrepancy. Let the empirical distribution be $F_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I\{x_i \leq x\}$. The *star discrepancy* of $\mathcal{P}_n$ is defined as $\mathcal{D}_\infty(\mathcal{P}_n) = \sup\limits_{x\in\mathcal{X}} |F_n(x) - F(x)|$. It measures the extent to which the set of design points $\mathcal{P}$ differs from the uniform distribution function.

An overview of methods for producing low-discrepancy sequential designs will now be provided. These designs allow points to be added sequentially and aim to spread any sub-sequence of points over the input space uniformly. They can be grouped into two classes - pseudorandom and quasirandom sequences. While pseudorandom sequences aim to produce sequences which "look" like sequences of realisations of i.i.d. uniform random variables to fill the space, quasirandom sequences aim to fill the space uniformly in a deterministic fashion. Popular sequences include the Halton, Niederreriter, Faure and Sobol' sequences etc.

Halton sequences (see Halton, 1960) are formed by reversing the digits in the representation of some sequence of integers in a given base. One way of forming these sequences is to first choose a prime number for base d. Write down the first m integers in the chosen base d. These m numbers are then "reflected" (i.e., reverse the digits) and converted back to base 10 format. To add another point, set $m = m + 1$ and continue. For more than one dimension, repeat these steps with a different base d. Even though Halton sequences perform very well in low dimensions, this uniformity property is difficult to maintain in dimensions of greater than 10.

Sobol' Sequences (see Sobol', 1993) are constructed first by determining a set of "directional" numbers $\{v_i = \frac{m_i}{2}\}$ and a choice of primitive polynomial of order d, given by $P = x^d + a_1 x^{d-1} + \cdots + a_{d-1}x + 1$ where $a_i = 0$ or 1. An initial set of integers m_i is chosen and a recursive relationship can be obtained for calculating subsequent m_i (using a binary operation). Using this algorithm, additional points can be added while maintaining the uniformity condition. This can be easily generalized to higher dimensions by repeating these steps (with distinct choices of m_i and the polynomial for each dimension).

Empirical studies to compare some of the quasirandom sequences listed above have been inconclusive in terms of the measure of discrepancy defined earlier in this subsection. The main drawback of Sobol' sequences and other quasirandom sequences is that they are not adaptive to what we learn about the response surface as we observe the code and may result in poor prediction accuracy and efficiency.

2.2 Criterion-based Optimal Designs

An alternative class of designs can be constructed based on statistical criteria. We begin by discussing two criteria based on the mean squared prediction error and the notion of entropy. Later in Section 2.3, we adapt these criteria to sequential designs. As we shall see in the next two subsections, these designs require knowledge of the unknown correlation parameters present in the GASP model (Section 1.3). One way of overcoming this problem might be to adopt a two-stage procedure: (i) estimate the parameters with observations at input points selected by a non-criterion based design, such as the LHD or maxmin design, through a small pilot study or knowledge from previous studies. (ii) use the estimated correlation parameters (and treating them as known) to subsequently select additional design points using a criterion-based strategy (see Sacks *et al.*, 1989).

Alternatively, some studies have used a fixed value for the correlation parameters for design purposes. Currin *et al.*(1991) used the exponential correlation function while advocating weak correlation strengths between the responses at different input points. For example, they used $e^{-\theta} = 0.0001$ in the study. Similarly, Mitchell and Morris (1992) favored using weak correlation parameters in the initial design phase although this might result in numerical difficulties in some cases. However, Sacks *et al.* (1989) and Lim *et al.* (2002) recommended using strong correlation strengths in the design criterion. In contrast, we choose to estimate the correlation parameters in this thesis and more details will be provided in Subsection 3.4.

2.2.1 Mean Squared Prediction Error Designs

Designs for the GASP model in (1.1) should spread out the input points across the entire input space to minimize the overall prediction error. The MSPE of the BLUP (1.8) is a measure of the prediction uncertainty of the GASP model and can be used as a design criterion. Recall in (1.8) that the MSPE is given by

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

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 $Y(\boldsymbol{x}_0)$ and observations at the previously sampled points, $\boldsymbol{Y}^n$. In practice, use of only a constant mean term, $\boldsymbol{\beta}$, in (1.1) has been found to produce an accurate predictive GASP model and leads to a simplified formula for the MSPE given by

$$MSPE[\hat{Y}(\boldsymbol{x}_0)] \;=\; \sigma^2 \left[1 - \boldsymbol{r}^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x_0}) + \frac{(1 - 1^T\boldsymbol{R}^{-1}r(\boldsymbol{x}_0))^2}{1^T\boldsymbol{R}^{-1}1}\right]. \qquad (2.1)$$

Box and Draper (1959) proposed the normalized integrated mean squared error (IMSE) as a design criterion for $\hat{Y}(\boldsymbol{x})$ over the $[0,1]^p$ domain. Sacks *et al.* (1989) considered the *integrated mean squared prediction error* (IMSPE) criterion

$$\int_{\mathcal{X}} \frac{\text{MSPE}[\hat{Y}(\boldsymbol{x})]}{\sigma^2}\, w(\boldsymbol{x})d\boldsymbol{x} \qquad (2.2)$$

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 (2.2) over the set of candidate points $\mathcal{X}$.

The *maximum mean squared prediction error* (MMSPE) criterion is defined by choosing the set of design points to find

$$\max_{\boldsymbol{x}\in\mathcal{X}} \frac{\text{MSPE}[\hat{Y}(\boldsymbol{x})]}{\sigma^2}. \qquad (2.3)$$

An n-point design is said to be MMSPE-optimal if it minimizes the above quantity (i.e. the worst prediction error) over $\mathcal{X}$.

To overcome the problem of the unknown correlation parameters present in the GASP model, one could use a two-stage procedure suggested at the beginning of this section. Another approach is proposed in Sacks *et al.* (1989) where they conducted a robustness study over a range of discrete correlation values to identify reasonable values to use for the correlation parameters so that the IMSPE design performs well. Their findings suggested that strong correlation (i.e., small correlation parameters

$\boldsymbol{\theta}$ for the Gaussian correlation function) seemed to have good relative efficiency for prediction. Lim *et al.* (2002) followed up from this study and showed that the BLUP with the Gaussian correlation function can be expressed as a polynomial as θ tends to zero (i.e., increasing correlation strength). An asymptotic IMSPE criterion based on this asymptotic form of the Gaussian correlation was derived and they showed that predictions using this design are better compared to the fixed-point LHD or the IMSPE criterion using correlation parameters estimated by maximum likelihood.

Sacks and Schiller (1988) implemented the IMSPE and MMSPE criteria for a discrete input space and commented that the design criterion optimization can be computationally formidable and suggested a sequential approach to these designs. However, it is noted that the IMSPE and MMSPE criteria cannot be carried out sequentially without modification as additional design points tend to clump around existing points (see Sacks *et al.*, 1989). To overcome this problem, the authors implemented an ad hoc sequential approach by dividing the input space into various subregions. The subregion with the largest contribution to the IMSPE criterion is identified and the point with the largest contribution to the criterion in that subregion will be selected as the next input point.

2.2.2 Maximum 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. Their idea is based on the measure of information introduced in Lindley (1956) and the Shannon's Entropy (Shannon, 1948). Currin *et al.* (1991) applied this design in the context of computer experiments.

Recall under the model (1.1) in Subsection 1.3.1 that the training data has the following conditional distribution

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

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

$$\sigma_z^2\boldsymbol{R} \; + \; \tau^2\boldsymbol{F}\boldsymbol{V}_0\boldsymbol{F}^T. \tag{2.4}$$

A design D_n^0 is said to be a *maximum entropy design* if

$$E_{\boldsymbol{Y}^n}[-ln\ P(\boldsymbol{Y}^n_{D_n^0})] \; = \; \underset{D_n}{min} E_{\boldsymbol{Y}^n}[-lnP(\boldsymbol{Y}^n_{D_n})],$$

where $P(\boldsymbol{Y}^n)$ is the probability density for the responses $\boldsymbol{Y}^n$ at n sampled points. One can show (see Koehler and Owen, 1996) that the *maximum entropy* design maximizes the determinant of the observation covariance matrix in (2.4).

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

$$\max(\ det(\boldsymbol{R})\ ). \tag{2.5}$$

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

$$\max(\ \det(\boldsymbol{R})\ \det(\boldsymbol{F}^T(\boldsymbol{R})^{-1}\boldsymbol{F})\ ). \tag{2.6}$$

Like the MSPE criterion in Subsection 2.2.1, the maximum entropy criterion depends on the unknown correlation parameters. Studies that used the maximum entropy design include Currin *et al.* (1991), and Mitchell and Scott (1987). Plots of examples of maximum entropy designs can be found in Koehler and Owen (1996). Maximum entropy designs have been shown to spread points out and often on the boundaries of the input space (see Koehler and Owen, 1996). In the limiting case of extremely weak correlation structures, entropy designs tend to become maximin distance designs according to Johnson *et al.* (1990). The maximum entropy criterion has also been introduced within the class of LHDs in Mitchell and Morris (1995).

2.3 Sequential Criterion-based Optimal Designs

In this section and the subsequent two sections, we introduce several sequential design criteria that can be used with the GASP model. Recall that the criterion-based designs in the previous section are not implementable without knowledge of the unknown correlation parameters. To overcome this problem, a sequential implementation of these criteria can be considered.

2.3.1 Sequential MSPE Criterion

A design based on the MSPE criterion (2.1) can be modified into a sequential design. 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,

$$\max_{\boldsymbol{x}_0} MSPE(\boldsymbol{x}_0) \;=\; \max_{\boldsymbol{x}_0} \left(\sigma^2 \left[1 - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1} r(\boldsymbol{x_0}) + \frac{(1 - 1^T\boldsymbol{R}^{-1} r(\boldsymbol{x}_0))^2}{1^T\boldsymbol{R}^{-1}1} \right] \right). \tag{2.7}$$

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.

Various sequential strategies are reviewed in Jin *et al.* (2002) and a new approach was proposed for the GASP model (1.1), which they termed the *kriging* model, and the radial basis function method (which is esssentially a stochastic model with the mean modeled by some basis function and the error term as *i.i.d.* noise). For the kriging model, the study compared a few sequential designs (namely MSPE, entropy, maximin distance and cross validation) with a fixed-point optimal LHD for several test functions. Their findings suggested there is no clear winner in terms of global model fit and their reasoning is that earlier information, from the fitted kriging model (i.e., using less observations), for sequential designs, might be misleading for design point selection and hence reduce the effectiveness of the sequential designs.

2.3.2 Sequential Maximum Entropy Criterion

The *maximum entropy* design criterion, either (2.5) or (2.6), can also be modified for use as a sequential algorithm. The correlation matrix $\boldsymbol{R}$, which now includes the candidate point $\boldsymbol{x}_0$, can be partitioned into

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_n & r_n(\boldsymbol{x}_0) \\ r_n^T(\boldsymbol{x}_0) & 1 \end{pmatrix}, \tag{2.8}$$

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 $r_n(\boldsymbol{x}_0)$. As a result, $\det(\boldsymbol{R})$ can be written as a product of $\det(\boldsymbol{R}_n)(1 - r^T(\boldsymbol{x}_0)\boldsymbol{R}_n^{-1}r(\boldsymbol{x}_0))$. Note that this a product of scalar terms. Hence, one can show the sequential maximum entropy criterion based on (2.5) reduces to selecting a new point that satisfies

$$\max_{\boldsymbol{x}_0}(1 - r^T(\boldsymbol{x}_0)\boldsymbol{R}_n^{-1}r(\boldsymbol{x}_0)). \tag{2.9}$$

Notice that (2.9) is very similar to (2.7) except for the last term. For (2.6), where the $\boldsymbol{\beta}$ coefficients have a diffuse prior distribution, the sequential maximum entropy criterion (2.9) is equivalent to the sequential MSPE criterion in (2.7). The proof is provided in Appendix A.2.

2.4 Sequential Designs for Model Fit of Response Surfaces

The designs considered so far are very general in that they do not have a specific objective, such as global optimization or model fit, associated with them. If the research objective is to achieve a good global model fit of the GASP model, the cross validation approach offers a promising design strategy.

2.4.1 Cross Validation Prediction Error (XVPE) Criterion

As mentioned in Section 1.6, we may use cross validation to come up with an alternative measure of the prediction error to the MSPE for the stochastic model specified in (1.8). In turn, this prediction error will be used as part of the design criterion to select additional input points. This approach is motivated by noting that the MSPE of the model depends directly on the distance between sampled input points, $\boldsymbol{x}$, and on the correlation function $\boldsymbol{R}(.)$, but indirectly on the response values observed at these input points or the predicted values given by the fitted surface (even though the responses are used to estimate the parameters in the fitted GASP model). By considering criteria based on cross validation, we use the observed and predicted responses.

This cross validation approach has been studied by Jin *et al.* (2002), Keijnen and Beers (2004), and Beers and Kleijnen (2004). Jin *et al.* (2002) evaluated the use of cross validation for design purposes using radial basis function modeling. Unlike the stochastic model in (1.1), radial basis function modeling does not have a prediction error associated with the point predictions. This may have motivated the authors to propose a cross validation approach to estimate the prediction error and they compared the performance of their proposed criterion against various sequential designs. Their studies reported that there was no clear winner between the cross validation method and the other designs, including a fixed-point design. They did not use the cross validation approach for their comparisons with the GASP model.

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 $\{x_j; j = 1, \ldots, n\}$ are the sampled points, while $\hat{Y}_n(\boldsymbol{x})$ denotes the EBLUP of $y(\boldsymbol{x})$ using all the data. To reduce the 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 (penalized) *arithmetic mean* in

$$XVPE_{\mathrm{A}}(\boldsymbol{x}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2 \times \min_{j}(d(\boldsymbol{x}_j, \boldsymbol{x}))} \qquad (2.10)$$

which was also considered in Jin *et al.*(2002) in the context of radial basis function modeling. A penalty term, $d(.)$, based on Euclidean distance, is incorporated to penalize candidate points that are close to existing sampled points to prevent the next point picked from being close to one of the existing design points. To illustrate the problem of this criterion without the penalty term, first suppose the response $y(\boldsymbol{x}_j)$ is at a high local peak (or a point that has a large effect on the fitted GASP model). Predictions at candidate points around the point $\boldsymbol{x}_j$ will change drastically if $\boldsymbol{x}_j$ is not included in the training set. As a result, these candidate points will have a large value for the $(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2$ component in (2.10) when the j^{th} observation is deleted and one is likely to pick $\boldsymbol{x}_j$, or an $\boldsymbol{x}$ close to $\boldsymbol{x}_j$, as the next design point.

Two other studies by Kleijnen and Beers (2004), and Beers and Kleijnen (2004) proposed the use of cross validation and jackknifing to select the next design point given n points have already been observed. They first pre-selected c candidate input points by using a space-filling criterion (ignoring points that are close to existing design points). For each of these candidate points, they obtained the $n-1$ cross validation predictions. Based on these predicted values, they obtained the jackknife estimate for each of the candidate points, $\boldsymbol{x}$, using

$$\tilde{y}^{(i)}(\boldsymbol{x}) = n\hat{Y}_n(\boldsymbol{x}) - (n-1)\hat{Y}^{(-i)}(\boldsymbol{x}) \tag{2.11}$$

where $\hat{Y}_n(\boldsymbol{x})$ denotes the EBLUP of $y(\boldsymbol{x})$ based on all the n observations, while $\hat{Y}^{(-i)}(\boldsymbol{x})$ denotes the EBLUP of $y(\boldsymbol{x})$ based on all the data except $\{\boldsymbol{x}_i, y(\boldsymbol{x_i})\}$ where $i = 1, \cdots, n$. The corresponding prediction variance is computed as follows:

$$\tilde{s}^2(\boldsymbol{x}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{y}^{(i)}(\boldsymbol{x}) - \bar{\tilde{y}}(\boldsymbol{x})\right)^2 \tag{2.12}$$

where

$$\bar{\tilde{y}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}^{(i)}(\boldsymbol{x}).$$

The point among the c candidates with the largest prediction error (2.12) was selected next. The authors claimed that their jackknifing approach outperformed the approach proposed by Jin *et al.* (2002). However, it is not clear how the choice of the space-filling criterion for the pre-selection of the c candidate points might affect the selection of the final design points and their effectiveness in leading to an accurate predictive model.

2.5 Sequential Designs for Global Optimization

One of the earlier statistical approaches to global optimization, based on the stochastic model outlined in Section 1.3, was introduced by Cox and John (1996). They proposed an algorithm to search for the gobal minimum based on the EBLUP of $y(\boldsymbol{x})$ and MSPE in (1.7) and (1.8), respectively. These values were used to compute a lower confidence bound for each of the candidate points over a regular grid. Suppose we have a set of candidate points $\mathcal{X}$, the lower confidence bound for each of the points $\boldsymbol{x}$ was given by

$$\text{lcb}(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) - b\sqrt{MSPE(\boldsymbol{x})}$$

where $\hat{y}(\boldsymbol{x})$ is the EBLUP of $y(\boldsymbol{x})$ and $b = 2$ or 2.5 are used (pre-specified by the authors). The isotropic power exponential correlation function (with a fixed weak correlation parameter $\theta = 2$) was used. The candidate point with the smallest lower bound will be selected if it is smaller than the existing observed response $y_{\min}$. The algorithm was terminated if it either reaches a user-defined maximum number of points added, or if $y_{\min} < \min\limits_{\boldsymbol{x} \in \mathcal{X}} \{\text{lcb}(\boldsymbol{x})\}$. The results were encouraging and the authors suggested a probablistic approach to this problem as future research.

2.5.1 Expected Improvement Algorithm

Schonlau (1997) and Jones *et al.* (1998) subsequently developed a sequential (probabilistic) design strategy to add one input site at every stage of the sequential algorithm to search for the global optimum of the response surface. The authors provided two reasons as motivation for the development of their algorithm. The search for a global optimum point can be done purely by using the EBLUP version of (1.7) and sequentially selecting the point that gives the maximum/minimum predicted value. However, this search is too local and will most likely result in a local minimum unless the surface is well-fitted. On the other hand, selecting the point with the largest MSPE (1.8) leads to an overly global search, which will push points out to the boundaries and will not neccessarily find the optimum point quickly.

Their proposed sequential design strategy for global minimization is follows:

1. Choose a small initial set of n_0 sampled input points using a space-filling design such as a LHD. Fit the GASP model (1.1) to this set of n_0 points using the power-exponential correlation function. The parameters are estimated by maximum likelihood. The fitted model gives a predictor of $y(\boldsymbol{x})$ at unobserved input points $\boldsymbol{x}$ and the corresponding MSPE of the predictor.

2. Diagnostic plots, based on the leave-one-out cross validation approach in Subsection 2.4.1, is used to assess the fit of the initial model. For each subset of the $n-1$ sampled points, predict y at each of the sampled points that is left out using the EBLUP version of (1.7). This gives $n-1$ predictions at each

sampled point and they are plotted against the response y from the computer simulator at the corresponding sampled point. The points should lie roughly along a 45 degree line if the model fits well. The authors suggested transforming the observed response y if the diagnostic plots fail to show a good fit. Unless one finds a transformation for which the subsequent fitted model fits well, one would use the original output response without transformation.

3. The algorithm then proceeds with the goal of finding the input point that gives the global minimum response.

The proposed sequential algorithm is based on a notion of "improvement" defined by

$$I(\boldsymbol{x}) = \begin{cases} f^n_{min} - y(\boldsymbol{x}), & \text{if } y(\boldsymbol{x}) < f^n_{min} \\ 0 & \text{otherwise,} \end{cases} \tag{2.13}$$

where f^n_{min} is the known minimum point of the response surface evaluated at the n design points and $y(\boldsymbol{x})$ is the random quantity in (1.3). Since $y(\boldsymbol{x})$ is unknown, it can be shown that the expected improvement at each candidate input point $\boldsymbol{x}$ can be expressed as

$$E[I(\boldsymbol{x})] \;=\; (f^n_{min} - \hat{y}(\boldsymbol{x}))\,\Phi\left(\frac{f^n_{min} - \hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right) \;+\; s(\boldsymbol{x})\,\phi\left(\frac{f^n_{min} - \hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right) \tag{2.14}$$

after integrating (2.13) on both sides with respect to the conditional distribution of

$$[y|y^n, \boldsymbol{\beta}, \boldsymbol{\theta}] \sim N(\hat{y}(\boldsymbol{x}), s^2(\boldsymbol{x}))$$

where $\hat{y}(\boldsymbol{x})$ denotes the EBLUP version of (1.7), $s^2(\boldsymbol{x})$ is the corresponding MSPE (1.8), and $\Phi(\cdot)$ and $\phi(\cdot)$ are the $N(0,1)$ distribution and density function, respectively. This expected improvement will give large values for candidate points with: (i) predicted values much less then f^n_{min} (sometimes referred to as the localized search component), or (ii) high uncertainty (large $s(\boldsymbol{x})$) about the prediction $\hat{y}(\boldsymbol{x})$ (sometimes referred to as the global search component). The search for the minimum point then proceeds by finding the point $\boldsymbol{x}$ that maximizes the expected improvement (EI) at each stage and the observing the response at this $\boldsymbol{x}$. It can be shown that $EI[\boldsymbol{x}] = 0$ if $\boldsymbol{x}$ is an existing design point.

The branch-and-bound algorithm is used for the numerical search for the maximum likelihood estimates of the correlation parameters. The algorithm is terminated if the expected improvement is less than a pre-specified cut-off of the best current function value, i.e., $\frac{max(EI)}{|f^n_{min}|} < \alpha$ where α is the pre-specified cut-off value.

Schonlau *et al.* (1998) extended the EI algorithm by incorporating an additional parameter, g, that systematically controls the balance between the global and local search range of the algorithm. The improvement function (2.13) becomes

$$I^g(\boldsymbol{x}) = \begin{cases} (f^n_{min} - y(\boldsymbol{x}))^g, & \text{if } y(\boldsymbol{x}) < f^n_{min} \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

where $g = 0, 1, 2, 3, \ldots$. Increasing values of g indicates a more global search. Taking $g = 0$ gives $\mathrm{E}[I^0(\boldsymbol{x})] = P(y(\boldsymbol{x}) < f^n_{min}) = \Phi\left(\frac{f^n_{min} - \hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right)$ which results in a very localized search and is not reccommended unless the surface is fitted well. A poor fit of the model, based on the diagnostic tool suggested in Jones *et al.* (1998), may indicate that the choice of $g = 1$ may be undesirable and g should be increased. Intuitively, this there is a need to improve the fit and hence search for better global model fit rather than a search for a global optimum. A recursive formula for computing $\mathrm{E}(I^g(\boldsymbol{x})|y^n)$ is given in Jones *et al.* (1998). It might be interesting to consider periodically assessing the current model fit and reducing g as the fit improves.

Schonlau *et al.* (1998) further extended the EI algorithm to add more than one input point at one time. Given n points, and if q points are to be added at each stage, the q-step "improvement" function becomes

$$I^g(\boldsymbol{x}) \;=\; [\max(0, y_{min} - y_{n+1}, \ldots, y_{min} - y_{n+q})]^g. \tag{2.16}$$

Computing the q-dimensional integration of $I^g(\boldsymbol{x})$ to get the expected improvement is a daunting task. As a simplification, the authors recommended (i) computing the expectations of $I^g(\boldsymbol{x})$ sequentially and (ii) updating the $s(\boldsymbol{x})$ term at each iterate but not the $\frac{f_{\min}-\hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}$ term. The rationale was that updating $\frac{f_{\min}-\hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}$ implied that the difference of $f_{\min} - \hat{y}(\boldsymbol{x})$ was known with greater certainty (which is not true). Prediction $\hat{y}(\boldsymbol{x})$ were not updated until the q runs are actually made. Schonlau *et al.* (1998) presented an example where they compared the "add one point" sequential design versus the "add many points" design in terms of how well each method located the global minimum. Their results showed that the designs were comparable.

2.5.2 Expected Improvement Algorithm with Noise Variables

Williams *et al.* (2000) futher extended the EI algorithm (2.14) to include input settings with control and environmental variables. The optimization procedure is to minimize a weighted average of the response over a discrete set of values for the environmental variables. Let $\boldsymbol{x}_c$ and $\boldsymbol{x}_e$ denote the control and environmental variables respectively and n_e denote the number of discrete levels for the environmental variable. The objective function is given by $l(\boldsymbol{x}_c) \;=\; \sum_{i=1}^{n_e} w_i y(\boldsymbol{x}_c, \boldsymbol{x}_{e,i})$. The goal is to identify the control variable setting $\boldsymbol{x}_c^*$ that minimizes $l(\boldsymbol{x}_c)$. This is analogous to finding the minimum $y(\boldsymbol{x})$ in the original EI algorithm.

The modified EI algorithm is summarized as:

1. The initial set of design points are selected using a space-filling maximin LHD, similar to the procedure in Jones *et al.* (1998) and Schonlau (1998).

2. The parameters in the model are estimated by maximizing the joint posterior distribution of the parameters. The key difference here is that Williams *et al.* (2000) adopted a Bayesian approach to this problem. The parameters were assumed to have non-informative prior distributions. Instead of the power exponential correlation function, the Matérn correlation function (1.13) was used in this study.

3. Proceed with finding the next control variable input $\boldsymbol{t}_{c,n+1}$ such that it maximizes the expected improvement

   $$\max_{\boldsymbol{x}_c} E\{I_n(\boldsymbol{x}_c)|\boldsymbol{Y}_{S_n}, \boldsymbol{\zeta}_n\}, \tag{2.17}$$

   where the maximization is taken over the posterior distibution $[.|\boldsymbol{Y}_{S_n}, \boldsymbol{\zeta}_n]$ and $\boldsymbol{\zeta}_n$ denotes the correlation parameter(s). And, let $\boldsymbol{S}_n = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n\}$ denote the sampled points, $\boldsymbol{S}_n^c = \{\boldsymbol{t}_{c,1}, \ldots, \boldsymbol{t}_{c,n}\}$ denote the control variable portion and $\boldsymbol{Y}_{S_n}$ denote the random vector of responses associated with the sampled points $\boldsymbol{S}_n$. The modified "improvement" function is

   $$I_n(\boldsymbol{x}_c) = \begin{cases} L_{1:n} - L(\boldsymbol{x}_c), & \text{if } L(\boldsymbol{x}_c) < L_{1:n} \\ 0 & \text{otherwise,} \end{cases} \tag{2.18}$$

   where $L_{1:n} = \min\{L(\boldsymbol{t}_{c,1}), \ldots, L(\boldsymbol{t}_{e,n})\}$ at the $\boldsymbol{x}_c$ that gives the minimum weighted average. The term $L_{1:n}$ is a random variable unlike in the original EI algorithm where the corresponding term f_{min} is known.

4. Choose the next environmental input corresponding to the control input $\boldsymbol{t}_{c,n+1}$ to minimize the *posterior mean square prediction error* given the current data

(i.e. including the new control input, $\boldsymbol{t}_{c,n+1}$), by

$$\underset{\boldsymbol{x}_e}{min}\, E\big\{[\hat{L}_{n+1}(\boldsymbol{t}_{c,n+1}) - L(\boldsymbol{t}_{c,n+1})]^2 \mid \boldsymbol{Y}_{S_n}, \boldsymbol{\zeta}_n\big\}$$

where $\hat{L}_{n+1}(\boldsymbol{t}_{c,n+1})$ is the posterior mean of $[L(\boldsymbol{t}_{c,n+1})|\boldsymbol{Y}_{S_n}, \boldsymbol{\zeta}_n]$.

5. The procedure for searching the minimum point is iterated until the stopping criterion is met.

CHAPTER 3

COMPARISON OF EXPERIMENTAL DESIGNS FOR RESPONSE SURFACE MODEL FIT

For the purpose of fitting a response surface to the output from computer simulations, a variety of designs based on a fixed number of runs have been proposed. As pointed out in Chapter 2, space-filling designs and (sequential) designs based on optimality criteria, such as MSPE and entropy, make no use of information gained about the shape of the predicted response surface from the observed responses and hence it is not clear if these designs will quickly result in an accurate predictive model.

In this chapter, we consider sequential adaptive designs as "efficient" alternatives to fixed-point designs and sequential designs that do not make direct use of the responses. New adaptive design criteria are proposed based on a cross validation approach and on an expected improvement criterion, the latter inspired by a criterion originally proposed for global optimization. While many sequential designs have been proposed (for example, in Chapter 2 and more in this chapter), it is not clear how the performance of these methods might be affected by the shape of response surface, choice of correlation function for the GASP model, size of starting designs etc. This chapter will address some of these issues and compare these new designs (i.e., the cross validation approach and the modified expected improvement criterion) with others in the literature in an empirical study (in Subsections 3.4.1 and 3.4.3).

In general, we are optimistic that sequential designs are 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 designs are sequential (e.g., Sobol sequences) but not adaptive.

Sequential algorithms have the desirable property that additional observations are naturally accomodated if the need to improve the accuracy of the GASP model arises. Although the diagnostic plots mentioned in Schonlau (1997) briefly discussed in Subsection 2.5.1 provide a way to assess the fit of the GASP model, it is not clear what can be done if a poor fit is observed even after taking transformation of the responses. This motivates a need for designs that will allow input points to be added to improve the response surface model fit.

3.1 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 thesis, attention is restricted 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 1.3. For our approach, the best linear unbiased predictor is used to predict the response at unobserved $\boldsymbol{x}$, based on the available training data.

3.1.1 Model and Best Linear Unbiased Predictors

Following the approach in Section 1.3, 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

$$Y(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{3.1}$$

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 (3.1) 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

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

The *best linear unbiased predictor* (BLUP) is given by

$$\hat{Y}(\boldsymbol{x}_0) = \boldsymbol{f}^T(\boldsymbol{x}_0)\hat{\boldsymbol{\beta}} + r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}\hat{\boldsymbol{\beta}}), \tag{3.2}$$

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 $1 \leq i \leq n$, $1 \leq j \leq k$. The *mean squared prediction error* (MSPE) of the

BLUP is then given by

$$\begin{aligned} MSPE[\hat{Y}(\boldsymbol{x_0})] &= \sigma^2[1 - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x_0}) + \qquad (3.3) \\ (f^T(\boldsymbol{x_0}) &- r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}(f^T(\boldsymbol{x_0}) - r^T(\boldsymbol{x_0})\boldsymbol{R}^{-1}\boldsymbol{F})^T]. \end{aligned}$$

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 (3.1) 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 if the response surface is not too highly non-stationary. The stochastic process $Z(\boldsymbol{x})$ captures the local trend which usually suffices to produce excellent fit.

Given that the correlation function $R(\cdot)$ is known, the BLUP can be easily calculated using (3.2). In this study, the correlation parameters will be estimated by the maximum likelihood approach and the resulting predictor is termed as the *empirical best linear unbiased predictor* (EBLUP).

3.1.2 Parametric Correlation Functions

As seen from the equations (3.2) and (3.3) above, the correlation function $R(\cdot)$ plays an important role and has to be specified by the user. Both the cubic and power exponential (with $p = 2$)correlation functions will be used for the examples in Section 3.4. We will also use the product correlation structure, $R(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{j=1}^{m} R(|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}| \, |\theta_j)$ where m denotes the number of dimensions, and let $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$. The one-dimensional forms of the cubic and power exponential correlation functions, from Subsection 1.3.3, are shown again below.

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

$$\begin{aligned} R(d) \quad &= 1 - 6\left(\tfrac{d}{2}\right)^2 + 6\left(\tfrac{|d|}{\theta}\right)^3, \quad && |d| < \tfrac{\theta}{2} \\ &= 2\left(1 - \tfrac{|d|}{\theta}\right)^3, && \tfrac{\theta}{2} \le |d| < \theta \\ &= 0, && |d| \ge \theta \end{aligned} \tag{3.4}$$

where $\theta > 0$ and d denotes the distance between two points (see Currin *et al.*, 1991 and Mitchell *et al.*, 1990). One 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. Subsection 3.4 will provide more details on how this feature of the cubic correlation is used for the examples in this thesis.

Power Exponential Correlation. Another very popular correlation function takes the form of

$$R(d) = exp(-\theta|d|^p), \tag{3.5}$$

where $0 < p \le 2$ and $\theta \in (0, \infty)$. For the special case of $p = 2$, this corresponds to the Gaussian correlation function which gives an EBLUP (and BLUP) that is infinitely differentiable.

3.1.3 Basic Algorithm for Constructing Sequential Designs

One can think of sequential designs in the context of either augmenting an existing network of input points or selecting a completely new set of input points. In both cases, it will be assumed that the starting designs will consist of at least two input points because it is not possible to fit a GASP model with a single point. As a result, there is really no difference in the implementation of sequential designs for these two cases.

The sequential algorithm for response surface model fit proceeds as follows:

1. Identify the set of existing n sampled design points $\boldsymbol{x}$ for either

   (a) Augmenting the network of design points: identify the existing design points.

   (b) Selecting a new set of design points: generate a small starting design spread over the input space $\mathcal{X}$. A space-filling design would be appealing for this initial fit of the response surface using the GASP model.

2. Run the computer code at the n input points identified in Step 1 and obtain the response, $y(\boldsymbol{x})$.

3. Estimate the correlation parameters (see Subsection 3.1.2).

4. Fit a GASP model (see Subsection 3.1.1) using the n observations from the computer simulator to predict $y(\boldsymbol{x}_0)$ at some untried $\boldsymbol{x}_0$.

5. Check the stopping rule, if available. If additional points are needed, search for the $\boldsymbol{x}_0$ that maximizes the design criterion (see Section 3.4 for criteria used in the simulation study) and add it to the existing set of sampled points giving a total of $n+1$ points. Repeat from Step 2 onwards with the $n+1$ points.

3.2 Proposed Sequential Design Criteria

3.2.1 Sequential Integrated Mean Squared Prediction Error Design

As alluded to in Subsection 2.2.1, direct implementation of the *integrated mean squared prediction error* (IMSPE) criterion sequentially 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 the $\boldsymbol{x}_0$ that

$$\min_{\boldsymbol{x}_0 \in \mathcal{X}} \{IMSPE(\boldsymbol{x}_0)/\min(d(\boldsymbol{x}_i, \boldsymbol{x}_0))\} \tag{3.6}$$

where $\boldsymbol{x}_i$ denotes an existing input point that is closest to $\boldsymbol{x}_0$. This is proposed as an alternative to the sequential approach suggested by Sacks, Schiller and Welch (1989) for (2.2) which is just minimizing the numerator in (3.6). The distance penalty is incorporated to push subsequent points away from existing input points and hence prevent the clumping problem.

3.2.2 Cross Validation Prediction Error Criteria

The criteria used in the *cross validation* approaches in (2.10) and (2.11) are based on the arithmetic mean of the cross validation prediction errors. The criterion based on (2.10) is penalized by distance, while (2.11) requires a pre-selection of candidate 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 in (2.7), maximum entropy (2.9), and the cross validation method (2.10), the new summaries 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. The criterion is to maximize

$$XVPE_{\mathrm{G}}(\boldsymbol{x}) = \sqrt[n]{\prod_{j=1}^{n}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2}. \qquad (3.7)$$

This avoids the need for pre-selection of candidate points as needed in the jackknifing approach in (2.11).

The *harmonic* mean of a sequence of numbers 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. The criterion is to maximize

$$XVPE_{\mathrm{H}}(\boldsymbol{x}) = \frac{n}{\sum\limits_{j=1}^{n} \frac{1}{(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2}}. \qquad (3.8)$$

As an aside, we note that the harmonic mean is always the smallest of the three types of means, while the arithmetic mean is always the greatest and the geometric mean is always in between.

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 which is to maximize

$$XVPE_{\mathrm{M}}(\boldsymbol{x}) = \min_{j}(\hat{Y}^{(-j)}(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2. \qquad (3.9)$$

3.2.3 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

$$I(\boldsymbol{x}) = (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 \tag{3.10}$$

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

$$E(I(\boldsymbol{x})) = (\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 + var(\hat{Y}(\boldsymbol{x})). \tag{3.11}$$

The derivation for the EIGF criterion is given in Appendix A.3. More details and illustrations of this criterion will be presented in Chapter 4.

3.3 Implementation of Sequential Designs

This section addresses some of the practical issues related to the implementation of sequential designs under the GASP model framework in general. Successful implementation of such designs depends on several factors, such as: (i) some knowledge of the features of the true response surface to be modeled (ii) choice of correlation function for the GASP model, (iii) choice of experimental design criterion, (iv) number of starting design points, (v) how to generate the starting design points, and (vi) the final number of input points.

Ideally, there should be *some knowledge of the response surfaces* even though the actual computer code used to generate the surface may not be known. For instance, it would be helpful to know whether the main features of the response surface are located at the boundaries or concentrated in the interior of the input space. This may help to determine some of the choices from (ii) to (vi) above. The *choice of correlation function* is an important component of the GASP model. A few choices are available (see Section 1.3.3) although the Matérn correlation, which is theoretically appealing, is typically not preferred because it is computationally more intensive and a power exponential correlation might do equally well in terms of prediction (see Lehman, 2002, Chapter 2) . The power exponential correlation with $p = 2$ (also called the Gaussian correlation) is a popular choice although we find, based on our experience, that estimation of the correlation parameters, using maximum likelihood estimation, tends to be unstable especially in the context of sequential designs. The cubic correlation function is an alternative that seems to perform rather well in our study.

We believe the key to a successful implmentation of sequential designs lies in the choice of the *experimental design criterion*. 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).

The *number of starting design points* is an important component of sequential designs, but it is generally not clear exactly what this number should be. 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 too. We will discuss an approach to decide on a stopping criterion in Chapter 5.

There exist a number of space-filling design criteria as mentioned in Section 2.1 but studies (e.g., Marin, 2005) suggest they perform similarly. Thus, in Sections 3.4 (and examples in subsequent chapters) we use a maximin LHD as representative of a space-filling design and also as the *starting designs* for the sequential methods.

Following this in Section 3.4, we present various sequential design criteria proposed in this chapter and give several examples to illustrate the effectiveness of the various designs.

3.4 Examples: Comparison of Design Criteria

The following examples illustrate the implementation and prediction performance of the various sequential designs and the fixed-points design. Various functions are used as "true" functions to compare the prediction performances of these designs using a small number of sampled points. There are two categories of response surfaces considered in this study: (i) random surfaces generated from a true stationary GASP model in Subsection 3.4.1, (ii) response surfaces, mostly generated from mathematical functions, which display a variety of features in the response in Subsection 3.4.3. In Subsection 3.4.5, an example is presented to compare the efficiency gained (in terms of number of observations to achieve a similar degree of predictive accuracy) across various designs.

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 parenthesis 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 20, 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 (MATLAB® codes for implementing the sequential designs presented in this thesis can be found on the department's computer experiment directory at /home/comp_exp/SOFTWARE/EIGF). For sequential designs, this means 30 different starting designs all approximately maximin LHDs (see Subsection 3.4.4 where we compared the predictive accuracy of the various designs based on the "best" maximin LHD as a starting design).

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, 20 or 30 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.

Estimation of Correlation Parameters

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.

In order for the sequential methods to work well, it is crucial to constrain the range of the numerical search for the maximum likelihood estimates. Although many studies have chosen to fix the values for these correlation parameters (as mentioned in Section 2.2), it is not clear what these values should be in general. Sacks *et al.* (1989) conducted a robust study to identify values for the correlation parameters in their IMSE criterion. The authors suggested very strong correlation parameter values for the Gaussian correlation function seem to have good relative efficiency for prediction. Lim *et al.* (2002) also commented that strong correlation seemed to work well in their study. In contrast, Currin *et al.* (1991) used the exponential correlation function with $e^{-\theta} = 0.0001$ (very weak correlation) for design purpose. And similarly, Mitchell and Morris (1992) favored using weak correlation parameters in the initial design phase, although they encountered some numerical difficulties.

Instead of fixing the value of the correlation parameters, we choose to carry out a numerical search across a wide range of correlation values and select the value that maximizes the likelihood function. The following constraints, for the cubic correlation (3.4), are used:

1. The two closest input points, in distance, must be at least weakly correlated. The lower limits for the range parameters $\boldsymbol{\theta}$, in (3.4), are set to be slightly larger than the distance between the two closest points. This is fixed at 1.1 times the minimum Euclidean distance between the points and this gives a correlation strength of about $R(\cdot) = 0.0015$ between the two closest points in a two-dimensional input space.

2. Points must have the chance to be highly correlated. The upper limits for the range parameters $\boldsymbol{\theta}$, in each dimension, are set to be four times the corresponding range of the input space.

For comparison purposes, similar constraints are specified for the Gaussian correlation function to match the correlation values used for the cubic correlation. Re-expressing (3.5) with $p = 2$ in one dimension, we have

$$\theta = \frac{-\ln R}{d^2}$$

where d denotes the range of the input space. The limits for the Gaussian correlation parameter θ can be computed by substituting the corresponding $R(\cdot)$ values based on the limits for the cubic correlation (in the above paragraph).

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

$$ERMSPE = \sqrt{\frac{\sum\limits_{i=1}^{m}(\hat{y}(\boldsymbol{x}_i) - y(\boldsymbol{x}_i))^2}{m}} \tag{3.12}$$

where $\boldsymbol{x}_i$, $i = 1, ...m$ $(m >> N_0)$ are a grid of equally spaced 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 at the same set of grid points. 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 spread out over $\boldsymbol{\mathcal{X}}$. Boxplots are used for each of the test functions to show the distribution of the ERMSPE for the 30 runs.

3.4.1 Test Functions: Random Surfaces Generated from Gaussian Stochastic Process (GASP) Model

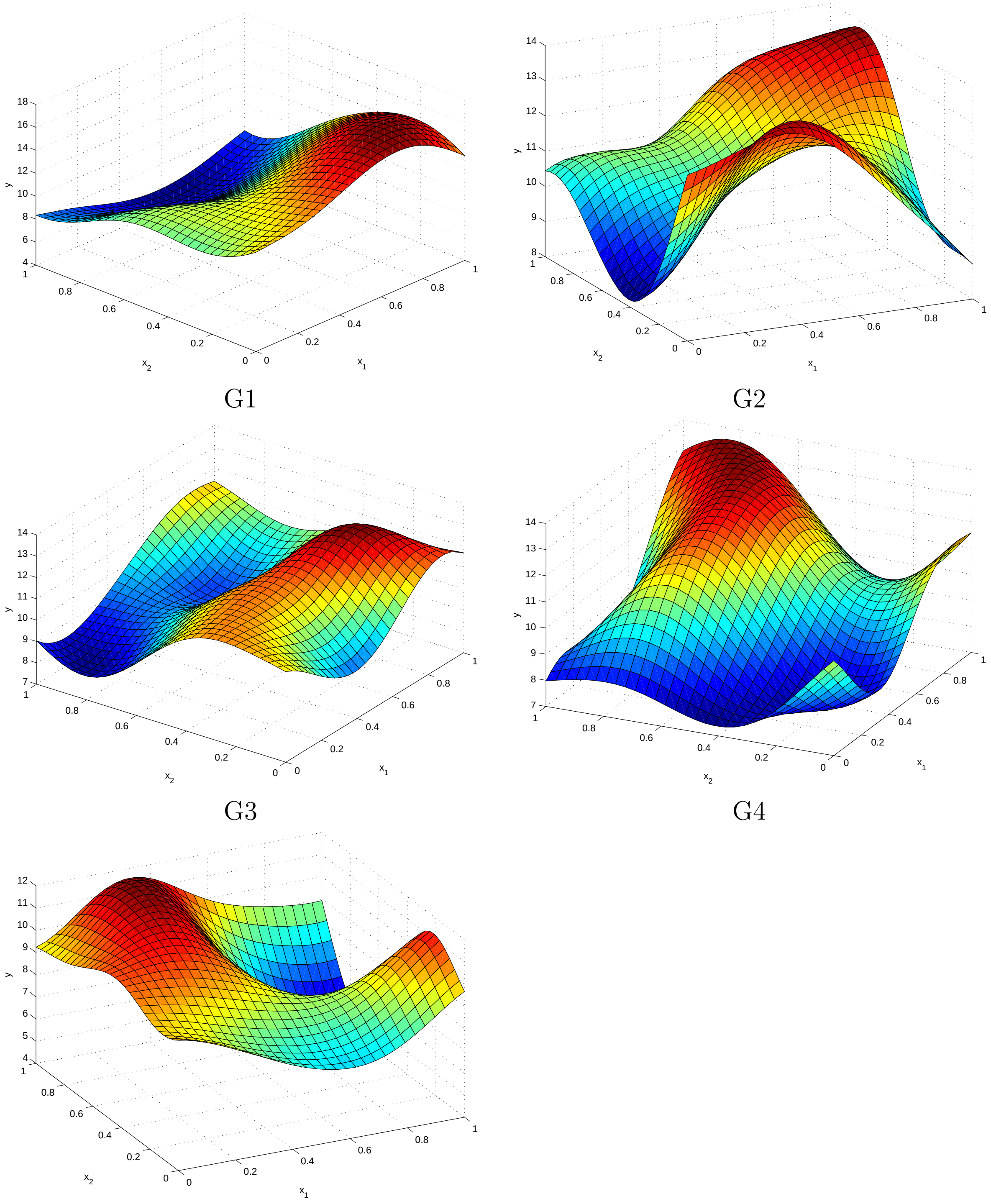
Random surfaces are generated from a stationary GASP model (3.1) with covariance matrix $\Sigma = \sigma^2 R(\boldsymbol{\theta})$. For $R(\boldsymbol{\theta})$, the Gaussian correlation function is used. The surfaces are generated with $\sigma^2 = 5$ and $\boldsymbol{\theta} = (5, 5)$, which give moderately strong correlation and result in rather smooth surfaces. These random surfaces will be referred to as GASP surfaces. Plots of the five random surfaces (denoted as surfaces G1 to G5) are shown in Figure 3.1 (page 56) on a grid of $m = 30 \times 30 = 900$ equally spaced points which coincide with the m points used to evaluate the designs in (3.12). These plots are used to illustrate the implementation and prediction performance of the various designs. All sequential designs have $N_0 = 5$ points for the starting designs. The comparison of predictive accuracy of the different design criteria is based on a final number of $N = 30$ points.

This example seeks to examine the impact of the choice of correlation function, initial starting design and design criterion on the predictive performance across the five surfaces (realizations) from the GASP model. In addition, we consider the following cases for the estimation of the correlation parameters for fitting the constant mean GASP model.

1. with the Gaussian correlation function and parameters estimated by maximum likelihood

2. with the cubic correlation function and parameters estimated by maximum likelihood

3. with the Gaussian correlation function and parameters known, $\boldsymbol{\theta} = (5, 5)$



G1

G2

G3

G4

G5

Figure 3.1: GASP surfaces: Plots of random surfaces generated from GASP model with $\boldsymbol{\theta} = (5, 5)$.

3.4.2 Results: GASP Model Response Surfaces

GASP surfaces: Figure 3.3 (page 69) shows comparative plots of the ERMSPE for the nine designs. It is not clear from Figures 3.3 (a) and (b) that designs with the Gaussian correlation will neccessarily result in better predictions than designs with the cubic correlation. The ERMSPE seems to be slightly lower with the Gaussian correlation but some extremely poor predictions can result, and this makes designs with Gaussian correlation less robust in terms of predictive accuracy. Due to their "adaptive" property, sequential adaptive designs run into problems with estimation of the Gaussian correlation parameters possibly due to the irregularity of inter-point distances between the sampled points (unlike the space-filling maximin LHD). The fixed-point maximin LHD (s) (the top five boxplots shown on each figure) seems to do better with the Gaussian correlation than with the cubic correlation. This is not surprising since the surfaces were generated using the Gaussian correlation function (Subsection 3.4.1).

In Figure 3.3 (c), all designs, except the EIGF criterion (ei), perform better with $\boldsymbol{\theta}$ known compared to (a) and (b) where the correlation parameters are estimated by maximum likelihood. This provides compelling evidence that some of these sequential designs fail to predict well due to difficulties in estimating the correlation parameters.

Overall, sequential designs, except the EIGF criterion (ei), tended to do better than the fixed-point maxmin LHD (s). However, none of these sequential designs seem to stand out with the Gaussian correlation in Figure 3.3 (a). However when the cubic correlation is used, the cross validation with the harmonic mean and the maximin criteria (xh and xm), and EIGF criterion (ei) are the worse performers among the sequential designs.

A detailed analysis reveals that the EIGF criterion (ei) tended to place too many points in regions with high variation in the response and suffers a loss of predictive accuracy over the other regions of the input space. A generalization of the the EIGF criterion (ei) is considered in Section 4.4 to address this problem.

3.4.3 Test Functions: Non-GASP Model Response Surfaces

In general, we not expect the simulator to produce GASP model response surfaces. We consider a few other test functions to evaluate the predictive performance of the various designs. In Subsections 3.4.3 and 3.4.4, four examples are used to evaluate the predictive performance 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 3.2 (page 60).

Boxplots are used for each of the test functions to show the distribution of the ERMSPE for the 30 runs. Plots of the worst case predicted surface (i.e., the surface with the maximum ERMSPE among the 30 runs), for both the Gaussian and cubic correlations, and different number of starting design points are also shown. These plots are shown on pages 70 to 81.

Function M1: Branin function

The two-dimensional Branin function, also considered in Schonlau (1997), is given by.

$$f(x_1, x_2) \;=\; (x_2 - \frac{5.1}{4\pi^2}x_1^2 + 5/\pi x_1 - 6)^2 + 10(1 - \frac{1}{8\pi}\pi)\; cos(x_1) + 10$$

where $x_1 \in [-5, 10],\; x_2 \in [0, 15]$. The true surface is plotted in Figure 3.2 on a fine grid of $m = 30 \times 30 = 900$ equally spaced points which coincide with the m points used to evaluate the designs in (3.12). This surface has peaks in the response at the boundaries of the input space. We follow 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. The final number of input points, N, is taken to be 20.

Function M2: 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 equally spaced points. This surface is constructed to examine the performances of the design strategies in a setting where the boundaries are "flat". Due to the complexity of the surface, the final number of input points, N, is taken to be 40.

Function M3: 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 3.2 on $m = 30 \times 30 = 900$ equally spaced points. The final number of input points, N, is 40.

Function M4: Non-polynomial surface

We consider a two-dimensional function that has highly correlated responses but relatively mild variation in the response. This relatively smooth response function is given by

$$f(x_1, x_2) \ = \ \frac{(30 + 5x_1 sin(5x_1)) \times (4 + exp^{-5x_2}) - 100}{6}$$

where $x_1, x_2 \in [0,1]$. This surface is considered in Lim *et al.* (2002). The true surface is plotted in Figure 3.2 on $m = 30 \times 30 = 900$ equally spaced points and the final number of input points, N, is 20.

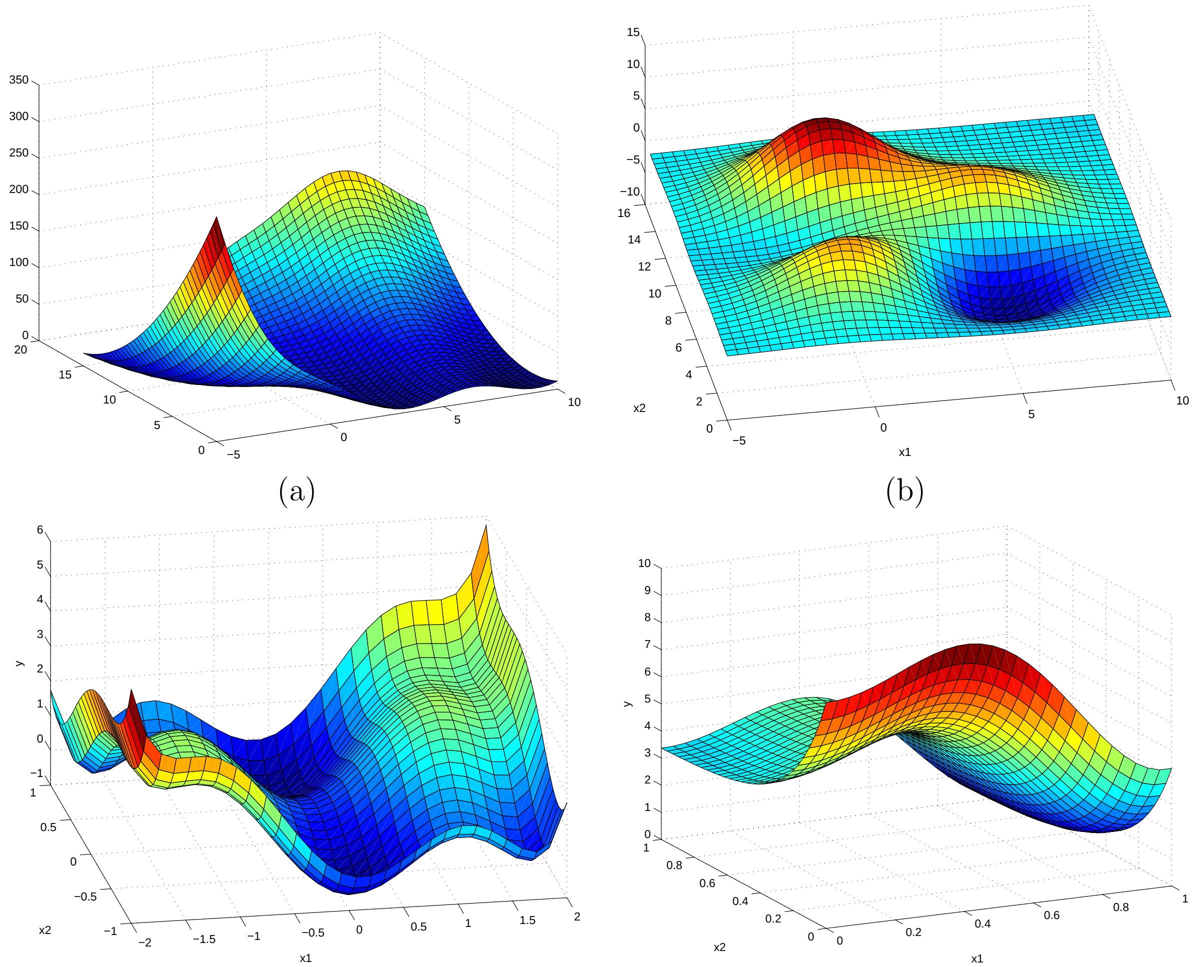


Figure 3.2: Surface plots of the true surfaces. (a) Function M1: Branin function, (b) Function M2: Simulated surface, (c) Function M3: Six-hump camel-back function, (d) Function M4: Non-polynomial surface.

3.4.4 Results: Non-GASP Model Response Surfaces

Among the 30 different starting designs (as mentioned on page 51), the "best" maximin LHD (i.e., the design, among the 30 runs, that has the largest maximin distance between the input points) is identified and the ERMSPE of the resulting final design with N points is shown as ($\times$) on the boxplots in Figures 3.4 (page 70), 3.7 (page 73), 3.10 (page 76) and 3.13 (page 79). From these plots, the "best" maximin LHD for each design criterion does not necessarily result in the most accurate predictive GASP model (i.e. the smallest ERMSPE shown in the boxplots).

Comparison of correlation function and number of starting design points

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.

Function M1 (Branin function): The boxplots in Figure 3.4 (page 70) show that although the cubic correlation function tends to result in designs with slightly larger ERMSPE median, the spread tends to be smaller and with fewer outliers. Figures 3.5 and 3.6 show that designs with the Gaussian correlation tend to result in relatively poorer predictions of the interior region of the input space.

Function M2 (Simulated surface): In Figure 3.7 (page 73), the Gaussian correlation function seems to perform better (with fewer outlying ERMSPE values) at least for smaller starting designs (i.e. $N_0 = 5$ and 10). Figures 3.8 and 3.9 show that some designs ($N_0 = 20$) with the cubic correlation can result in very poor predictions. Overall with $N_0 = 20$, predictions using the two types of correlation are comparable.

Function M3 (Six-hump camel-back function): 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 3.10 on page 76). Figures 3.11 and 3.12 show the worst case predicted surfaces and the Gaussian correlation is found to result in poor predictions in some cases.

Function M4 (Non-polynomial surface): The use of the cubic correlation tends to results in relatively smaller ERMSPE median and spread. However, the difference is not too significant and designs with both correlation functions give comparable predictions. Designs with the cubic correlation tend to predict the boundaries more accurately as shown in Figures 3.14 and 3.15 on pages 80 and 81 respectively. This is somewhat surprising since this is a very smooth surface and the Gaussian correlation is expected to be more suitable. We suspect that because of the sharp and rapidly changing edges of some of the surfaces, the cubic correlation tends to give better predictions since it is a piece-wise cubic spline interpolator. These extreme behaviors may not be sufficiently captured by the Gaussian correlation function.

In conclusion, for the "best" designs, the cubic performs as well as the Gaussian. Based on various examples that we have examined, 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.

Comparison of design criteria

Function M1 (Branin function): The sequential designs are clearly superior to the fixed-point maximin LHD (s) in Figure 3.4. The EIGF criterion (ei) outperforms all the other sequential designs and has a bigger advantage with $N_0 = 5$. The fact that the maximum MSPE (m) and maximum entropy (e) criteria select more points on the boundaries seems to give them a slight advantage in this example. Among cross validation methods, the harmonic mean and the maximin criteria (xh and xm) look the worst for both $N_0 = 5$ and $N_0 = 10$. Figure 3.5 shows that the surface is reasonably well predicted by all designs.

Function M2 (Simulated surface): 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 M3 (Six-hump camel-back function): The maximin LHD design (s) is the worst performer among all the designs (see Figure 3.10). 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 M4 (Non-polynomial surface): Overall, the sequential designs outperform the fixed-point maximin LHD. The 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. With relatively less points placed on/near the boundaries of the input space, the fixed-point maximin LHD appears to give poor predictions at the edge of the top right corner and bottom center regions, shown in Figures 3.14 and 3.15.

Overall conclusion and other issues

Among the sequential adaptive designs for the four functions considered in this subsection, the cross validation prediction error criterion with the arithmetic mean and geometric mean (both with larger starting designs), and EIGF criterion with smaller starting designs are the better performers (in terms of both ERMSPE median and spread) using the GASP model. In contrast, the fixed-point LHD (s) and cross validation prediction error criterion with the harmonic mean and maximin criterion tend to have larger ERMSPE median and/or spread (i.e., less robust to the starting design).

We have chosen to compare the predictive performances of the designs based on a fixed sample size, N. It would be interesting to explore the possibility using a sequential designs with fewer design points to achieve the same predictive performance as a fixed-point design. An example will be presented next in Subsection 3.4.5.

3.4.5 Test Function: Comparing The Efficiency of Experimental Designs

In this subsection, we wish to compare the predictive accuracy and efficiency (in terms of sample size) of various designs as we increase the final number of sampled points.

Function M5: Hartman 3 Function

We consider the three-dimensional Hartman 3 function given by

$$f(x_1, x_2, x_3) = \sum_{i=1}^{4} \alpha_i exp \left[\sum_{j=1}^{3} A_{ij} (\boldsymbol{x}_j - P_{ij})^2 \right]$$

where $\alpha = [1, 1, 2, 3, 3, 2]^T$, $\boldsymbol{A} = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$ and $\boldsymbol{P} = 10^{-4} \begin{pmatrix} 6890 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{pmatrix}$ where $0 < \boldsymbol{x}_j < 1, j = 1, 2, 3$. A total randomly selected $m = 4000$ points (out of the $30 \times 30 \times 30 = 2,7000$ regularly spaced grid points) are used to evaluate the designs.

Suppose we follow the rule of thumb suggested in Jones *et al.* (1998) for selecting an initial fixed-point design of $N_0 = 30$ points for the Hartman 3 function. A GASP model with the cubic correlation function is fitted using the output from the computer code at these 30 sampled points. We then compare the predictive accuracy of various designs as we increase the final number of sampled points from $N = 40$ to $N = 90$. The designs to be compared are the EIGF criterion, cross validation with the arithmetic mean and the geometric mean, and the fixed-point maximin LHD. For the maximin LHD, this means re-generating a new set of design points for any increase in N. Only the cubic correlation function will be used in this example.

3.4.6 Results: Three-Dimensional Hartman 3 Function

Results from this simulation study shows that sequential designs tend to be more efficient in terms of the number of sampled points to achieve the same predictive accuracy. The cross validation with the arithmetic mean and the geometric mean (xa and xg) criteria consistently outperform the fixed-point maximin LHD (s) as we increase the number of sampled points from $N = 40$ to $N = 90$ (see Figure 3.16 on page 82). In addition, the cross validation criteria (xa and xg) tend to require about 10 points fewer than the fixed-point maximin LHD (s) to achieve the same predictive accuracy. For example with $N = 60$, the medians of the ERMSPE for the cross validation criteria (xa and xg) and the maximin LHD (s) are about 0.8, 0.75 and 0.9 respectively. It takes $N = 70$ for the maximin LHD (s) to reduce the median of the ERMSPE to about 0.75.

The boxplots in Figure 3.16 for the EIGF criterion (ei) show mixed results. It performs better than the maximin LHD (s) initially with $N = 40$ to $N = 60$, but is outperformed by the maximin LHD (s) with $N = 70$ and more. As the sample size increases, the EIGF criterion (ei) tends to concentrate its sampling effort in regions with more variation in the response (although not too significant) and this results in poorer fit for the flatter region.

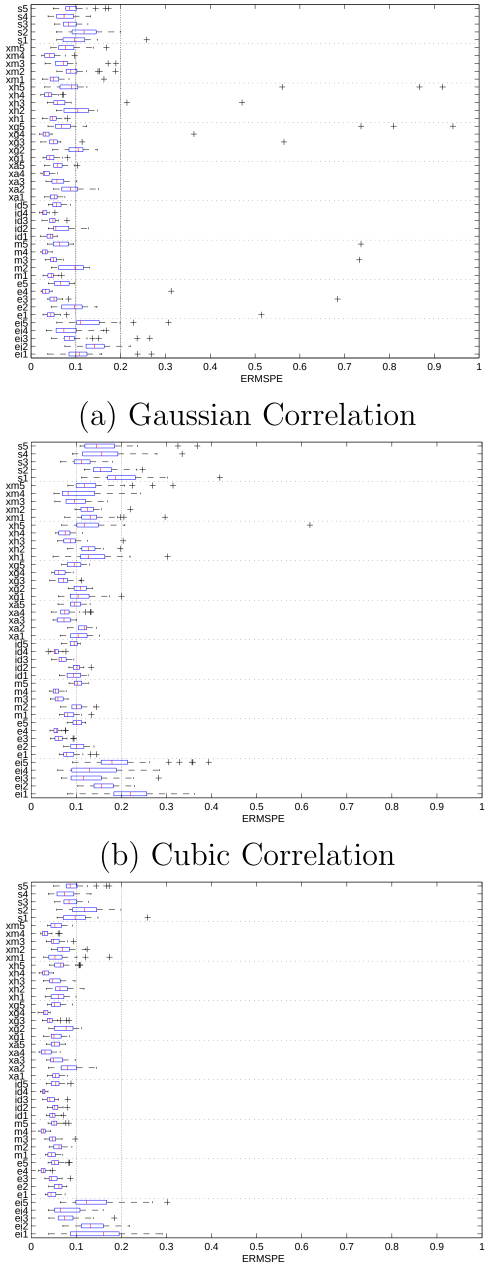
3.5 Conclusion

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) suggest similar results will occur if other fixed-point designs are used. Among the sequential adaptive designs for the examples considered, the cross validation prediction error criterion with the arithmetic mean and geometric mean (both with 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. These three 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 observations. Subsection 3.4.6 shows potential gains in efficiency (in terms of number sampled points) from using sequential designs, particularly the cross validation prediction error criterion with the arithmetic mean and the geometric mean considered in our example, over the fixed-point maximin LHD.

More examples are presented in Chapter 4 to compare the predictive performances of the sequential adaptive designs for *non-stationary* looking response surfaces, while a single stationary GASP model is still fitted across the entire input space.

Other issues do arise during the implementation of these sequential designs. For instance in Subsection 3.4.1, it is clear that the estimation of correlation parameters has a huge impact on the prediction accuracy using the GASP model. Another key issue in sequential design is the number of starting design points (Subsection 3.4.5). This is crucial to the success of the sequential adaptive designs in surface predictions. Being the three 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 and geometric are 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 view of these issues, Chapter 5 presents a newly proposed method for improving the estimation of the correlation parameters which, in turn, would lead to a more accurate predictive GASP model. This modified estimation approach also provides measures for assessing model fit as well as a guide for stopping the design algorithm.

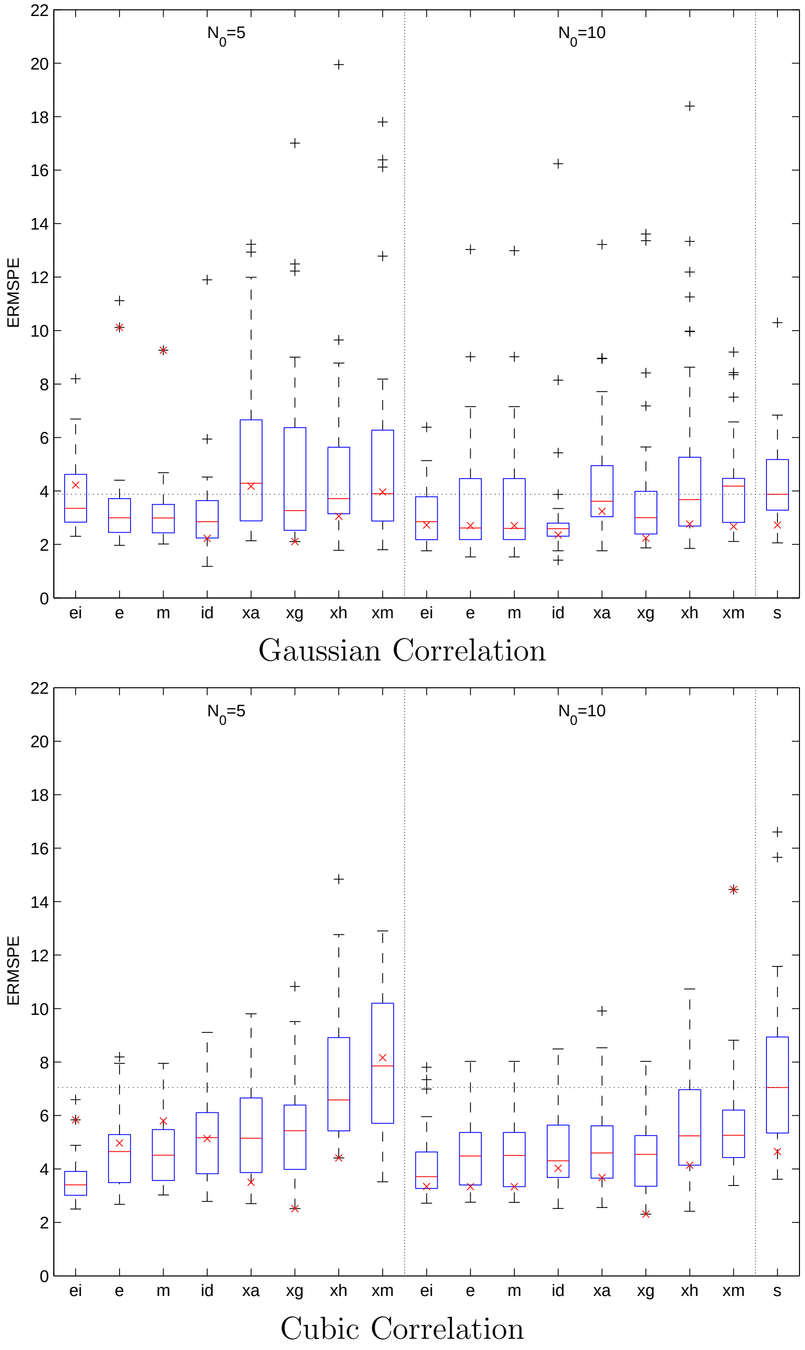


(a) Gaussian Correlation

(b) Cubic Correlation

(c) Gaussian Correlation, $\boldsymbol{\theta} = (5, 5)$ assumed known

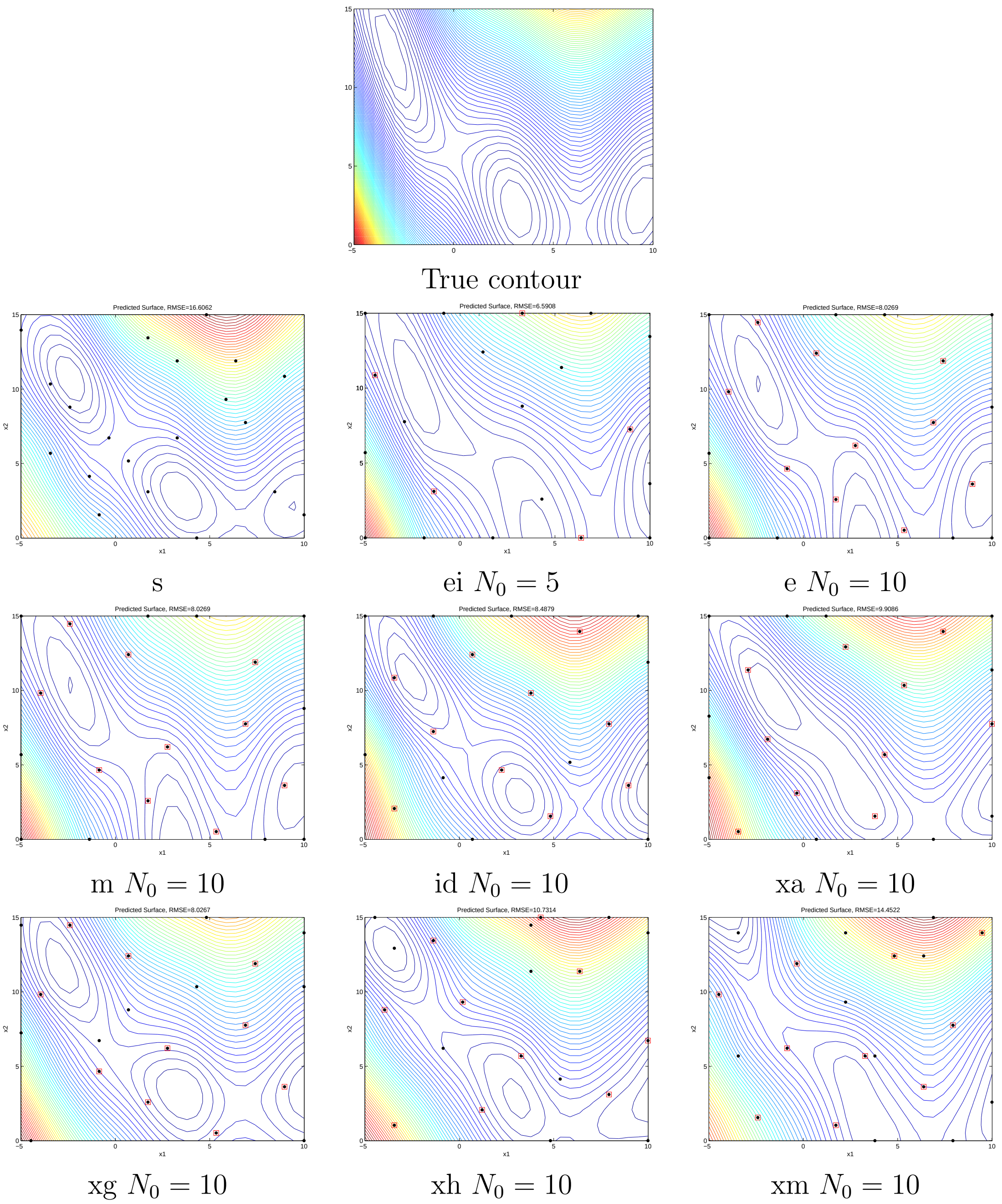
Figure 3.3: (GASP surfaces: G1 to G5) Boxplots of the empirical root mean squared prediction error (ERMSPE) for 30 different starting designs and Gaussian/cubic correlation functions - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error 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). Vertical axis labels represent design and surface (e.g., ei1 denotes design points selected using EIGF criterion for surface G1). The number of starting input points is $N_0 = 5$ for the sequential methods. A total of $N = 30$ design points are selected in all the cases.



Gaussian Correlation

Cubic Correlation

Figure 3.4: (Function M1, Branin 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. (×) denotes ERMSPE for the best (starting) maximin LHD (see page 51). A total of $N = 30$ design points are selected in all the cases.



True contour

s

ei $N_0 = 5$

e $N_0 = 10$

m $N_0 = 10$

id $N_0 = 10$

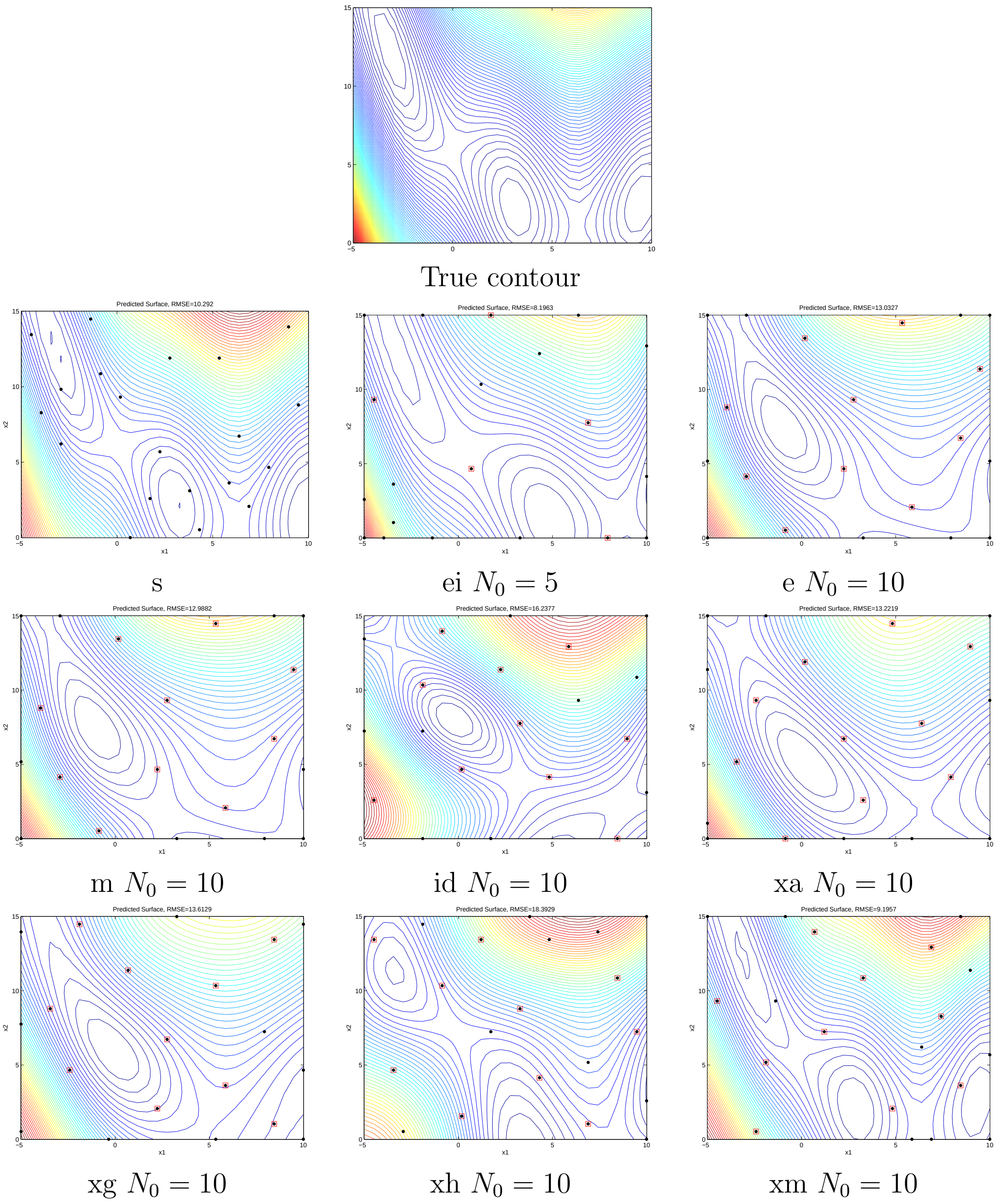
xa $N_0 = 10$

xg $N_0 = 10$

xh $N_0 = 10$

xm $N_0 = 10$

Figure 3.5: (Function M1, Branin function) Worst case prediction with maximum ERMSPE among 30 different starting designs and cubic correlation - 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 design points for the sequential methods. A total of $N = 20$ design points are selected in all the cases. N_0 denotes number of starting design points.



True contour

s

ei $N_0 = 5$

e $N_0 = 10$

m $N_0 = 10$

id $N_0 = 10$

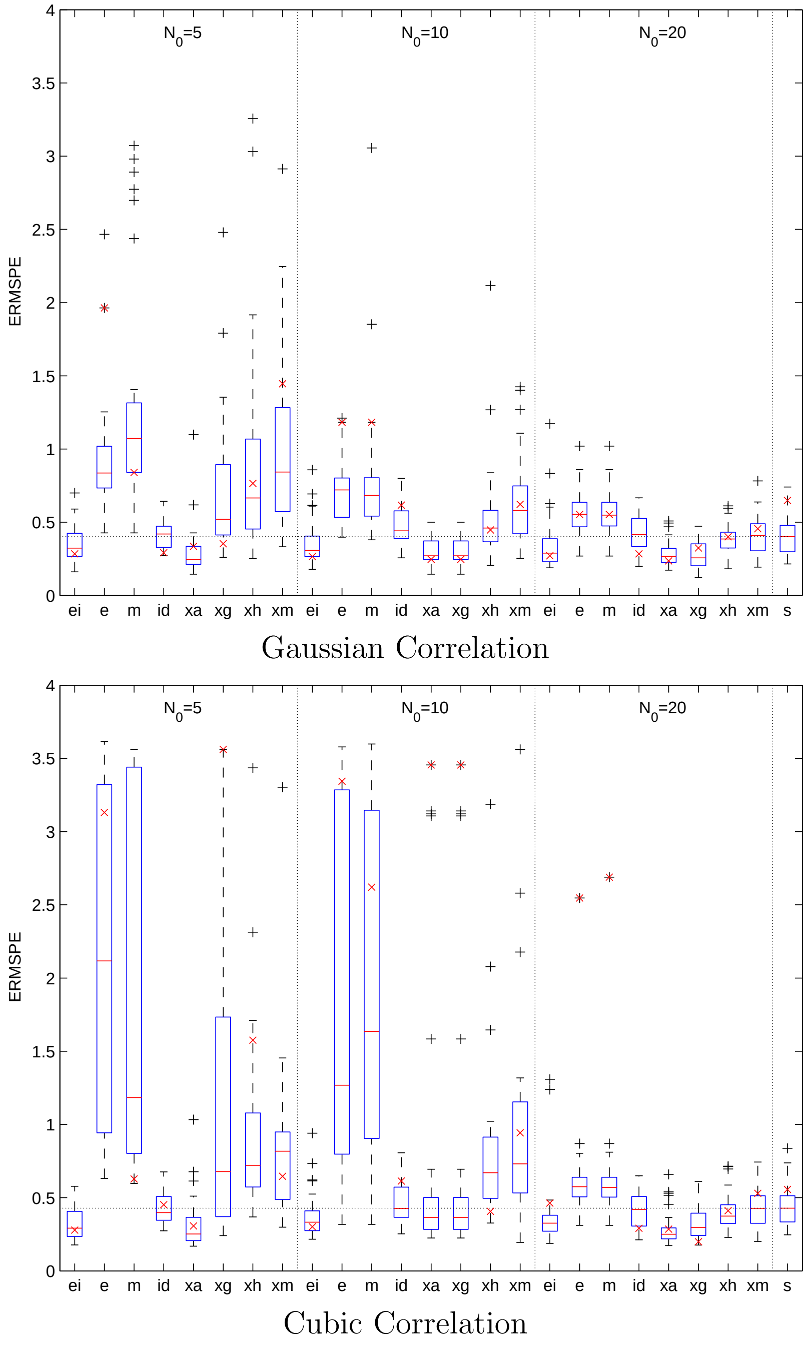
xa $N_0 = 10$

xg $N_0 = 10$

xh $N_0 = 10$

xm $N_0 = 10$

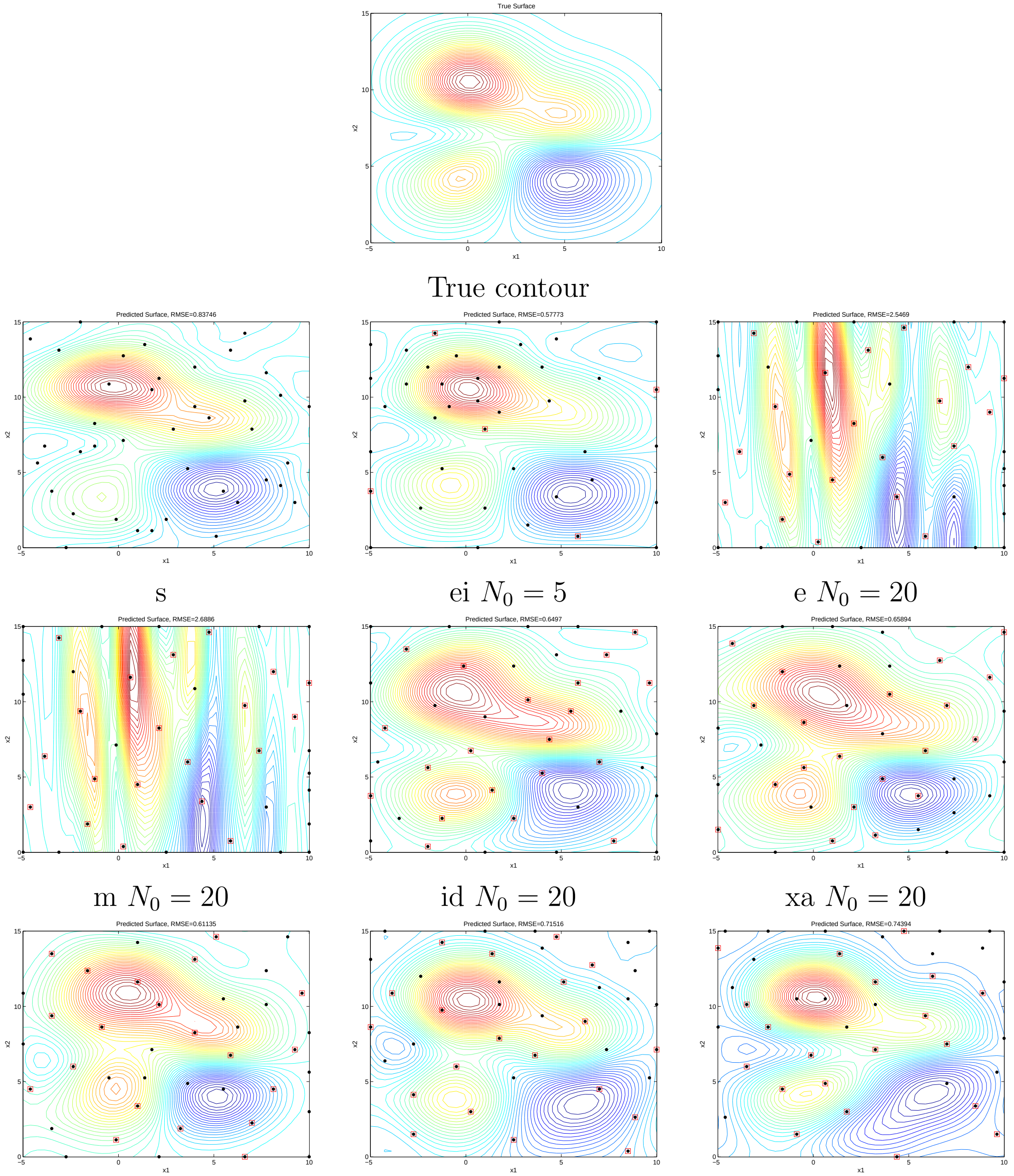
Figure 3.6: (Function M1, Branin function) Worst case prediction with maximum ERMSPE among 30 different starting designs and Gaussian correlation - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross validation prediction 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 design points for the sequential methods. A total of $N = 20$ design points are selected in all the cases. N_0 denotes number of starting design points.



Gaussian Correlation

Cubic Correlation

Figure 3.7: (Function M2, 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. (×) denotes ERMSPE for the best (starting) maximin LHD (see page 51). A total of $N = 40$ design points are selected in all the cases.



True contour

s

ei $N_0 = 5$

e $N_0 = 20$

m $N_0 = 20$

id $N_0 = 20$

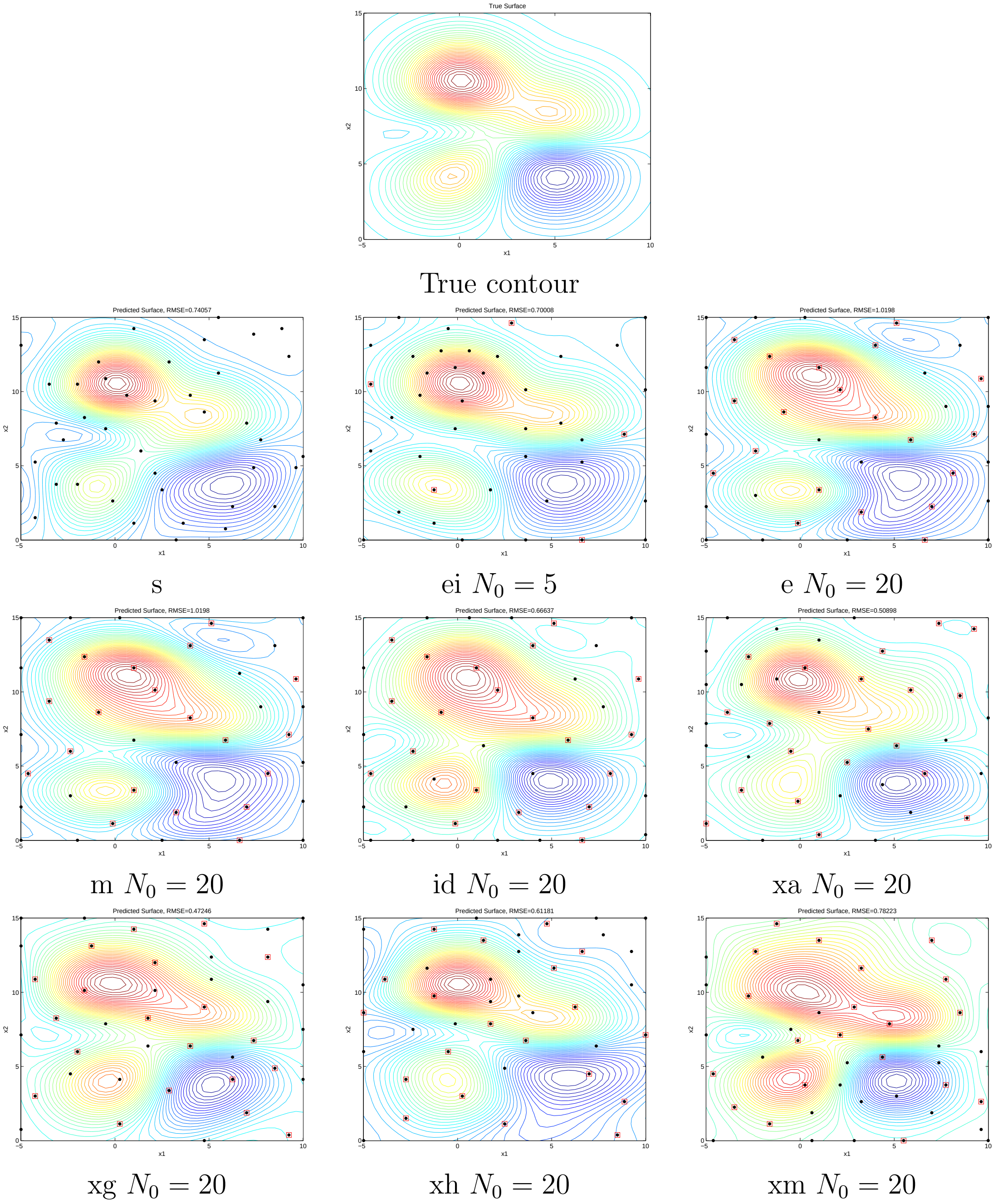
xa $N_0 = 20$

xg $N_0 = 20$

xh $N_0 = 20$

xm $N_0 = 20$

Figure 3.8: (Function M2, Simulated surface) Worst case prediction with maximum ERMSPE among 30 different starting designs and cubic correlation - 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 design points for the sequential methods. A total of $N = 40$ design points are selected in all the cases. N_0 denotes number of starting design points.



True contour

s

ei $N_0 = 5$

e $N_0 = 20$

m $N_0 = 20$

id $N_0 = 20$

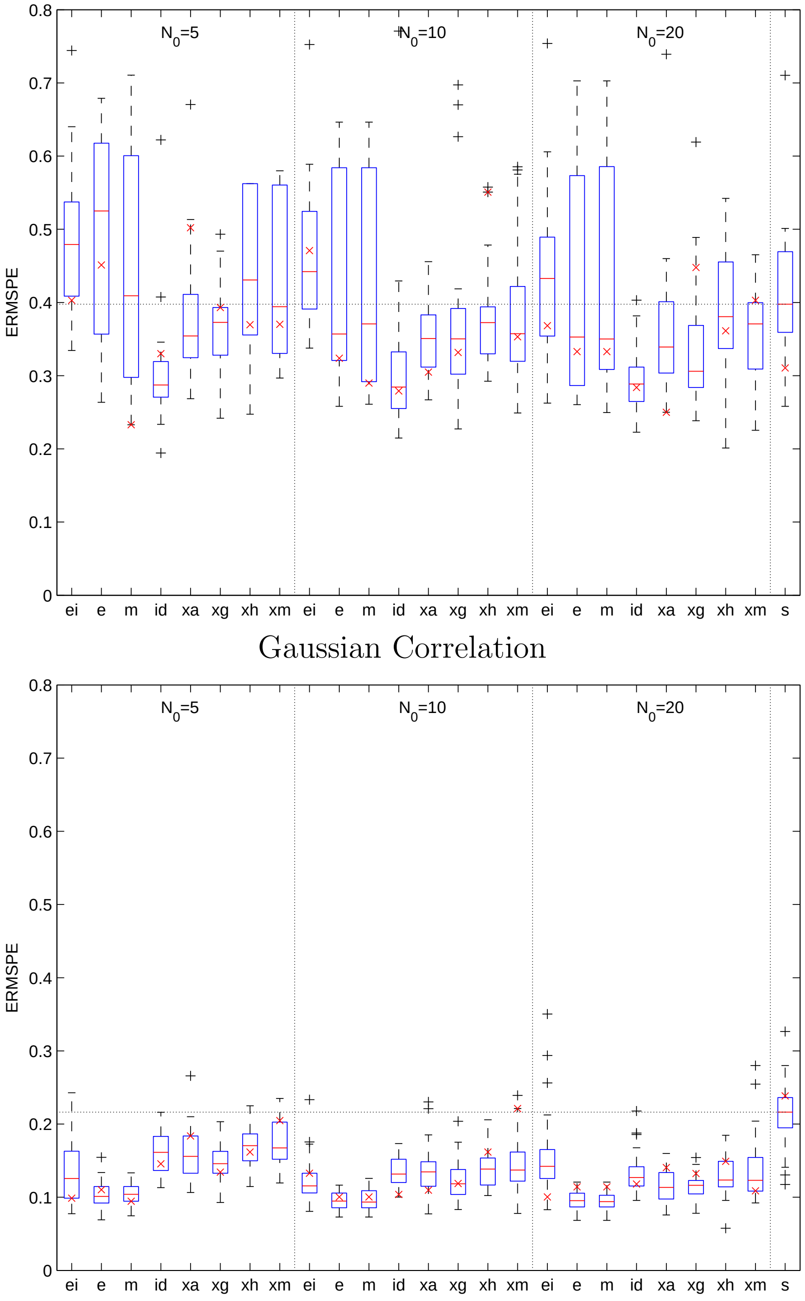
xa $N_0 = 20$

xg $N_0 = 20$

xh $N_0 = 20$

xm $N_0 = 20$

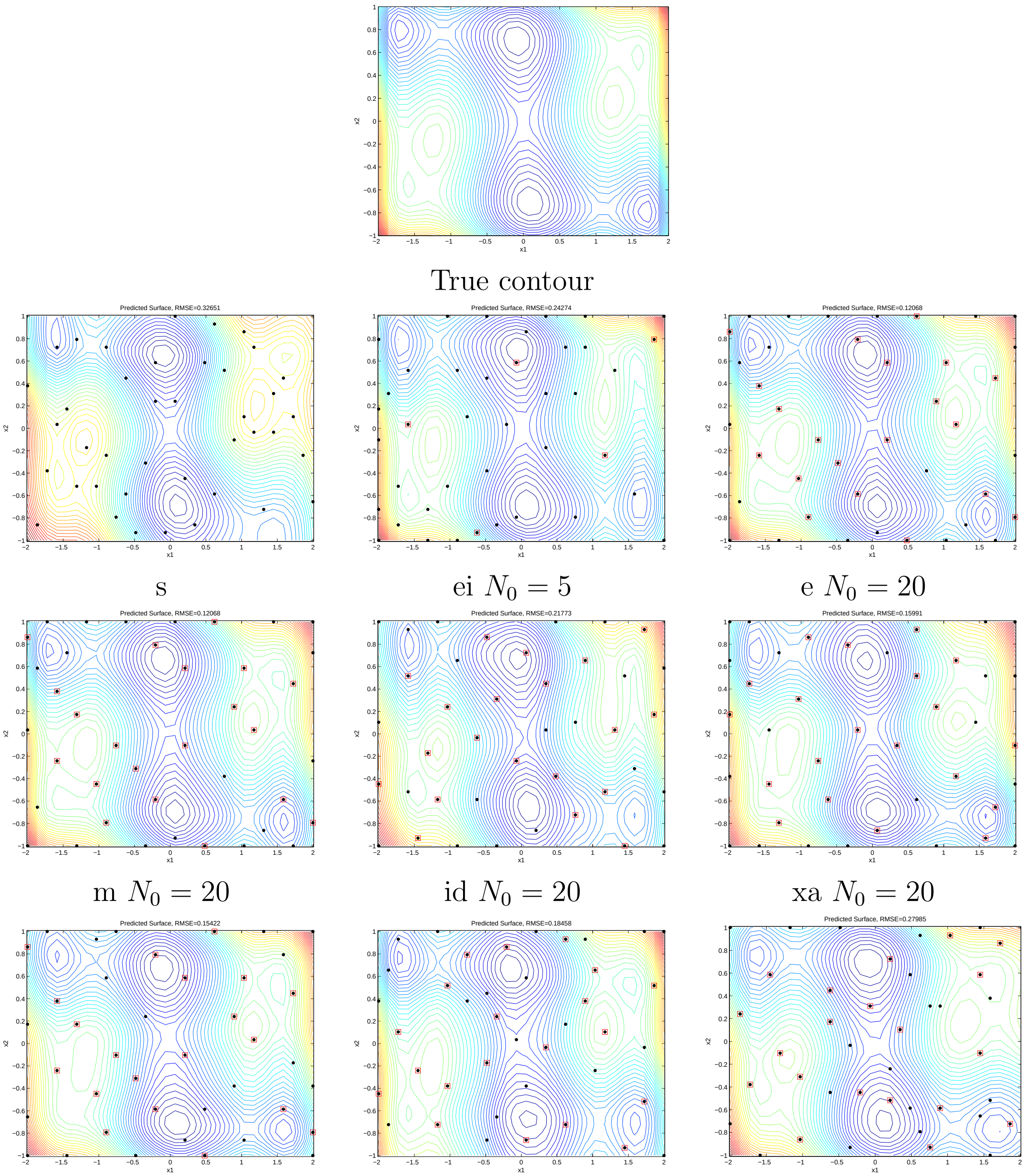
Figure 3.9: (Function M2, Simulated surface) Worst case prediction with maximum ERMSPE among 30 different starting designs and Gaussian correlation - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross validation 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 design points for the sequential methods. A total of $N = 40$ design points are selected in all the cases. N_0 denotes number of starting design points.



Gaussian Correlation

Cubic Correlation

Figure 3.10: (Function M3, 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. (×) denotes ERMSPE for the best (starting) maximin LHD (see page 51). A total of $N = 40$ design points are selected in all the cases.



True contour

s

ei $N_0 = 5$

e $N_0 = 20$

m $N_0 = 20$

id $N_0 = 20$

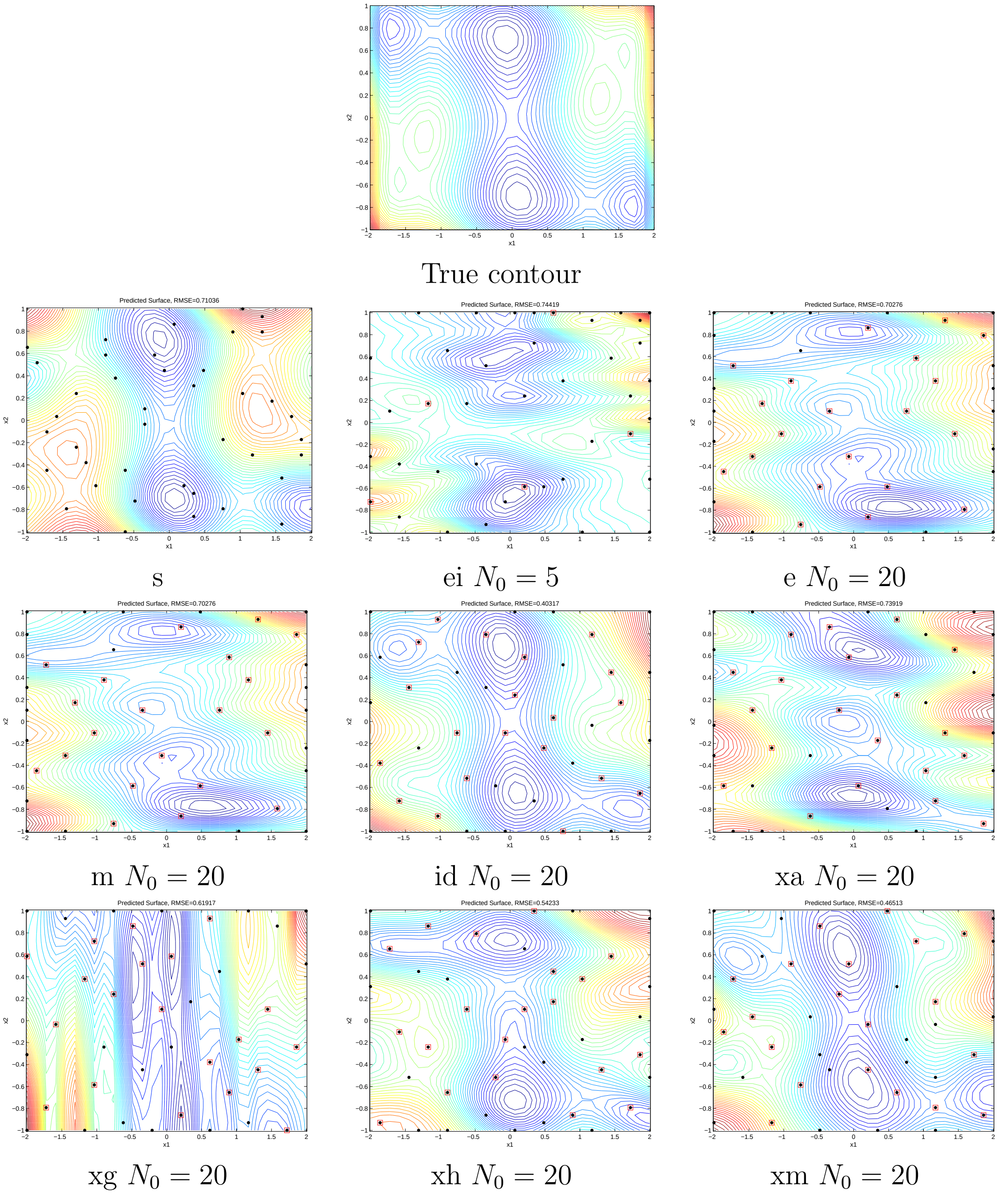
xa $N_0 = 20$

xg $N_0 = 20$

xh $N_0 = 20$

xm $N_0 = 20$

Figure 3.11: (Function M3, Six-hump camel-back function) Worst case prediction with maximum ERMSPE among 30 different starting designs and cubic correlation - 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 design points for the sequential methods. A total of $N = 40$ design points are selected in all the cases. N_0 denotes number of starting design points.



True contour

s

ei $N_0 = 5$

e $N_0 = 20$

m $N_0 = 20$

id $N_0 = 20$

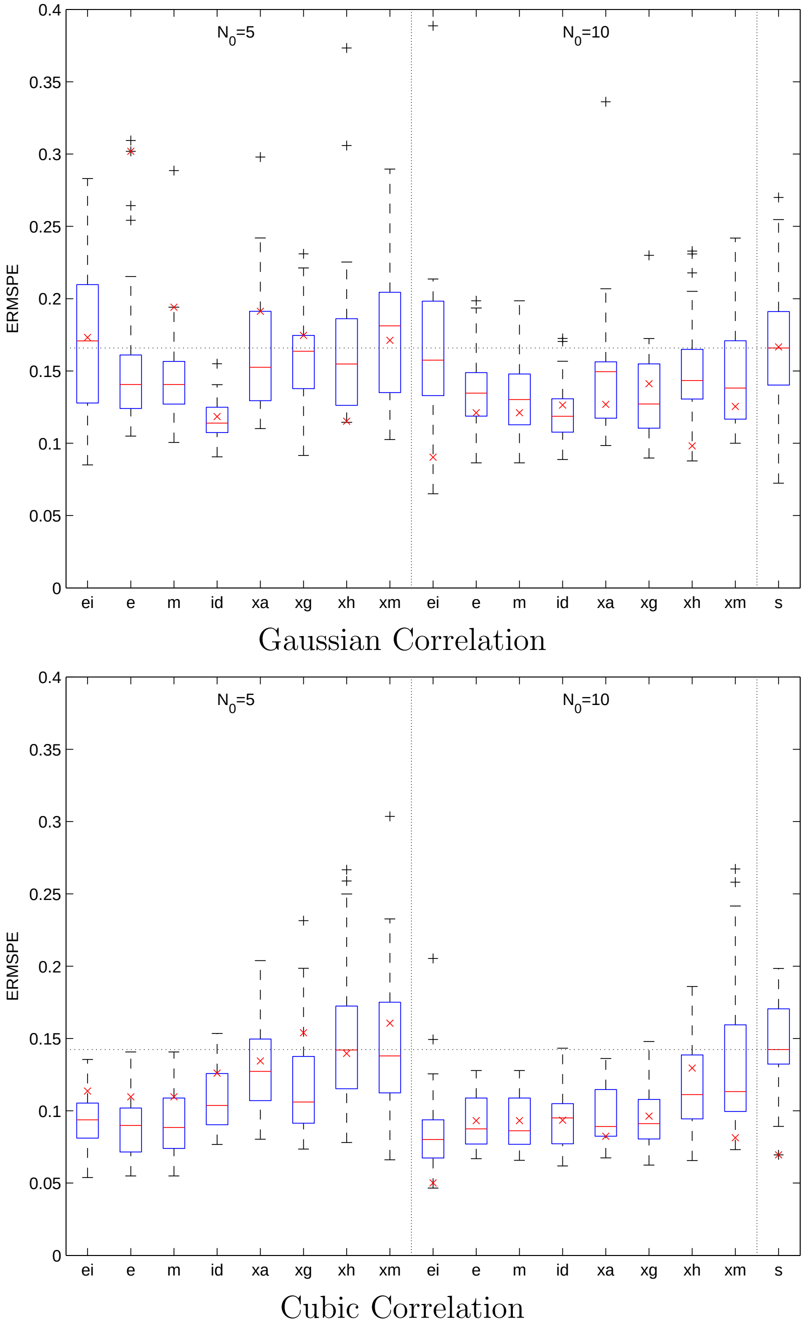
xa $N_0 = 20$

xg $N_0 = 20$

xh $N_0 = 20$

xm $N_0 = 20$

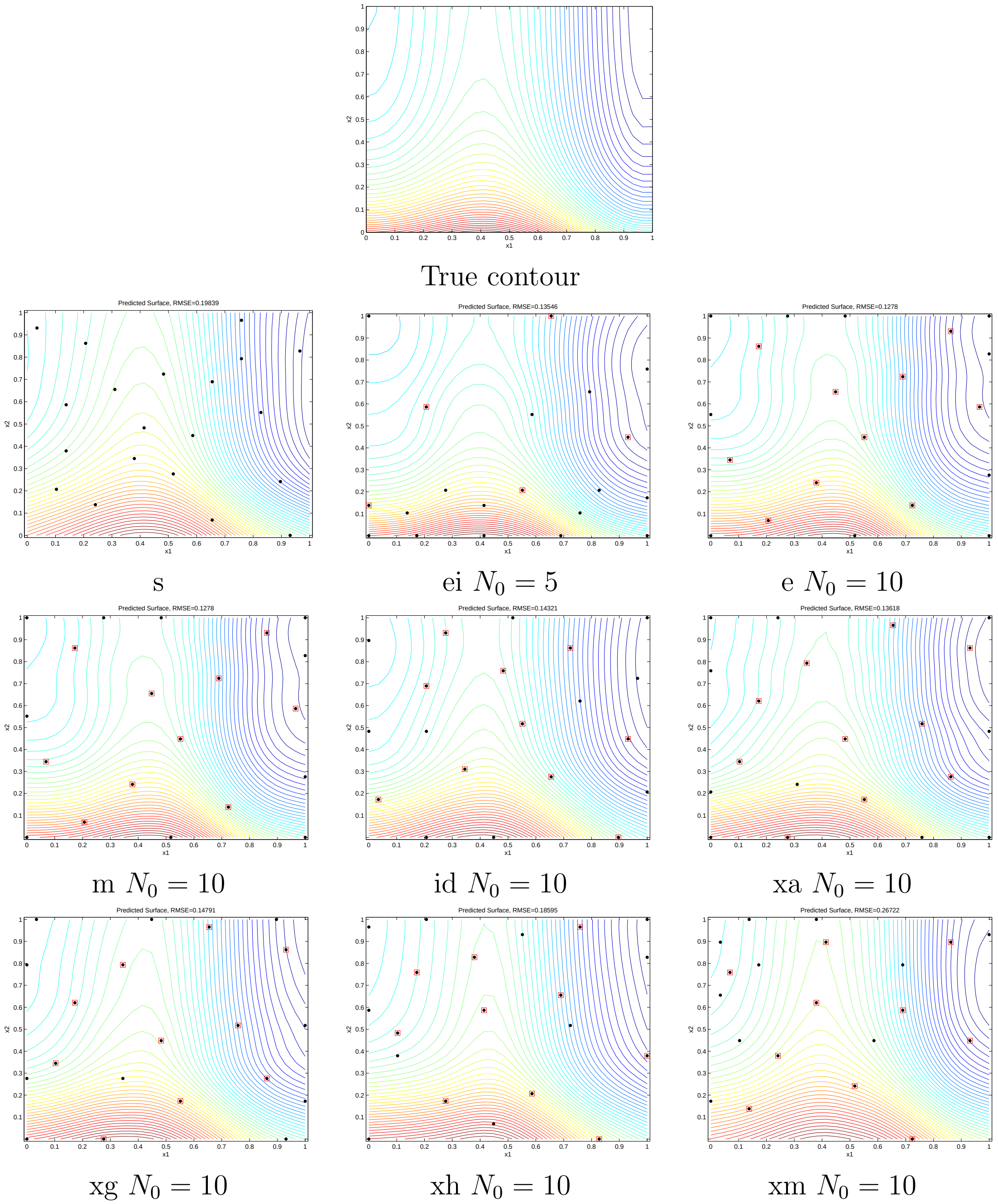
Figure 3.12: (Function M3, Six-hump camel-back function) Worst case prediction with maximum ERMSPE among 30 different starting designs and Gaussian correlation - EIGF (ei), maximum entropy (e), maximum mean squared prediction error (m), integrated mean squared prediction error with penalty (id), cross 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 design points for the sequential methods. A total of $N = 40$ design points are selected in all the cases. N_0 denotes number of starting design points.



Gaussian Correlation

Cubic Correlation

Figure 3.13: (Function M4, Non-polynomial 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. (×) denotes ERMSPE for the best (starting) maximin LHD (see page 51). A total of $N = 20$ design points are selected in all the cases.



True contour

s

ei $N_0 = 5$

e $N_0 = 10$

m $N_0 = 10$

id $N_0 = 10$

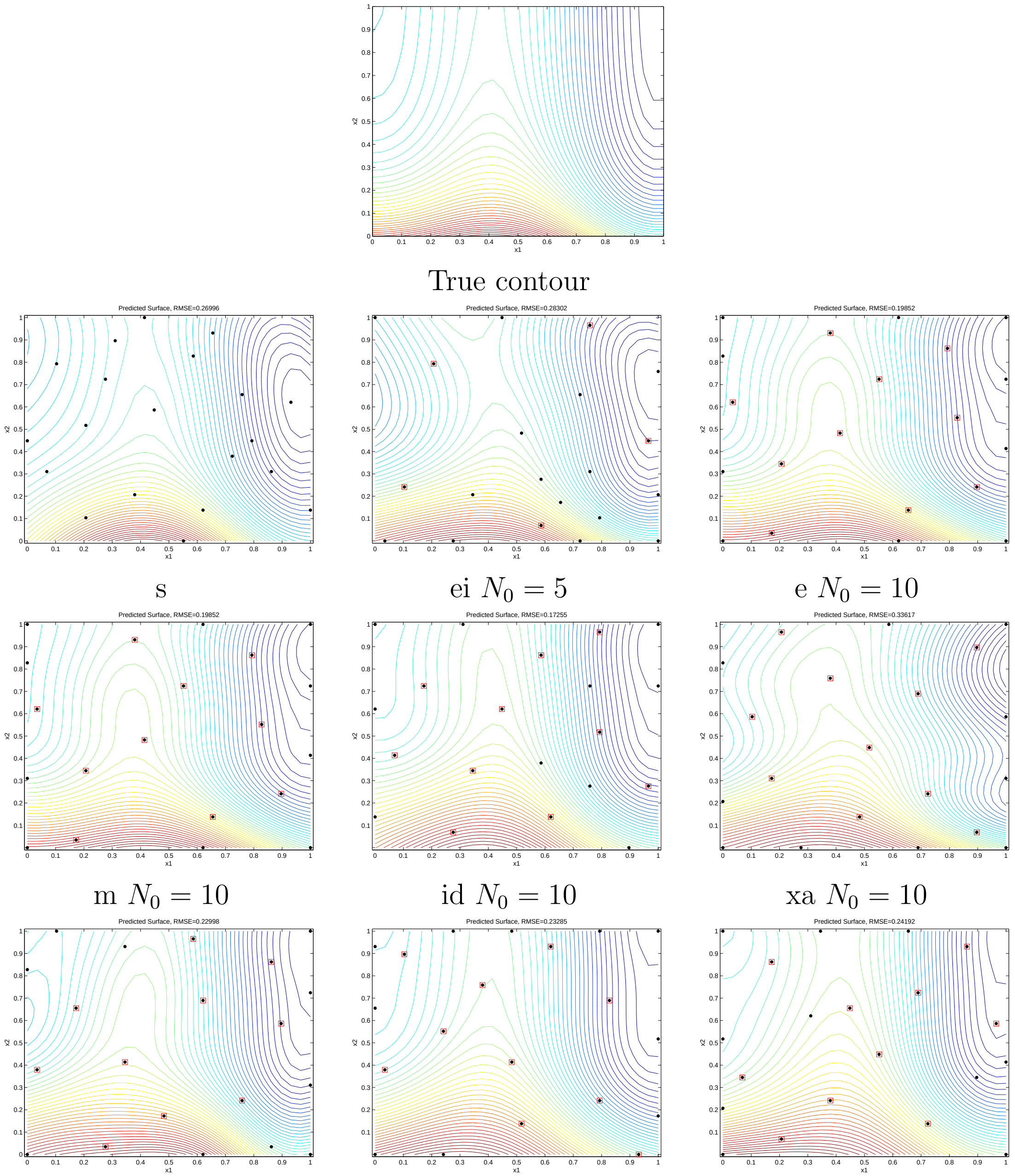
xa $N_0 = 10$

xg $N_0 = 10$

xh $N_0 = 10$

xm $N_0 = 10$

Figure 3.14: (Function M4, Non-polynomial surface) Worst case prediction with maximum ERMSPE among 30 different starting designs and cubic correlation - 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 design points for the sequential methods. A total of $N = 20$ design points are selected in all the cases. N_0 denotes number of starting design points.



True contour

s

ei $N_0 = 5$

e $N_0 = 10$

m $N_0 = 10$

id $N_0 = 10$

xa $N_0 = 10$

xg $N_0 = 10$

xh $N_0 = 10$

xm $N_0 = 10$

Figure 3.15: (Function M4, Non-polynomial surface) Worst case prediction with maximum ERMSPE among 30 different starting designs and Gaussian correlation - 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 design points for the sequential methods. A total of $N = 20$ design points are selected in all the cases. N_0 denotes number of starting design points.

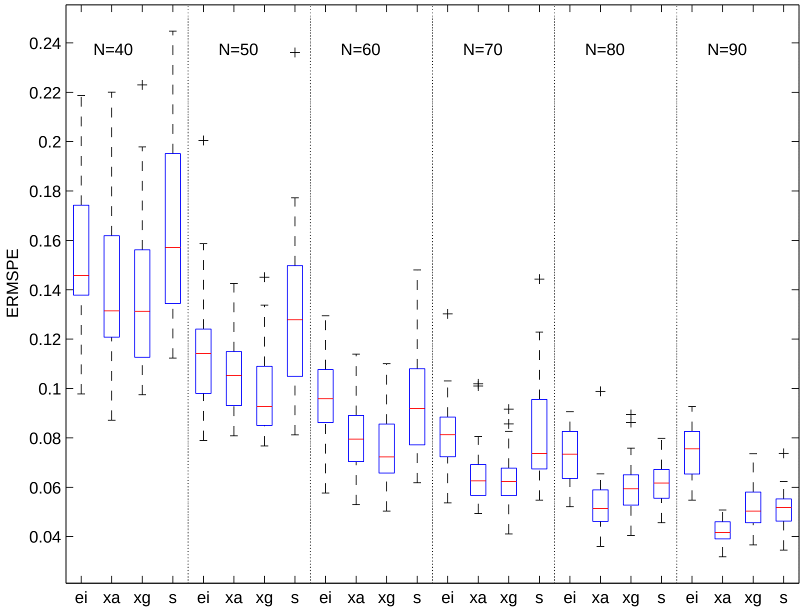


Figure 3.16: (Function M5, Hartman 3 function) Boxplots of the empirical root mean squared prediction error (ERMSPE) for 30 different starting designs and cubic correlation functions - EIGF (ei), cross validation prediction error (using arithmetic mean (xa), geometric mean (xg)), fixed-point maximin LHD (s). The number of starting input points for the sequential methods is $N_0 = 30$. N denotes the final number of design points.

CHAPTER 4

SEQUENTIAL ADAPTIVE DESIGNS FOR FITTING NON-STATIONARY RESPONSE SURFACES

4.1 Introduction

Much of the recent work in the design and analysis of computer experiments has involved global optimization or achievement of a good response surface model fit. Most, if not all, have adopted the approach of assuming a single stationary Gaussian Stochastic Process (GASP) model across the entire input space in the design and analysis stages. Because they are based on using a stationary model, application of the various designs from the previous chapters may not be effective when the surface appears to be highly non-stationary. These designs may not be able to target regions with high variation in the response and hence may suffer in terms of prediction ability.

In this chapter, we investigate a sequential adaptive design, the expected improvement for global fit (EIGF) criterion, for non-stationary looking surface model fit. This work is in part motivated by the work of Gramacy (2005) where he developed the Bayesian Treed Gaussian process as a surrogate model for fitting non-stationary response surfaces.

In Section 4.2, we present the EIGF criterion and illustrate its implementation in a simple one-dimensional example. Following this in Section 4.3, an example is given to compare the EIGF criterion and other designs from Chapter 3. A generalization of the EIGF criterion is considered in Section 4.4. We conclude with a discussion of the predictive performance of the various designs based on our simulations.

A brief review of two studies is now given to highlight recent work in fitting GASP models for non-stationary looking response surfaces in computer experiments. We view these studies as approaching the problem of fitting a GASP model to a non-stationary looking surface in two different perspectives: (1) modeling, versus (2) design.

(1) Modeling: Gramacy (2005) and Gramacy and Lee (2006) proposed the Bayesian Treed Gaussian process model as a general methodology for fitting non-stationary response surfaces using the GASP model and they also discussed strategies for experimental design. Their work involved three main stages: (i) use of trees and recursive partitioning to identify distinctive subregions of the input space where separate stationary GASP models can be fitted within each partition. Partitioning was done by making binary splits on the value of a single input variable so that partition boundaries were parallel to coordinate axes. The partitioning was recursive such that each new partition became a sub-partition within the previous one. They imposed a requirement that there were at least five data points in each new subregion. A Bayesian approach was adopted using the Classication and Regression Trees (CART) methodology (see Breiman *et al.*, 1984, Chipman *et al.*, 1998 and Denison *et al.*, 1998). (ii) fitted a stationary GASP model within each partition of the input space. A Gaussian correlation function was used and estimation of the correlation parameters was carried out using a Markov chain Monte Carlo (MCMC) approach. (iii) specified an adaptive sampling design to guide the choice of the next design point to be added. They took a two-stage approach to selecting the next design point. First, a set of candidate points was selected by an optimal design (such as D-optimal, maximin or Latin hypercube). Given that their design scheme was to be implemented in an asynchronous parallel supercomputing environment, they mentioned that having a dense grid for candidate points will lead to clumping of design points. As a solution, they proposed to select a subset of well-spread out candidate points using a sequential D-optimal design. Next in the second stage, the design point was chosen from the subset of candidate points via the ALC (Active Learning-Cohn) or ALM (Active Learning-McKay) algorithm from the machine learning literature. These criteria appear to be similar to the IMSPE and MMSPE criteria described in Subsection 2.2.1.

(2) Design: In another study by Farhang-Mehr and Azarm (2005), the authors were interested in fitting a GASP model (also called meta-model) that predicted the response surface at unobserved sites everywhere in the input space by incorporating an adjustment factor (into the covariance function) to take into account irregularity in the response of the surface. Their goal was to identify the subregions in which the correlation decayed faster with distance due to the high variation in the response in those subregions.

The basic idea behind their approach is as follows: (i) they started with a small initial set of design points selected using the maximum entropy criterion (2.6) where diffuse prior distributions were assumed for the parameters, (ii) the responses at these sampled points were obtained from running the computer code, (iii) a GASP model with the Gaussian correlation function was fitted using the sampled points to predict the response $y(\boldsymbol{x}_0)$ at same untried $\boldsymbol{x}_0$, (iv) the set of input points associated with all local optima based on the fitted GASP model were obtained denoted as P, (v) for each candidate point, $\boldsymbol{x}$, a characteristic certainty width (CCW), $L(\boldsymbol{x})$, was computed as the length of the diagonal of the smallest hyper-rectangle with its vertices formed by two of the input points in P. A large adjustment factor, $\left(\frac{L(\boldsymbol{x})}{L_0}\right)$, where L_0 denotes the diagonal distance of the entire rectangular input space, implies a relatively flat region. This adjustment factor will be small if $\boldsymbol{x}$ lies in region where the response is highly varying, i.e. there are two optima close to $\boldsymbol{x}$ in distance. Instead of maximizing the determinant of $\sigma_z^2\boldsymbol{R}$ in (2.8), the criterion chooses the point that maximizes $\sigma_z^2\left(\frac{L(\boldsymbol{x})}{L_0}\right)\left(\frac{L(\boldsymbol{x}_j)}{L_0}\right)\boldsymbol{R}$ where $\boldsymbol{x}_j$ denotes an existing sampled point, as the next design point. The modifed maximum entropy criterion will concentrate its sampling effort in regions where candidate points have larger adjustment factors.

Both studies have claimed that their approaches performed well in identifying regions with high variation in the response for non-stationary looking surfaces and had targeted their sampling efforts in these regions sufficiently well to obtain a good fit of the GASP model.

4.2 Sequential Adaptive Design for Non-stationary Looking Response Surfaces

Results from Farhang-Mehr and Azarm (2005) suggest that the identification of regions of high variation in the response provides a good indication of where the sampling effort should be concentrated. Drawing from this idea, we modify the EI criterion by Schonlau (1997) to search for input points where the response varies significantly instead of searching for the global optimum or optima, so as to achieve a fit of the GASP model. Unlike Gramacy (2005), we continue to specify a single stationary GASP model across the entire input space of a clearly non-stationary surface, and avoid the need for partitioning of the input space. Our rationale is that although fitting a separate stationary GASP model within each distinct partition of the input space may give more accurate predictions, it is not clear that this can be achieved using a small number of samples — this is one of the key challenges in designs for computer experiments.

4.2.1 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

$$I(\boldsymbol{x}) = (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 \tag{4.1}$$

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 $\boldsymbol{x}$ that maximizes the expected improvement

$$E(I(\boldsymbol{x})) = (\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 + var(\hat{Y}(\boldsymbol{x})). \tag{4.2}$$

The expected improvement in (4.2) 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 (1.8), i.e., points about which there is large uncertainty and, as mentioned in Subsection 2.3.1, these tend to be far from existing sampled points. The EIGF algorithm follows Subsection 3.1.3 and its derivation is given in Appendix A.3.

4.2.2 Characteristics of EIGF Criterion

Suppose we start the EIGF algorithm with N_0 points and a fine regular grid of candidate points is laid across the input space. The expected improvement in (4.2) can be computed for each point in the grid. If the predicted response surface (with the GASP model) using the N points is smooth, the first (local) component of the expected improvement in (4.2) tends to be larger for candidate points close to the midpoint (in Euclidean distance) of any two existing design points. This always happens if the predicted response function is monotone. For non-monotone surfaces, this will also happen unless there is a significant optimum at a candidate point that is not one of the midpoints. For example, the top left plot in Figure 4.1 (page 91) shows additional point, labeled as 1, is close to the midpoint of two existing sampled points. However, the final design point to be selected is also affected by the magnitude of the *global* component, $\mathrm{var}(Y)$. This will prevent design points from clumping in areas with steep gradients.

This "close to midpoint" feature of the EIGF criterion seems to make it more robust with smaller starting designs. If we start the algorithm with a small number of points from a maximin LHD, for example, the predicted response surface is typically smooth with minimal variation in the reponse. As a result, the EIGF criterion tends to spread out the input points to the midpoint of the two existing sampled points (in one dimension) and produce a crude "space-filling" design. In higher dimensions, the EIGF criterion still mantains this nice property of spreading out points although the notion of "midpoint" does not generalize to higher dimensions. The merit of this feature is that the EIGF criterion does not suffer as much with problems in estimating the correlation parameters which in turn negatively impact the predictive performance of the other sequential criteria, namely the MSPE, maximum entropy and cross validation. It is clear from the examples in Chapter 3 that the other sequential designs perform better with larger starting designs.

4.2.3 Illustration of EIGF Criterion

In this subsection, we present a detailed illustration of the EIGF criterion and evaluate its predictive performance as more input points are added.

Function N1: One-dimensional non-stationary sine-cosine function

We use a simple one-dimensional function taken from Gramacy (2005),

$$y(x) = \begin{cases} \sin\left(\frac{\pi x}{5}\right) + \frac{1}{5}\cos\left(\frac{4\pi x}{5}\right) & x < 10 \\ x/10 - .8 & \text{otherwise.} \end{cases} \tag{4.3}$$

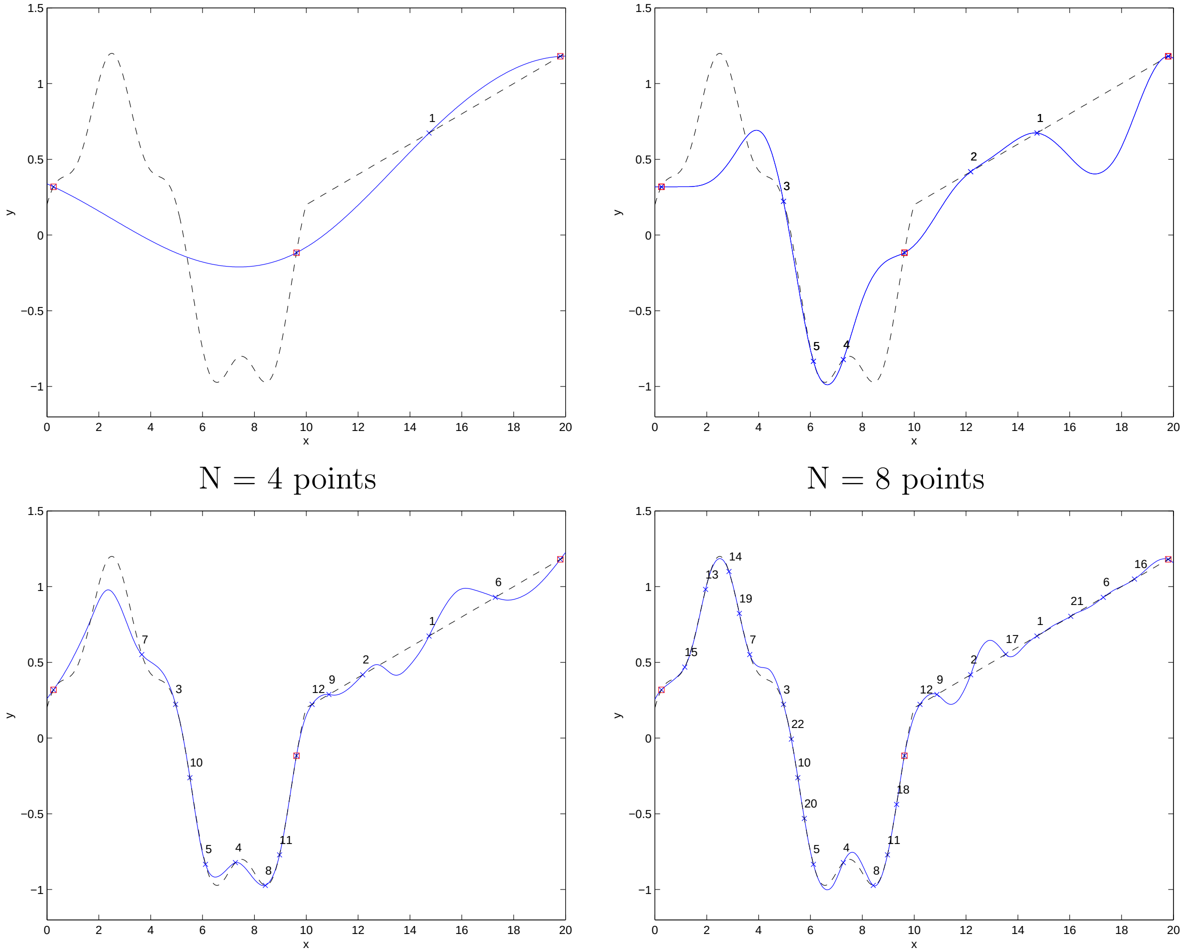
A plot of the true function is shown in Figure 4.1. Unlike in Gramacy (2005), we do not add noise to the response. There is a clear change point in the response at around $x = 10$.

We illustrate the implementation of the EIGF criterion in detail. Suppose we have $N_0 = 3$ observations from the computer simulator (4.3) and the points are denoted by the □ in Figure 4.1. The points are generated from one run of our software using the maximin LHD. For this run, the three initial points are close to the two endpoints and the midpoint of the input space x. Figure 4.1 shows a few iterations of the algorithm with plots of the true and predicted functions based on $N = 4, 8, 15$ and 25 input points.

The first plot (top left) shows the true function (dashed line) and predicted function (solid line) using the GASP model with 4 input points. The first point added is labeled as 1. Notice that this point is located at about the midpoint of two existing design points. The second plot (top right) shows the predicted function with 8 points and the response varies significantly for the region where $x < 10$. As the number of input points increases, the EIGF criterion starts to focus its sampling effort in the region to the left of the change point where the function is more irregular. With a total of 15 points (bottom left plot), main features of the function are rather accurately predicted by the GASP model. The last plot (bottom right) in Figure 4.1 shows the EIGF criterion selecting 14 out of the 22 points (not counting the initial 3 points) in the region left of the change point. The overall fit of the predicted function resembles the true function except for the small hump to right of the change point. This example shows that the EIGF criterion is able to target the region where the response varies significantly and produces an accurate prediction of the function. This is achieved with many fewer observations than the Bayesian Treed approach in Gramacy (2005) which required close to 100 observations.

4.3 Test Function: Two-dimensional Non-stationary Looking Response Surface

The following example illustrates the implementation and prediction performace of the various sequential designs (including the EIGF criterion) and the fixed-point maximin LHD. The design strategies to be compared are:



N = 4 points

N = 8 points

N = 15 points

N = 25 points

Figure 4.1: Plots of true and predicted curve for Function N1 (One-dimensional sine-cosine function) using EIGF criterion. True function (dashed lines), predicted function (solid lines). Initial input points (□), remaining added points (×) labeled according to their sequence.

- 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. Due to the complexity of the response surface used in this section, the final number of input points is taken to be 30. Similar to the examples in Section 3.4, our comparisons are based on 30 runs of our software for generating maximin LHDs (MATLAB$^{\copyright}$ codes for implementing the sequential designs presented in this thesis can be found on the department's computer experiment directory at /home/comp_exp/SOFTWARE/EIGF). 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 generated using an N_0-point (nearly) maximin LHD, with N_0 chosen to be 5 and 15.

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),

$$ERMSPE = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}(\boldsymbol{x}_i) - y(\boldsymbol{x}_i))^2}{m}} \tag{4.4}$$

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 at the same set of grid points. 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 spread out over $\mathcal{X}$. Boxplots are used for each of the test functions to show the distribution of the ERMSPE for the 30 runs.

Function N2: 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

$$y = x_1 \exp(-x_1^2 - x_2^2) \tag{4.5}$$

for $x_1, x_2 \in [-2, 6]$. This surface (shown in Figure 4.2) 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 $m = 30 \times 30 = 900$ equally spaced points which coincide with the m points used to evaluate the designs in (4.4). 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 Subsection 4.3.1, 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.

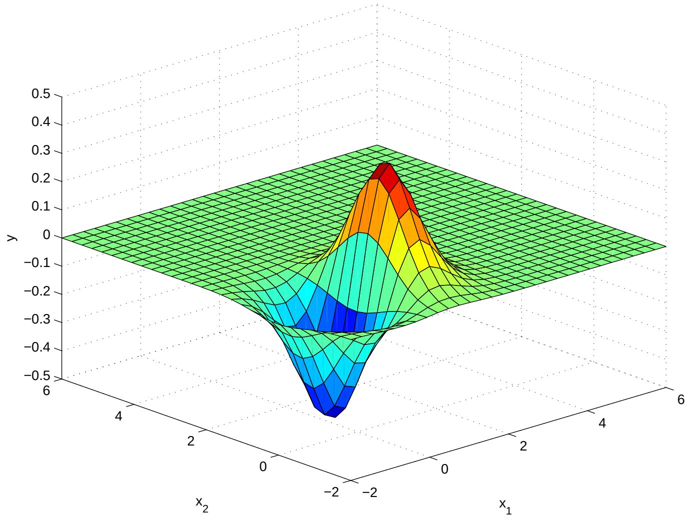


Figure 4.2: Surface plot of true surface for Function N2 (Two-dimensional exponential function)

4.3.1 Results: Two-dimensional exponential function

Boxplots of the ERMSPE for the various designs, in Figure 4.3 on page 100, show similar results for using the Gaussian and cubic correlation.

Here, for both correlation functions, the EIGF criterion (ei) with $N_0 = 5$ stands out as the best (see Figure 4.3). 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 criterion, 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 4.3) 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 4.4 shows comparative plots of predicted surfaces of Function N2 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 4.4.

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 Figure 4.12 in Gramacy, 2005, where the predicted surface with 123 points is shown). 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.

4.4 Generalization of EIGF Criterion

The examples in Sections 3.4, 4.2 and 4.3 have shown that the EIGF criterion is able to target regions with high variation in the response and gives an accurate prediction of the response surface using a single stationary GASP model. However, some of the 30 runs (in Figures 3.3 (page 69) and 3.4 (page 70) for the GASP surfaces and Branin function, respectively) have relatively large ERMSPE.

For the GASP surfaces in Subsection 3.4.1, Figure 4.6 (page 103) shows the true surfaces (left column) and the predicted surfaces from one run of the EIGF criterion with the Gaussian correlation (middle column) for each of the G1-G5 surfaces. Some of the surface features are not predicted well. Another example is shown in Figure 4.7 for the Branin function where responses in the interior region of the input space are not very well predicted using the GASP model with the EIGF criterion and cubic correlation.

4.4.1 Generalized EIGF criterion

We consider an extension of the EIGF criterion to allow for a more global search capability. Following the generalization of the expected improvement criterion in Schonlau (1997) (reviewed on page 36), we consider an additional parameter g for $\mathrm{E}(I^g(\boldsymbol{x}))$ where $g = 2, 4, \cdots$, and recall that the original EIGF criterion is denoted by

E$(I(\boldsymbol{x}))$ with $g = 1$. Suppose we have the computer outputs $y(\boldsymbol{x}_j)$ at sampled points $\boldsymbol{x}_j$, $j = 1, ..., n$. For each candidate input point $\boldsymbol{x}$, its improvement with $g = 2$ for a more global search is defined as

$$E(I^2(\boldsymbol{x})) = (\hat{Y} - y(\boldsymbol{x}_{j^*}))^4 + \ 6\ (\hat{Y} - y(\boldsymbol{x}_{j^*}))^2\ [var(\hat{Y}(\boldsymbol{x})]\ +\ 3\ [var(\hat{Y}(\boldsymbol{x})]^2 \quad (4.6)$$

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 modified criterion is to choose the next input point that maximizes the modifed expected improvement (4.6). The derivation of the generalized EIGF criterion is shown in (A.12).

Instead of two components in the original EIGF in (4.2), there is interaction between the change in the response, $(\hat{Y} - y(\boldsymbol{x}_{j^*}))$, and the prediction error term $var(\hat{Y}(\boldsymbol{x}))$ in (4.6). This drives the search to be more global by giving more weight to the prediction uncertainty component.

4.4.2 Examples: Generalized EIGF Criterion

GASP model surfaces: (Subsection 3.4.1) The plots in the middle column of Figure 4.6 show the predicted surfaces using the EIGF criterion and Gaussian correlation. Compared to the true surfaces (left column), some finer features of the response surfaces are not predicted well. For example, the predicted G2 surface has an extra hump in the middle of the input space at $(x_1, x_2) = (0.6, 0.3)$.

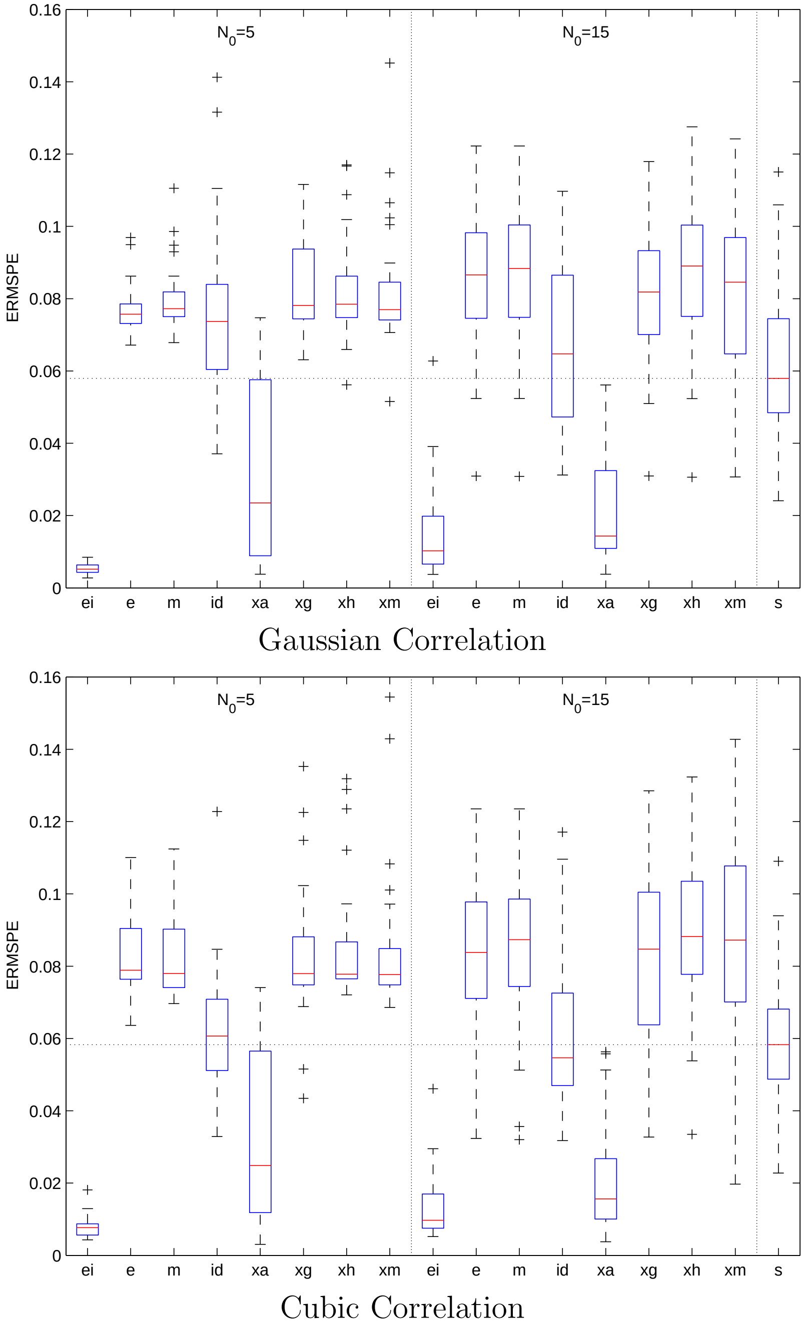
Applying the generalized EIGF criterion (4.6) to the same run of the GASP surfaces, we show the predicted surfaces in the right column of Figure 4.6. Overall, the predicted surfaces looks more like the true surfaces (graphically) compared to those using the EIGF criterion but the ERMSPEs do not neccessarily decrease (compare the ERMSPEs for each row across the middle and right columns in Figure 4.6).

Function M1, Branin function: (Subsection 3.4.3) The true surface and predicted surface using the EIGF criterion (with the cubic correlation) is shown in the left and middle columns of Figure 4.7 respectively. It is obvious that the responses at around $(x_1, x_2) = (4, 3)$ and $(x_1, x_2) = (10, 3)$ are not well predicted with 20 input points. Using the generalized EIGF criterion(4.6), the predicted surface looks more accurate (right column) with a reduction in ERMSPE from 6.59 to 5.84.

4.5 Conclusion

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 thesis. 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.

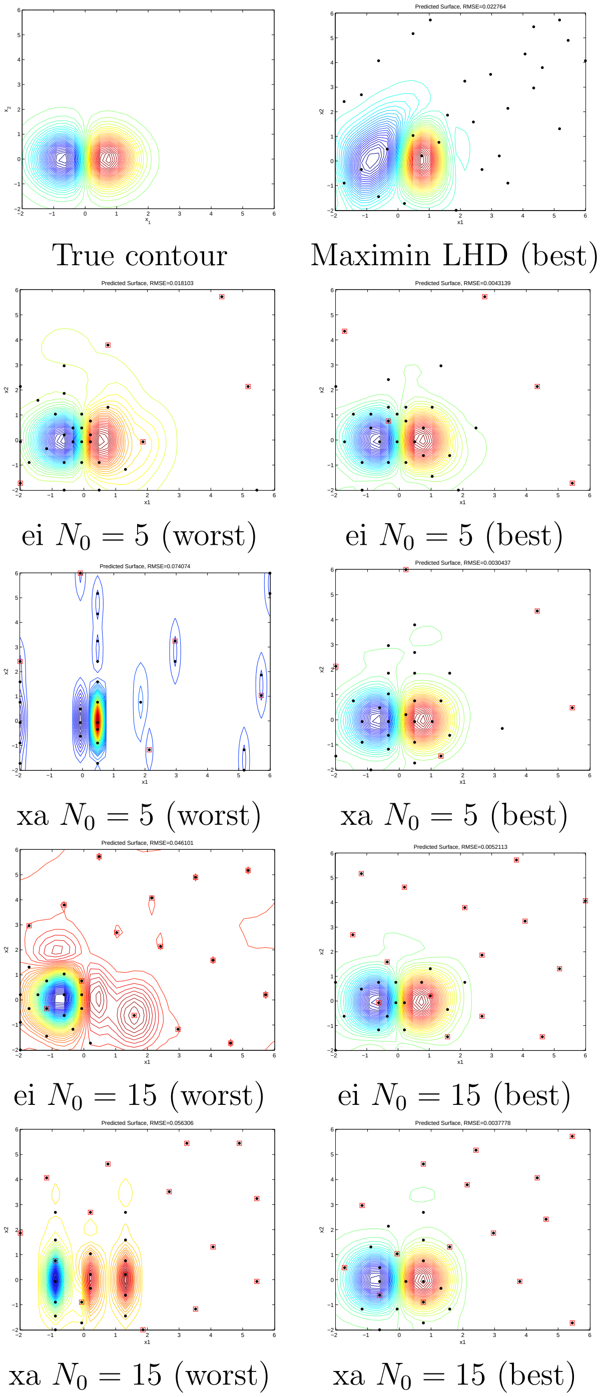
Results from our simulation study shows that the EIGF criterion seems to be the best performer in predicting non-stationary looking response surfaces. Generalization of the EIGF criterion for a more globalized search also shows promising results in that it performs well in examples. Further work might consider other approaches for controlling/varying the degrees of “globalized search” as input points are being added.



Gaussian Correlation

Cubic Correlation

Figure 4.3: (Function N2, 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.



True contour

Maximin LHD (best)

ei $N_0 = 5$ (worst)

ei $N_0 = 5$ (best)

xa $N_0 = 5$ (worst)

xa $N_0 = 5$ (best)

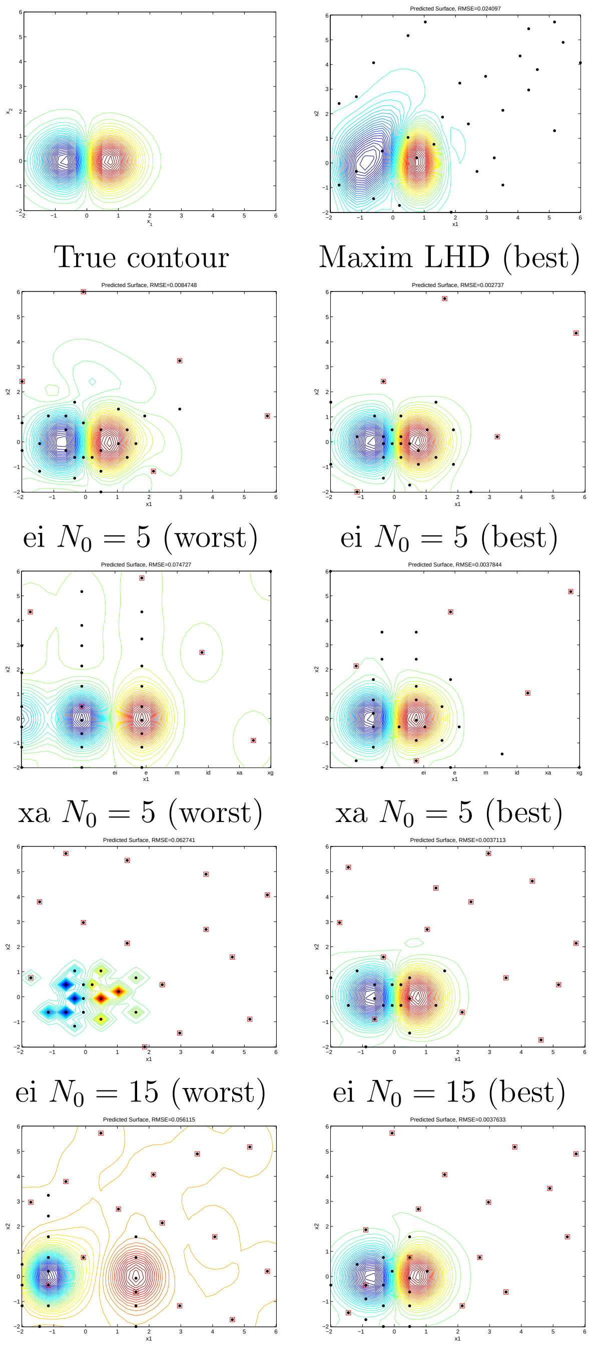
ei $N_0 = 15$ (worst)

ei $N_0 = 15$ (best)

xa $N_0 = 15$ (worst)

xa $N_0 = 15$ (best)

Figure 4.4: (Function N2, Two-dimensional exponential function) Contour plots of the true surface and predicted surfaces with cubic correlation 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.



True contour

Maxim LHD (best)

ei $N_0 = 5$ (worst)

ei $N_0 = 5$ (best)

xa $N_0 = 5$ (worst)

xa $N_0 = 5$ (best)

ei $N_0 = 15$ (worst)

ei $N_0 = 15$ (best)

xa $N_0 = 15$ (worst)

xa $N_0 = 15$ (best)

Figure 4.5: (Function N2, Two-dimensional exponential function) Contour plots of the true surface and predicted surfaces with Gaussian correlation 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.

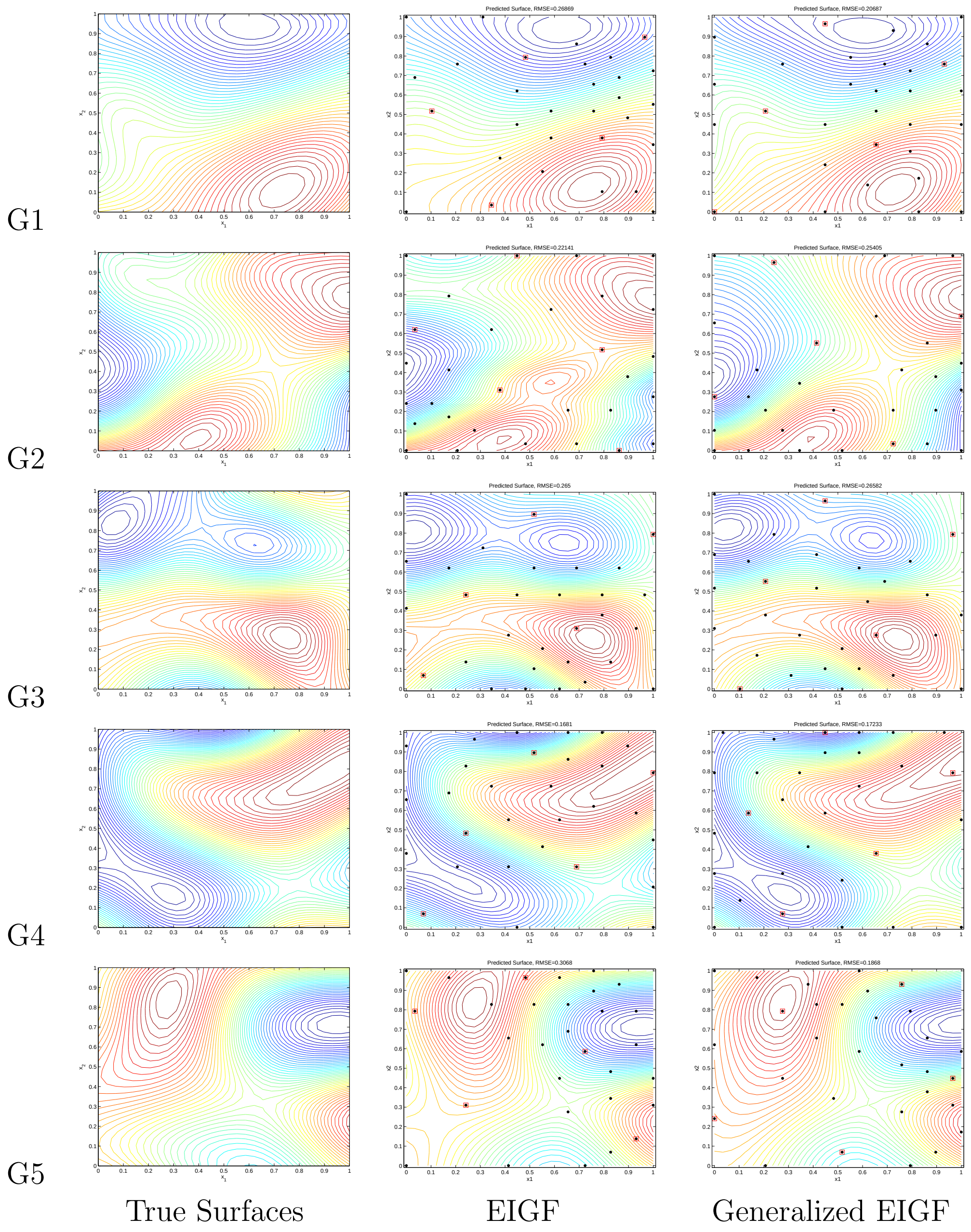
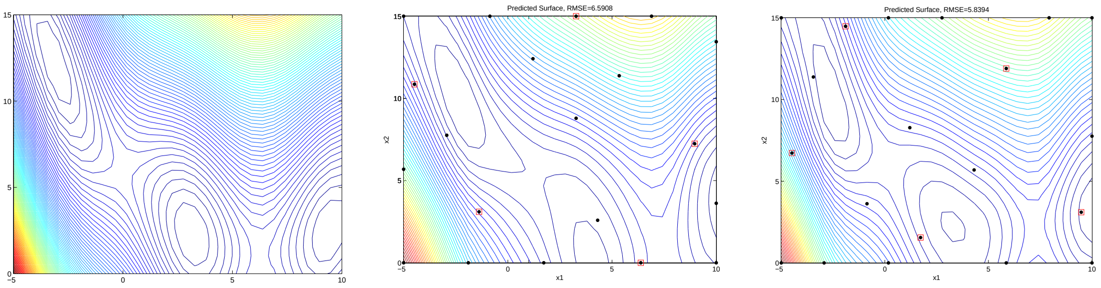


Figure 4.6: (GASP surfaces from Subsection 3.4.1: G1 to G5, from top to bottom row) Contour plots of true surfaces and predicted surfaces using EIGF and Generalized EIGF. The plots show the worst predicted surfaces based on ERMSPE among the 30 runs for each surface. Number of starting input points is $N_0 = 5$ and final number of design points is $N = 30$. True surfaces (left column), predicted surfaces with EIGF (middle column), predicted surfaces with generalized EIGF (right column).



True Surface EIGF Generalized EIGF

Figure 4.7: (Function M1, Branin function from Subsection 3.4.3) Contour plots of true surface and predicted surfaces using EIGF and generalized EIGF. The plots show the worst predicted surfaces based on ERMSPE among the 30 runs. Number of starting input points is $N_0 = 5$ and final number of design points is $N = 20$. True surface (left column), predicted surface with EIGF (middle column), predicted surface with generalized EIGF (right column).

CHAPTER 5

DIAGNOSTIC CHECKS FOR SEQUENTIAL DESIGNS

Several sequential as well as adaptive designs based on cross validation and a modified expected improvement criterion have been proposed and are shown in the previous two chapters to give accurate prediction of the response surfaces using the GASP model.

In this chapter, we present an exploratory tool to assess the goodness of fit of the fitted GASP model. In turn, this tool can be used to improve the predictive accuracy of the GASP model and also to provide a stopping criterion for the sequential designs. The examples in Subsections 3.4.1 and 3.4.3 reveal that that the GASP model with the Gaussian correlation function sometimes result in poor predictions as shown by the relatively larger empirical root mean squared prediction error (ERMSPE) as compared to the cubic correlation function. For examples, see Figures 3.3, 3.4, 3.10 and 3.13. Specifically, there is compelling evidence from Figure 3.3 (GASP surfaces) that the performances of the sequential designs are related to the choice of the correlation function and it can be seen that many runs of the various designs with the Gaussian correlation have relatively larger ERMSPE. The sequential designs run into numerical problems in estimating the correlation parameters, and the effectiveness and efficiency of the sequential procedures can be negatively affected. This also happens when the cubic correlation function is used, although the impact is not as bad as using the Gaussian correlation.

The results from our simulation study (described in Section 3.4) have been run with no user intervention, even if some of the estimated correlation parameters may not be reasonable. The hope is that adding additional design points will correct the problem. To reduce/eliminate this instability for sequential designs using the Gaussian correlation function, we present the use of *cross validation*, described in Section 1.5, as a diagnostic tool to check the fit of the GASP model at each iteration when a point is added and to improve the accuracy of the fitted GASP model.

The outline of this chapter is as follows. In Section 5.1, we present the diagnostic tools used in our simulation study and incorporate them into the sequential design algorithm outlined in Subsection 3.1.3. Five examples are given in Section 5.2 to illustrate the effectiveness of the modified design algorithm with the diagnostic checks. We conclude with a discussion of the proposed diagnostic checks and simulation results.

5.1 Diagnostic Tools for Assessing Response Surface Model Fit

Suppose we have the computer outputs $y(\boldsymbol{x}_i)$ at sampled points $\boldsymbol{x}_i$, where $i = 1, \cdots, n$. For each of the sampled point $\boldsymbol{x}_i$, we denote $\hat{\boldsymbol{Y}}_i^{(-i)}$ as the EBLUP of $y(\boldsymbol{x}_i)$ based on all the data except $\{\boldsymbol{x}_i, y(\boldsymbol{x}_i)\}$, i.e. removing the sampled point itself. We define the *cross validation prediction error* (XVPE) for each of the $\boldsymbol{x}_i$ as

$$XVPE(\boldsymbol{x}_i) \; = \; [\hat{Y}(\boldsymbol{x}_i)^{(-i)} - y(\boldsymbol{x}_i)]^2. \tag{5.1}$$

Two measures of goodness of fit of the GASP model using the n sampled points are given by:

(i) cross validation sum-of-squares of prediction error

$$XVSSPE_n = \sum_{i=1}^{n} XVPE(\boldsymbol{x}_i) = \sum_{i=1}^{n} (\hat{Y}(\boldsymbol{x}_i)^{(-i)} - y(\boldsymbol{x}_i))^2, \tag{5.2}$$

(ii) median cross validation prediction error

$$MXVPE_n = \underset{i=1,\cdots,n}{\text{median}}\{XVPE(\boldsymbol{x}_i)\} = \underset{i=1,\cdots,n}{\text{median}}\{(\hat{Y}(\boldsymbol{x}_i)^{(-i)} - y(\boldsymbol{x}_i))^2\}. \tag{5.3}$$

5.1.1 Modified Design Algorithm with Diagnostic Checks

Similar to the basic algorithm for sequential designs in Subsection 3.1.3, the modified algorithm proceeds as follows with the diagnostic checks included in Step 3. A flowchart of the modified algorithm using the GASP model with the Gaussian correlation function is shown in Figure 5.1 (page 110).

1. Identify the set of n sampled design points $\boldsymbol{x}$ for either

   (a) Augmenting the network of design points: identify the existing design points..

   (b) Selecting a new set of design points: generate a small starting design spread over the input space of $\boldsymbol{x}$. A space-filling design would be appealing for this initial fit of the response surface using the GASP model.

2. Run the computer code at the $\boldsymbol{x}$ input points and obtain the response, $y(\boldsymbol{x})$.

3. Estimate the correlation parameters (see Subsection 3.1.2).

   (a) First estimate the correlation parameters using the n observations as before in Subsection 3.1.2. Recall that, for the Gaussian correlation, the upper limit of the numerical search is determined by the Euclidean distance between the two closest sampled points and the lower limit depends on the range of the input space. The MLE of the correlation parameters using these limits is denoted as $\boldsymbol{\theta}_F$ (i.e., "full range" estimates).

Then compute the relative change in $\boldsymbol{\theta}_F$ and the previous estimates $\boldsymbol{\theta}_{n-1}$ based on $n-1$ observations. If the change is significantly large, we compute the $MXVPE_F$ (5.3) using the n observations and compare it to $MXVPE_{n-1}$ based on $n-1$ observations. If there is a large increase in $MXVPE_F$ over $MXVPE_{n-1}$, the range of the numerical search for the MLE will be reduced according to Step 3(b). (Note: if this is the first iteration of the algorithm where there is no $MXVPE_{n-1}$, then skip Steps 3(b) and (c), and proceed with Step 4).

(b) restrict the numerical search to a smaller range. We consider lowering the upper limit for the Gaussian correlation in our examples (see Subsection 5.1.2). The MLE of the correlation parameters using these limits is denoted as $\boldsymbol{\theta}_R$. As in Step 3(a), compare the $\boldsymbol{\theta}_R$ versus $\boldsymbol{\theta}_{n-1}$, and $MXVPE_R$ versus $MXVPE_{n-1}$. Proceed to Step 3(c) if needed.

(c) further restrict the range or do not update the estimates of the correlation parameters. The MLE of the correlation parameters using these limits is denoted as $\boldsymbol{\theta}_L$. For the later case, $\boldsymbol{\theta}_L$ is set to be the previous estimates $\boldsymbol{\theta}_{n-1}$, based on $n-1$ observations (if $\boldsymbol{\theta}_{n-1}$ is available).

4. Fit a GASP model (see Subsection 3.1.1) with $\boldsymbol{\theta}_n = \{\boldsymbol{\theta}_F, \boldsymbol{\theta}_R \text{ or } \boldsymbol{\theta}_L\}$ using the n observations from the computer simulator to predict $y(\boldsymbol{x}_0)$ at some untried $\boldsymbol{x}_0$.

5. Check stopping rule, if available. In this chapter, we consider the $MXVPE_n$ (5.3) and $XVSSPE_n$ (5.2) as an informal guide on the decision of whether to

stop. If additional points are needed, search for the $\boldsymbol{x}_0$ that maximizes the design criterion (see Section 3.4 for criteria used in the study) and add it to the existing set of sampled points giving a total of $n+1$ points. Repeat from Step 2 onwards with the $n+1$ points.

5.1.2 Details on Diagnostic Checks in Step 3

The diagnostic checks for Step 3 proposed in Subsection 5.1.1 are intended to improve the predictive performance of the sequential designs from the previous chapters. Recall in Section 3.4, the MLEs of the correlation parameters are estimated using a numerical search over a range of correlation values. The range for the search is set to be very wide to accomodate various types of response surfaces (i.e., varying strengths for the correlation parameters in the GASP model). We noted that this presents problems in some of the runs (i.e. the 30 different designs) where the estimates are pushed to the limit of the constraints and hence result in poor predictions using the GASP model. Here, we focus on the case where the Gaussian correlation parameters are too large.

Suppose we have the n computer outputs $y(\boldsymbol{x})$ and are estimating the correlation parameters in Step 3. For Steps 3 (a), (b) and (c), we consider restricting the range of the numerical search for the MLEs of the correlation parameters. For implementing our modified design algorithm (Subsection 5.1.1), we consider the following constraints and definitions in the subsequent examples in this chapter:

- *Step 3 (b) in Subsection 5.1.1* (restricted range): Specify a smaller range of the Gaussian correlation parameter values for the numerical optimization. For

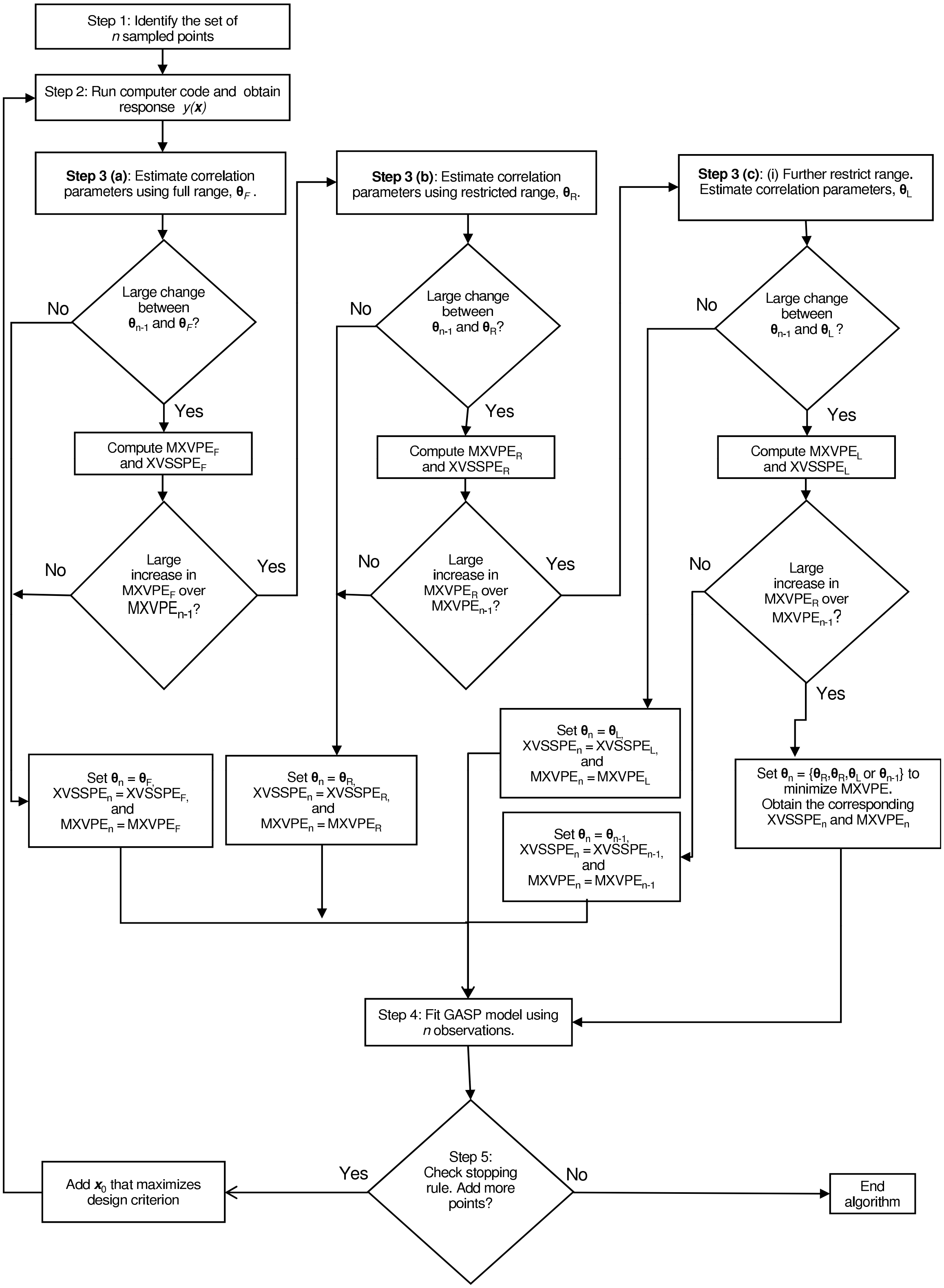

Step 1: Identify the set of n sampled points
Step 2: Run computer code and obtain response $y(\boldsymbol{x})$
Step 3 (a): Estimate correlation parameters using full range, $\boldsymbol{\theta}_F$.
Large change between $\boldsymbol{\theta}_{n-1}$ and $\boldsymbol{\theta}_F$?
No
Yes
Compute $MXVPE_F$ and $XVSSPE_F$
Large increase in $MXVPE_F$ over $MXVPE_{n-1}$?
No
Yes
Step 3 (b): Estimate correlation parameters using restricted range, $\boldsymbol{\theta}_R$.
Large change between $\boldsymbol{\theta}_{n-1}$ and $\boldsymbol{\theta}_R$?
No
Yes
Compute $MXVPE_R$ and $XVSSPE_R$
Large increase in $MXVPE_R$ over $MXVPE_{n-1}$?
No
Yes
Step 3 (c): (i) Further restrict range**.** Estimate correlation parameters, $\boldsymbol{\theta}_L$
Large change between $\boldsymbol{\theta}_{n-1}$ and $\boldsymbol{\theta}_L$?
No
Yes
Compute $MXVPE_L$ and $XVSSPE_L$
Large increase in $MXVPE_R$ over $MXVPE_{n-1}$?
No
Yes
Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{L,}$ $XVSSPE_n = XVSSPE_{L,}$ and $MXVPE_n = MXVPE_L$
Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{F,}$ $XVSSPE_n = XVSSPE_{F,}$ and $MXVPE_n = MXVPE_F$
Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{R,}$ $XVSSPE_n = XVSSPE_{R,}$ and $MXVPE_n = MXVPE_R$
Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1,}$ $XVSSPE_n = XVSSPE_{n-1,}$ and $MXVPE_n = MXVPE_{n-1}$
Set $\boldsymbol{\theta}_n$ = {$\boldsymbol{\theta}_R$,$\boldsymbol{\theta}_R$,$\boldsymbol{\theta}_L$ or $\boldsymbol{\theta}_{n-1}$} to minimize MXVPE. Obtain the corresponding $XVSSPE_n$ and $MXVPE_n$
Step 4: Fit GASP model using n observations.
Step 5: Check stopping rule. Add more points?
Yes
No
Add $\boldsymbol{x}_0$ that maximizes design criterion
End algorithm

Figure 5.1: Flowchart of modified design algorithm with diagnostic checks

example, an upper bound of $e^{-\theta} = 0.0001$ is set for the numerical optimization for the examples in Section 5.2.

- *Step 3 (c)* (Further restriction): If the fit of the GASP model is still poor (based on the $MXVPE$), we then set the upper limit of each of the correlation parameters to $\min(\boldsymbol{\theta}^*_{3b}, 2 \times \boldsymbol{\theta}_{n-1})$, where $bt\theta^*_{3b}$ denotes the upper limit set in Step 3(b) and $\boldsymbol{\theta}_{n-1}$ denotes the previous estimate of $\boldsymbol{\theta}$ based on $n-1$ observations.

- *Step 3 (c)* (Do not update the MLEs at this stage): If the MLE of the correlation parameters $\boldsymbol{\theta}$ based on the n observations are not reasonable in the sense that they result in a poorly fitted GASP model, we choose not to update the MLE but use the estimates, $\boldsymbol{\theta}_{n-1}$. A rationale is that the newly added n^{th} point may be very influential in affecting the predicted surface and might have caused problems for estimating the correlation parameters. This results in inaccurate predictions for the other regions of the input space. Since our goal is an accurate GASP model over the entire input space, temporarily “ignoring” this observation would seem reasonable.

- *Definition of large change in MLE of correlation parameters* (Steps 3 a-c): We define a change in $\boldsymbol{\theta}_F$, $\boldsymbol{\theta}_R$ or $\boldsymbol{\theta}_L$ as large if any of the estimated parameters, based on n observations for each dimension, changes by more than 100 percent from $\boldsymbol{\theta}_{n-1}$ based on $n-1$ observations.

- *Definition of large increase in $MXVPE$* (Steps 3 a-c): We define a change in $MXVPE_F$, $MXVPE_R$ and $MXVPE_L$ as large if it increases by more than 100 percent over $MXVPE_{n-1}$.

5.2 Examples: Illustrations of Modified Algorithm

The following examples illustrate the implementation of the modified design algorithm with the diagnostic checks outlined in Section 5.1.1. The cross validation with the geometric mean criterion (3.7) will be used in the examples. Recall in Section 3.4 that the examples in Chapters 3 and 4 are based on 30 runs of different designs using the maximn LHD generated from our software. For the examples in this section, we identify some of the runs that have relatively large ERMSPE to demonstrate the modified algorithm. A detailed illustration using the six-hump camel-back function is given first, followed by four other examples.

- *Function M3, six-hump camel-back function*: We consider the worst predicted surface based on ERMSPE (of 0.62) among the 30 runs in Figure 3.10 (page 76) for the cross validation with the geometric mean (xg) and $N_0 = 20$ starting design points. Plots of the $MXVPE$ and $XVSSPE$ against the final number of sampled points N (from 20 to 60) are shown in Figure 5.2(a) and (b) respectively (page 118), for both the Gaussian and cubic correlation functions.

  After adding the $27th$ point, there is an increase in the corresponding $XVSSPE_F$ (Step 3a) without a significant change in the estimated correlation parameters $\boldsymbol{\theta}_F$ over $\boldsymbol{\theta}_{n=26}$ with 26 points. A plot of the predicted surface (not shown) shows that the increase in the $XVSSPE_F$ is due to the $27th$ point being added at the top-right corner of the surface at inputs $(x_1, x_2) = (2, 1)$. No adjustment is made to the sequential criterion at this point and we proceed with adding more points. A large change is noted in $\boldsymbol{\theta}_F$ with 38 points over $\boldsymbol{\theta}_{n=37}$. After carrying out Step 3(a), Figure 5.2(a) shows a large increase in the $MXVPE_F$

(see Figure 5.2(c) for comparative plot of $XVPE(\boldsymbol{x})$ for each sampled point $\boldsymbol{x}$ for $n = 37$ and $n = 38$). The upper limit for the numerical search for $\boldsymbol{\theta}$ is reduced by restricting the maximum θ such that $e^{-\theta} = 0.0001$ (based on Step 3b). The estimated $\boldsymbol{\theta}_R$ still changed significantly from $\boldsymbol{\theta}_{n=37}$ with no significant dcrease in the $MXVPE_R$. We proceed to Step 3(c) and set the upper limit to 2 $\times$ $\boldsymbol{\theta}_{n=37}$ and it was found that the $MXVPE_L$ decreased, shown by the dashed line (-×-) in Figure 5.2(a). The final estimate, $\boldsymbol{\theta}_{n=38}$, is set to be $\boldsymbol{\theta}_L$.

It is noted that abrupt changes in $\boldsymbol{\theta}$ and $MXVPE$ are not clearly reflected in the corresponding $XVSSPE$ plot in Figure 5.2(b). As a comparison, plots of the predicted surfaces (using the modified algorithm with the Gaussian correlation, and the original algorithm with the Gaussian and cubic correlation functions) are shown in Figure 5.2(d), (e) and (f) respectively with a final sample of $N = 40$ points. Their ERMSPE are 0.6197, 0.25295 and 0.12541 in the same order.

This example also illustrates the usefulness of using the $XVSSPE$ as a check of the fitted model in a sequential setting. In this case, the plot reveals very early in the algorithm that the GASP model with the Gaussian correlation function might not be a good fit and addition of more design points should proceed with care on how to deal with poor predictions such as those when the $38th$ to $40th$ points are added. Together, the $MXVPE$ and $XVSSPE$ also offers a guide to when to stop the sequential algorithm. In Figure 5.2(b), we continue to add more points beyond $N = 40$ until $N = 60$, and the $MXVPE$ and $XVSSPE$ seem to level off at around the $55th$ sampled point, suggesting a good fit of the GASP model.

- *Function G5a, GASP model surface*: (worst case predicted surface based on ERMSPE) Recall in Figure 3.3(a) that the cross validation with the geometric mean criterion (xg) and the Gaussian correlation has a few relatively large ERMSPE for the GASP model surface $G5$. The worst case predicted surface has a ERMSPE of 0.9417 (for the design labeled as xg5 in Figure 3.3).

  Figure 5.3 (left column) shows the $MXVPE$ and $XVSSPE$ increased significantly when the $26th$ point is added. The plot in the third row (left column) shows the predicted surface with $N = 30$ points with the original EIGF algorithm. To improve the prediction, we carry out the modified algorithm and we constrain the numerical search to an upper limit of $e^{-\theta} = 0.0001$ (Step 3b) starting from the 26th point onwards. Significant improvement is evident in the predicted surface (plot in fourth row, left column) with a signficantly smaller ERMSPE of 0.0654. And, the $MXVPE$ and $XVSSPE$ plots (based on the modified algorithm) seem to stabilize after the 30th point suggesting a good fit of the GASP model.

- *Function G5b*: (second worst case predicted surface based on ERMSPE) We use Function G5 again but for the 2nd worst predicted surface with an ERMSPE of 0.8095. The $MXVPE$ and $XVSSPE$ plots (see Figure 3.3 right column), based on the original algorithm, show large ERMSPE values for the 22th, 23th, 24th and 30th observations. The final predicted surface with $N = 30$ points is displayed in the third row (right column). As with the previous example, we constrain the search to an upper bound of $e^{-\theta} = 0.0001$ (Step 3b of the modified algorithm) from the $22th$ observation onwards when neccessary (Note: Steps 3 (b) and (c) are not implemented when the $26th$ and $28th$-$30th$ observations are added). The predicted surface, based on the modified algorithm, is shown in the bottom row with a smaller ERMSPE of 0.0907. Again, the $MXVPE$ and $XVSSPE$ plots (based on the modified algorithm) seem to stabilize after the 30th point suggesting a good fit of the GASP model.

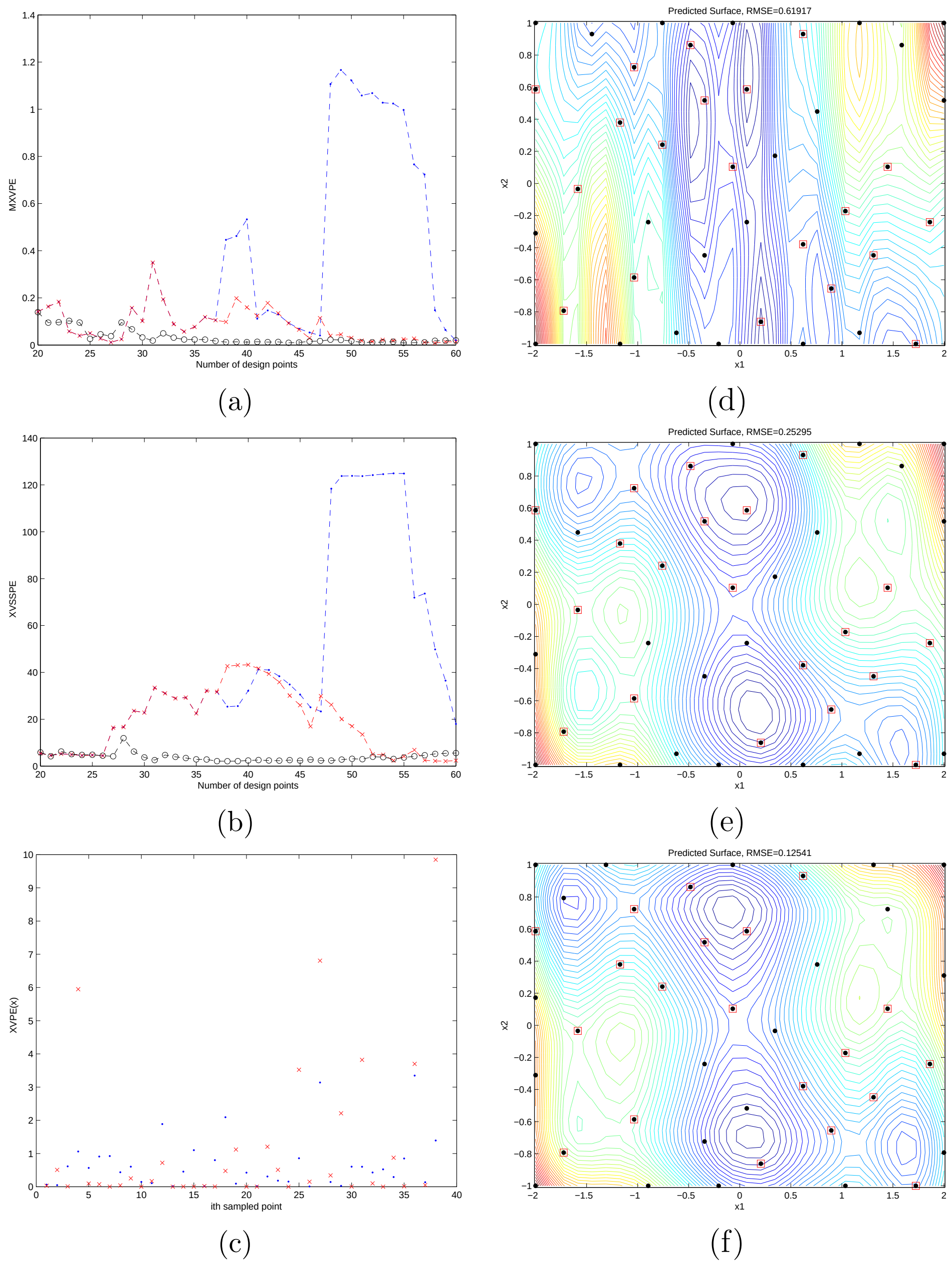
The significant decreases in ERMSPE for these two GASP model surfaces (Function G5a and G5b) are not surprising since the random surfaces were generated from the GASP model using the Gaussian correlation function. With the modifed algorithm, the cross validation with the geometric mean criterion and the Gaussian correlation outperforms the same criterion with the cubic correlation.

- *Function M4, non-polynomial surface*: The boxplot in Figure 3.13 shows an outlier ERMSPE value of 0.2988 for the cross validation with the geometric mean (xg) criterion with the Gaussian correlation and $N_0 = 10$. The predicted surface for this run is shown in Figure 3.15 and also in Figure 5.4 (left column). The $MXVPE$ and $XVSSPE$ plots show the large errors when th e $11th$ to $13th$ observations are added. Despite the large errors, the estimated the correlation parameters did not change significantly with each additional point. The increases in the $MXVPE$ and $XVSSPE$ are due to the GASP model detecting new features of the response surface and not due to the instability of estimating the correlation parameters. Steps 3 (b) and (c) are not implemented. Additional points are then added and it is noted that the $MXVPE$ and $XVSSPE$ gradually decrease after the $13th$ observation. The addition of the $20th$ observation caused a significant change in the estimated correlation parameters, $\boldsymbol{\theta}_F$, over $\boldsymbol{\theta}_{n=19}$, and the corresponding $MXVPE$ increased from about $MXVPE_{n=19} = 4$ to $MXVPE_F = 14$. Carrying out Steps 3(b) and (c) of the modified algorithm, the final $\boldsymbol{\theta}_{n=20}$ is set to $\boldsymbol{\theta}_{n=19}$ based on the previous 19 observations. The predicted surface (shown in the last row) using the modified algorithm shows a smoother predicted surface with lower ERMSPE of 0.0824. According to thhe $MXVPE$ and $XVSSPE$ plots, it would be more conservative to take a final sample of at least $N = 25$ points for a better fit of the GASP model.

- *Function M1, Branin function*: The worst predicted surface, based on the ERMSPE in Figure 3.4 for the Gaussian correlation function and a initial design of $N_0 = 10$ points, is shown in Figure 5.4 (right column). Large values of $MXVPE$ and $XVSSPE$ can be seen in the initial stage of the algorithm and gradually decrease as more observations are added. The large values are primarily due to the addition of points at the bottom left of the Branin surface (presence of a sharp peak) but the estimated correlation parameters did not change significantly. The modified algorithm (i.e., Steps 3b and 3c) is not implemented for this example. The $MXVPE$ and $XVSSPE$ plots suggest stopping the algorithm with about $N = 25$ observations.

5.3 Discussion and Conclusion

The examples in Section 5.2 show that the modified design algorithm with the diagnostic checks improves the predictive performance of the cross validation with the geometric mean criterion and the Gaussian correlation function. We have considered the use of the median cross validation prediction error ($MXVPE$) and cross validation sum-of-squares of prediction error ($XVSSPE$) as diagnostic tools in this study for improving response surface model fit. Further refinements can be made to consider other error summaries. The primary advantage of the proposed diagnostic checks is that we can quantify the goodness of fit of the GASP model at each iteration when an observation is added without having to plot the predicted surfaces, which can be difficult to visualize in higher dimensions. The $MXVPE$ and $XVSSPE$ can also be used as indicators for stopping the design algorithm. The idea is to predict the response surface with the GASP model and stop adding obseravtions when the $MXVPE$ and $XVSSPE$ are "small". More extensive simulation studies are needed to provide more insight and directions on how to extend the methodologies.



(a)

(d)

(b)

(e)

(c)

(f)

Figure 5.2: Diagnostic summaries and predicted surfaces for Function M3 (six-hump camelback function). (a) MXVPE, (b) XVSSPE for original algorithm and Gaussian correlation (-·-), modified algorithm and Gaussian correlation (-×-), original algorithm and cubic correlation (-o-), (c) XVPE($\boldsymbol{x}$) for each sampled point $\boldsymbol{x}$, (×) for $XVPE_F(\boldsymbol{x})$, (·) for $XVPE_{n=37}(\boldsymbol{x})$.

Predicted surfaces with EIGF and $N = 40$ points for (d) original algorithm, Gaussian correlation, (e) modified algorithm, Gaussian correlation, (f) original algorithm, cubic correlation.

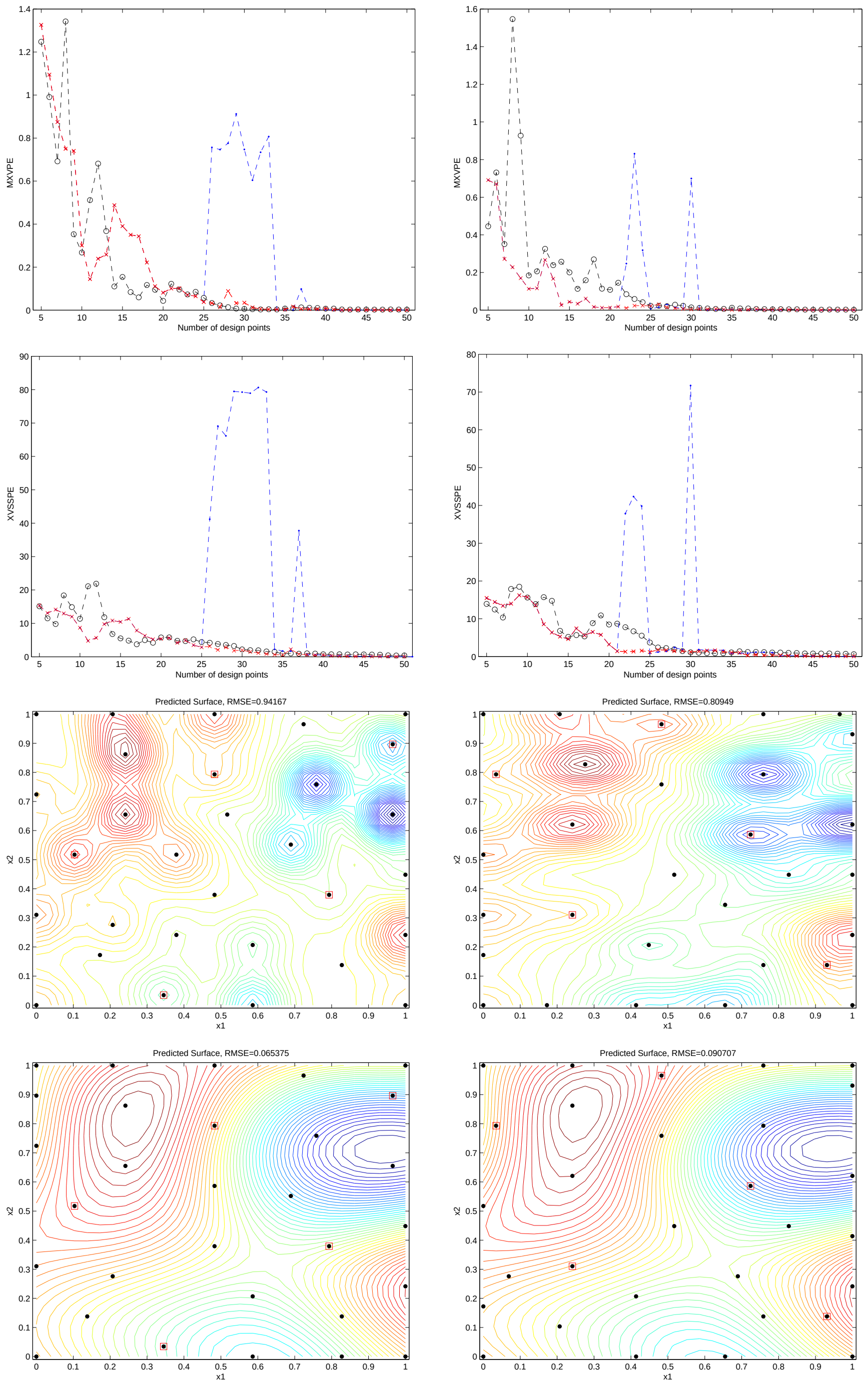


Figure 5.3: Plots of sequential diagnostics and predicted surfaces for GASP Function G5a (left column) and Function G5b (right column).
(1st row) $MXVPE$, (2nd row) $XVSSE$ for original algorithm with Gaussian (-·-) and cubic (-o-) correlations, modified algorithm with Gaussian(-×-), (3rd row) predicted surface with original algorithm with $N = 30$ points, (4th row) predicted surface with modified algorithm.

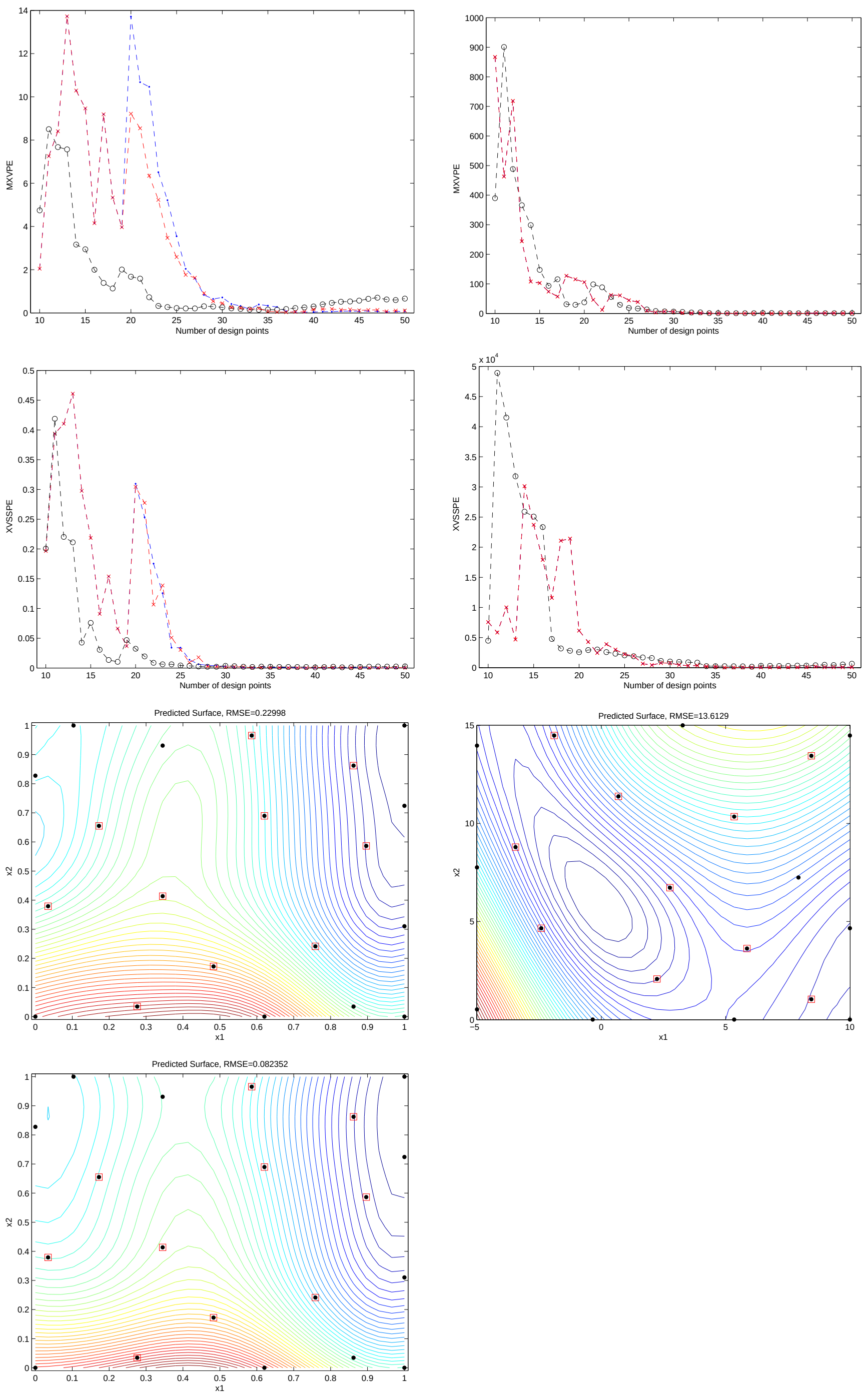


Figure 5.4: Plots of sequential diagnostics and predicted surfaces for Function M4, Non-polynomial surface (left column) and Function M1, Branin function (right column) . (1st row) $MXVPE$, (2nd row) $XVSSE$ for original algorithm with Gaussian (-·-) and cubic (-o-) correlations, modified algorithm with Gaussian(-×-), (3rd row) predicted surface with original algorithm with $N = 20$ points, (4th row) predicted surface with modified algorithm (not available for Branin function).

CHAPTER 6

SEQUENTIAL ADAPTIVE DESIGNS FOR MODEL FIT OF INTEGRATED RESPONSE SURFACES

This chapter considers sequential design criteria for response surface model fit in situations where the input variables in computer experiments consist of both *control* and *environmental* variables. We consider control variables as those that can be precisely controlled by the experimenter in a physical experiment, or can be precisely controlled in a manufacturing process, while environmental variables are beyond the experimenter's or manufacturer's control. However, in a lab setting (as assumed in this chapter), we assume that the experimenter has control over the environmental variables and that the values of the input points can be described by probability density functions.

We define the input points to be $\boldsymbol{x} = \{\boldsymbol{x}_c, \boldsymbol{x}_e\}$ where $\boldsymbol{x}_c$ and $\boldsymbol{x}_e$ denotes the control and environmental input points, respectively. We assume that the values of the environmental input points vary according to some specified probablity density function $F(\boldsymbol{x}_e)$. In the presence of both control and environmental variables, we focus on the selection of input points (for both control and environmental variables) at which to run the simulations so as to obtain good fit of the GASP model for the integrated response surface,

$$\mu_F(\boldsymbol{x}_c) = E_{F(\boldsymbol{x}_e)}[y(\boldsymbol{x}_c, \boldsymbol{x}_e)] \tag{6.1}$$

For computational feasibility, we further assume that the environmental variables have finite support. Suppose $\boldsymbol{x}_c$ is one of the control variable input point. Let $\{\boldsymbol{x}_c, \boldsymbol{x}_{e,j}\}$ denote the j^{th} support point for the environmental variables, $\boldsymbol{x}_{e,j}$, paired with $\boldsymbol{x}_c$, where $j = 1, \cdots, n_e$ and n_e denotes the number of support points for the environmental variables. In this chapter, we consider the specific case where the quantities of interest are

$$\mu(\boldsymbol{x}_c) = \sum_{j=1}^{n_e} w_j \; y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) \tag{6.2}$$

where $\{w_j\}_{j=1}^{n_e} = P(\boldsymbol{x}_e = \boldsymbol{x}_{e,j})$ denotes the corresponding non-negative weights and such that $\sum\limits_{j=1}^{n_e} w_j = 1$. Our objective is to obtain a good model fit of the response surface of $y(\cdot)$ averaged over the distribution of the environmental variables $\boldsymbol{x}_e$ using (6.2).

6.1 Statistical Model

As before, it is assumed that the deterministic output $y(\boldsymbol{x})$ is a realization of a stochastic process (or random function), $Y(\boldsymbol{x})$. The model used is

$$Y(\boldsymbol{x}) = f^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{6.3}$$

where $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x}))^T$ 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 (6.3) and the correlation function $R(\cdot)$ essentially controls the smoothness of the process.

We consider a Bayesian approach which provides a more general approach to the prediction problem in computer experiments. Recall in Subsection 1.3.2, we have the joint distribution of the responses $\boldsymbol{Y}^n$ at the sampled points and some untried $\boldsymbol{x}_0$ given by

$$\begin{pmatrix} \boldsymbol{Y}^n \\ Y(\boldsymbol{x}_0) \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{F} \\ f^T(\boldsymbol{x}_0) \end{pmatrix} \boldsymbol{\beta}, \sigma^2 \begin{pmatrix} \boldsymbol{R} & r(\boldsymbol{x}_0) \\ r(\boldsymbol{x}_0) & 1 \end{pmatrix} \right]. \tag{6.4}$$

Assuming a non-informative prior distribution of $[\boldsymbol{\beta}] \propto 1$ with $R(.)$ and σ^2 known, we obtain the conditional posterior distribution $[Y(\boldsymbol{x}_0)|\boldsymbol{Y}^n]$ by applying Theorem (A.1)

$$\begin{aligned} [Y(\boldsymbol{x}_0)|\boldsymbol{Y}^n] \sim \quad & N\Big(\boldsymbol{\beta} + r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}\boldsymbol{\beta}), \\ & \sigma^2 \left[1 - r^T(\boldsymbol{x}_0)\boldsymbol{R}^{-1}r(\boldsymbol{x}_0) + \frac{\left(1 - \mathbf{1}^T\boldsymbol{R}^{-1}r(\boldsymbol{x}_0)\right)^2}{\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1}}\right]\Big) \end{aligned} \tag{6.5}$$

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 $Y(\boldsymbol{x}_0)$. One can recognize that the posterior distribution, $[Y(\boldsymbol{x}_0)|\boldsymbol{Y}^n]$, is Gaussian with mean and variance identical to the BLUP and MSPE in (1.7) and (1.8) respectively for a constant mean term model.

We adopt an empirical Bayesian approach by substituting for the unknown parameters their maximum likelihood estimates. If $R(\cdot)$ is known and the estimate of $\boldsymbol{\beta}$ is taken to be $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{Y}^n$ (generalized least-squares estimate).

6.2 Design Criteria for Integrated Response Surfaces

In the previous chapters, the *expected improvement for global fit* (EIGF) criterion was introduced as one of the sequential design criteria for response surface model fit. The EIGF criterion can be easily implemented in this situation, where both control and environmental variables are present, by simply treating the environmental variables as additional variables to be evaluated as part of the criterion computation. We do not distinguish between the types of variables, i.e., whether they are control or environmental variables (see Subsection 6.2.1).

Under the assumption that the environmental variables vary according to some specified probablity density function, we extend the EIGF criterion by incorporating some form of weighting in the selection of the input points in Subsection 6.2.2 for predicting the integrated response surface, $\mu(\boldsymbol{x_c})$, given in (6.2).

In Subsection 6.2.3, an extension of the EIGF criterion is proposed where the improvement function is taken over the sum of the response over the support points of the environmental variables according to the w_j in (6.2).

6.2.1 EIGF Criterion

Following Subsection 4.2.1, we have the computer outputs $y(\boldsymbol{x}_k)$ at sampled points $\boldsymbol{x}_k$, $k = 1, \cdots, n$. For each potential input point $\boldsymbol{x}$, its improvement is defined as

$$I(\boldsymbol{x}) = (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{k^*}))^2 \tag{6.6}$$

where $y(\boldsymbol{x}_{k^*})$ refers to the observed output at the sampled point, $\boldsymbol{x}_{k^*}$, that is closest (in Euclidean distance) to the candidate point $\boldsymbol{x}$. The EIGF criterion is to choose the next input point that maximises the expected improvement

$$E(I(\boldsymbol{x})) = (\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{k^*}))^2 + var(\hat{Y}(\boldsymbol{x})). \tag{6.7}$$

6.2.2 Weighted EIGF Criterion

We consider an extension of the EIGF criterion in (6.7) by incorporating the weights, w_j, from the objective function (6.2) into the selection of the the next input point. If the probablity distribution $F(\boldsymbol{x}_e)$ is not uniform, we would like the design criterion to place more "emphasis" on the candidate point where its environmental variable input point has more weight (i.e., larger values of w_j). We consider a simple case where

$$E(I(\boldsymbol{x})) = w_j\big\{(\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{k^*(j)}))^2 + var(\hat{Y}(\boldsymbol{x}))\big\}. \tag{6.8}$$

where $y_{k^*(j)}$ denotes the observed output at the sampled point, $\boldsymbol{x}_{k^*(j)}$, that has the same j^{th} support point for the environmental variable inputs and is closest (in Euclidean distance) to the candidate point $\boldsymbol{x}$. The key difference is that the set of $\{y_{k^*(j)}\}$ is a subset of $\{y_{k^*}\}$ in (6.7). The idea is to indirectly select $\boldsymbol{x}$ by considering the fitted response surface for each support point of the environmental variables separately but still making use of the GASP model fitted with all inputs.

6.2.3 Integrated EIGF Criterion

Williams *et al.* (2000) extended the expected improvement criterion to situations where both control and environmental variables are present (an overview of the criterion is given in Subsection 2.5.2). Unlike their algorithm, which is intended for global optimization, we modify the criterion for response surface model fit. The proposed *integrated* EIGF criterion consists of the following steps:

1. Generate an initial (space filling) design.
2. Run the computer code at the $\boldsymbol{x}$ input points and obtain the response, $y(\boldsymbol{x})$.
3. Estimate the correlation parameters (see Subsection 3.1.2), as before.
4. Fit a GASP model using the n observations from the computer simulator to predict $y(\boldsymbol{x}_0)$ at some untried $\boldsymbol{x}_0$.

5. Suppose we have the computer outputs $y(\boldsymbol{x}_k)$ at sampled points $\boldsymbol{x}_k$, $k = 1, \cdots, n$. The *integrated* EIGF criterion is used to choose the next input point $\boldsymbol{x} = \{\boldsymbol{x}_c, \boldsymbol{x}_e\}$ in two stages as follows:

   (a) Select the control variable input point: For each potential control input point $\boldsymbol{x}_c$, its *integrated* expected improvement will be computed and the next point that maximizes the criterion will be selected (details to follow after these steps).

   (b) Select the environmental variable input point: Search for the $\boldsymbol{x}_{e,n+1}$ corresponding to $\boldsymbol{x}_{c,n+1}$ that satisfies a specified criterion. The criterion to be evaluated will depend on the desired robust design formulation (details to follow after these steps).

6. Check the stopping rule, if available. If additional points are needed, repeat from step 2 onwards.

More on Step 5(a): Suppose we have the computer outputs y_k at sampled points $\boldsymbol{x}_k$, $k = 1, ..., n$. For each candidate control input point $\boldsymbol{x}_c$, its improvement function ("averaged" over the distribution of the environmental variables) is defined by

$$I(\boldsymbol{x}_c) = \Big[\sum_{j=1}^{n_e} w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \Big(\sum_{\substack{j=1 \\ j \neq m}}^{n_e} w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}) + y_{k^*(m)}\Big)\Big]^2 \tag{6.9}$$

where $y_{k^*(m)}$ refers to the observed output at the sampled point, $\boldsymbol{x}_{k^*} = \{\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,k^*}\}$, of which its control variable input $\boldsymbol{x}_{c,k^*}$ is closest (in distance) to the control input of the candidate point $\boldsymbol{x}_c$, and the m in parenthesis is used to index the environmental input $\boldsymbol{x}_{e,k^*}$. As before, we shall determine this nearest sampled design point $\boldsymbol{x}_{k^*}$ using Euclidean distance.

The *integrated* expected improvement criterion is to choose the next control input point that maximizes the expected improvement

$$\begin{aligned}
E(I(\boldsymbol{x}_c)) \;&=\; E\left(\sum_{j=1}^{n_e} w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \Big(\sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}) + y_{k^*(m)}\Big)\right)^2 \\
&= \left[E\Big(\sum_{j=1}^{n_e} w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j})\Big) - y_{k^*(m)}\right]^2 \\
&\quad + Var\left(\sum_{j=1}^{n_e} w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}) - y_{k^*(m)}\right) \\
&= \left[\sum_{j=1}^{n_e} w_j \hat{Y}(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j \hat{Y}(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}) - y_{k^*(m)}\right]^2 \\
&\quad + Var\left(\sum_{j=1}^{n_e} w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j})\right) \qquad\qquad (6.10) \\
&= \left[\sum_{j=1}^{n_e} w_j \hat{Y}(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}) - \sum_{\substack{j=1\\ j\neq m}}^{n_e} w_j \hat{Y}(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}) - y_{k^*(m)}\right]^2 \\
&\quad + \sum_{j=1}^{n_e} Var(w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j})) + \sum_{\substack{j=1\\ j\neq m}}^{n_e} Var(w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j})) \\
&\quad + \sum_{\substack{j=1\\ j<m}}^{n_e} 2\; Cov(w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}), w_h Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,h})) \\
&\quad + \sum_{\substack{j=1\\ j<h,\; h\neq m}}^{n_e} 2\; Cov(w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,j}), w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,h})) \\
&\quad - \sum_{j,h=1,\; h\neq m}^{n_e} 2\; Cov(w_j Y(\boldsymbol{x}_c, \boldsymbol{x}_{e,j}), w_j Y(\boldsymbol{x}_{c,k^*}, \boldsymbol{x}_{e,h})).
\end{aligned}$$

Similar to the EIGF criterion in (4.2) and (6.6), the expected improvement in (6.10) consists of two search components — local and global. The first (local) component, given by the $[\,\cdot\,]^2$ term, will tend to be large at input point $\boldsymbol{x}_c$ where it has the largest "averaged" (response) increase over its sampled control input point. The second (global) component is determined by the variance and covariance terms remaining in (6.10).

Step 5(b): Let $\boldsymbol{x}_e$ denote the set of environmental inputs corresponding to the selected control variable $\boldsymbol{x}_c^*$ selected in step 5(a). For each potential input point $\{\boldsymbol{x}_c^*, \boldsymbol{x}_e\}$, the design criterion is to choose the next environmental input point that

1. maximizes the prediction error in (6.5),

$$\boldsymbol{x}_e^* = \max_{\boldsymbol{x}_c^*, \boldsymbol{x}_e} (Var(\hat{Y}(\boldsymbol{x}_c^*, \boldsymbol{x}_e))) \tag{6.11}$$

2. minimizes the average prediction error (6.5) among the remaining $n_e - 1$ candidate input points,

$$\boldsymbol{x}_e^* = \min_{\boldsymbol{x}_c, \boldsymbol{x}_e} \Big(\frac{1}{n_e - 1} \sum_{\boldsymbol{x}_{e,j} \neq \boldsymbol{x}_e^*} Var(\hat{Y}(\boldsymbol{x}_c^*, \boldsymbol{x}_{e,j})) \Big) \tag{6.12}$$

3. maximizes the expected improvement given by (4.2), or

$$\boldsymbol{x}_e^* = \max_{\boldsymbol{x}_c^*, \boldsymbol{x}_e} E(I(\boldsymbol{x}_c^*, \boldsymbol{x}_e)) \tag{6.13}$$

4. maximizes the *weighted* expected improvement,

$$\boldsymbol{x}_e^* = \max_{\boldsymbol{x}_c^*, \boldsymbol{x}_e} w_j E(I(\boldsymbol{x}_c^*, \boldsymbol{x}_{e,j})). \tag{6.14}$$

Together, the selected values for the control and environmental variables, $\boldsymbol{x}_c^*$ and $\boldsymbol{x}_e^*$, respectively, are used as the (new) additional design point and the sequential design algorithm proceeds as before.

6.3 Examples

The following examples are intended to illustrate the implementation and prediction performance of the various design strategies. The design strategies to be compared are:

- EIGF criterion in (6.7)
- weighted EIGF criterion (wEIGF) in (6.8)
- Integrated EIGF criterion with the environmental input points selected by maximizing the expected improvement (IEIGF) in (6.13)
- Integrated EIGF criterion with the environmental input points selected by maximizing the weighted expected improvement (IEIGFw) in (6.14)

(the abbreviations 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) of selecting a fixed-point design, namely with $N = 10 \times p$ points, where p is the number of dimensions of the input space.

The number of starting design points is determined by the number of levels of the environmental variables. Given that the wEIGF criterion in (6.8) requires at least one sampled point from each environmental input, the number of starting design points N_0 will depend on the number of support points for the environmental variable. For example, $N_0 = 3$ will be used for our first example with the 2-dimensional Branin function.

Since the input points generated by maximin LHD are not unique, our comparisons are based on 30 runs of different designs. For each of the 30 runs, we first obtain the initial starting design using a N_0-point approximate maximin LHD for the control variables only. We then cycle through different permutations of the selected control variable input with all the support points of the environmental variable and pick the starting design as the one that gives the maximin distance among the $\boldsymbol{x}$ inputs.

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. Only the cubic correlation function will be used for the examples in this section.

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

$$
\begin{aligned}
ERMSPE &= \sqrt{\frac{\sum_{i=1}^{m}\left(\sum_{j=1}^{n_e}\hat{Y}(\boldsymbol{x}_{i,j})-\sum_{j=1}^{n_e}y(\boldsymbol{x}_{i,j})\right)^2}{m}} \\
&= \sqrt{\frac{\sum_{i=1}^{m}\left(\hat{\mu}(\boldsymbol{x}_{c,i})-\mu(\boldsymbol{x}_{c,i})\right)^2}{m}}
\end{aligned}
\qquad (6.15)
$$

where $\boldsymbol{x}_{i,j}$, $i = 1, ...m$ and $j = 1, ..., n_e$ are the grid points used for evaluating the prediction accuracy and m is the total number of grid points; $\hat{y}(\boldsymbol{x}_{i,j})$ is the predicted response at $\boldsymbol{x}_{i,j}$; $y(\boldsymbol{x}_{i,j})$ are the true values at the same set of grid points; $\mu(\boldsymbol{x}_{c,i})$ and $\hat{\mu}(\boldsymbol{x}_{c,i})$ is the observed output and predicted response, respectively, averaged over the support of the environmental variables using (6.2).

6.3.1 Test Functions and Features

In this study, four examples are used to evaluate the predictive performance of the GASP model with the input points chosen by the various design criteria. Details of the functions are given below and plots of the true response surfaces for the different support of the environmental variables are shown in Figure 6.1 (left column) on page 137.

Function 1: Two-dimensional Branin function

We assume the response is evaluated via

$$f(x_1, x_2) \;=\; (x_2 - \frac{5.1}{4\pi^2}x_1^2 + 5/\pi x_1 - 6)^2 + 10(1 - \frac{1}{8\pi}\pi)\; cos(x_1) + 10$$

where $x_1 \in [-5, 10],\; x_2 \in [0, 15]$. We take x_1 to be the control variable, x_c, and x_2 to be the environmental variable x_e at support points $\{0.25, 0.5, 0.75\}$. Three different distributions for the environmental variable x_2 are considered (in Table 6.1) and are labeled as unequal, moderate and equal according to the degree of uniformity of the weights for each of the support points. The entries in the table denote the weights w_j in the objective function (6.2).

	x_2		
	0.25	0.5	0.75
unequal	1/10	8/10	1/10
moderate	3/10	4/10	3/10
equal	1/3	1/3	1/3

Table 6.1: (Two-dimensional Branin function) Probablity distributions for environmental variable x_2

The true two-dimensional surfaces, for each support point of the environmental variable, are plotted in Figure 6.1(a) (page 137) on $m = 100$ equally spaced points which coincide with the m points used to evaluate the designs in (6.15) and the input space of the control variable x_1. The number of starting design points, N_0, is 3 and the final number of input points, N, is taken to be 20.

Function 2: Four-dimensional Branin function

Next, we have a surface with two control variables and two environmental variables (with a total of twelve support points). The response for this surface is evaluated via $y(\boldsymbol{x}) = f(x_1, x_2) \times f(x_3, x_4)$ and

$$f(u, v) \;=\; (v - \frac{5.1}{4\pi^2}u^2 + 5/\pi u - 6)^2 + 10(1 - \frac{1}{8\pi}\pi)\; cos(u) + 10$$

where $u \in [-5, 10]$, $v \in [0, 15]$. We take $\{x_1, x_4\}$ to be the control variables, $\boldsymbol{x}_c$, and $\{x_2, x_3\}$ to be the environmental variables $\boldsymbol{x}_e$. The joint distribution for the environmental variables is given in Table 6.2 (this is considered a set of moderate weights). For equal weights, the entries are replaced by 1/12 (not shown in table). As before, the input domain for the control variables is finely divided into $m = 30 \times 30 = 900$ equally spaced points. The number of starting design points, N_0, is 12 and the final number of input points, N, is taken to be 40.

		x_3			
		0.2	0.4	0.6	0.8
x_2	0.25	0.0375	0.0875	0.0875	0.0375
	0.5	0.0750	0.1750	0.1750	0.0750
	0.75	0.0375	0.0875	0.0875	0.0375

Table 6.2: (Four-dimensional Branin function) Joint probablity distribution for environmental variables x_2 and x_3

Function 3: Three-dimensional exponential-sine function

We consider the function

$$y(x_1, x_2, x_3) = c_1 exp\left(c_2 \frac{x_1 + x_2}{3}\right) + c_3 x_2 sin(x_3 x_1) + c_2 x_3 x_2$$

where $c_1 = 2, c_2 = 1, c_3 = 100$ are taken as known constants and $x_1, x_2 \in [0, 3]$ are the control variables. The distributions of the environmental variable x_3 are given in Table 6.3. The input domain for the control variables is finely divided into $m = 30 \times 30 = 900$ equally spaced points. The number of starting design points, N_0, is 5 and the final number of input points, N, is taken to be 30.

	x_3				
	0.5	1.0	1.5	2.0	2.5
unequal	1/28	3/28	20/28	3/28	1/28
moderate	2/18	4/18	6/18	4/18	2/18
equal	1/18	1/18	1/18	1/18	1/18

Table 6.3: (Three-dimensional exponential-sine function) Probablity distributions for environmental variable x_3

Function 4: Three-dimensional sine function

Finally, we consider the function

$$y(x_1, x_2, x_3) = \theta_1 sin(\theta_2 \pi x_1) + \theta_3 sin(x_3 \pi x_2)$$

where $\theta_1 = 100,\ \theta_2 = 2,\ \theta_3 = 100$ are taken as known constants and $x_1, x_2 \in [0, 3]$ are the control variables. The distributions of the environmental variable x_3 are given in Table 6.4. The input domain for the control variables is finely divided into

$m = 30 \times 30 = 900$ equally spaced points. As with Function 3, $N_0 = 5$ and N, is taken to be 30.

			x_3		
	1.0	1.5	2.0	2.5	3.0
unequal	1/28	3/28	20/28	3/28	1/28
moderate	2/18	4/18	6/18	4/18	2/18
equal	1/18	1/18	1/18	1/18	1/18

Table 6.4: (Three-dimensional sine function) Probablity distributions for environmental variable x_3

6.3.2 Results: Comparison of Design Criteria

Results from our simulation studies, in Figure 6.2 (page 138), show that there are differences in the predictive accuracy of the different designs depending on the shape of the response surfaces and the distribution of the environmental variables.

We first consider the case of a uniform distribution for the environmental variable where $w_j = \frac{1}{n_e}, j = 1, \cdots, n_e$ (it should be noted that the IEIGF and IEIGFw criteria are equivalent in this case). The leftmost region of the plots in Figure 6.2 (c), (e), (h) and (k), on page 138, separated by the dotted vertical lines show equal proportions both positive and negative differences in the ERMSPE. This suggests that there do not seem to be any significant differences between the EIGF and wEIGF criteria. The plots suggest that the IEIGF criterion outperforms both the EIGF and wEIGF criteria. The center and right regions of the plots in Figure 6.2 (c), (e), (h) and (k) separated by the dotted vertical lines show larger positive difference in ERMSPE for the EIGF versus. IEIGF and wEIGF versus IEIGF comparisons.

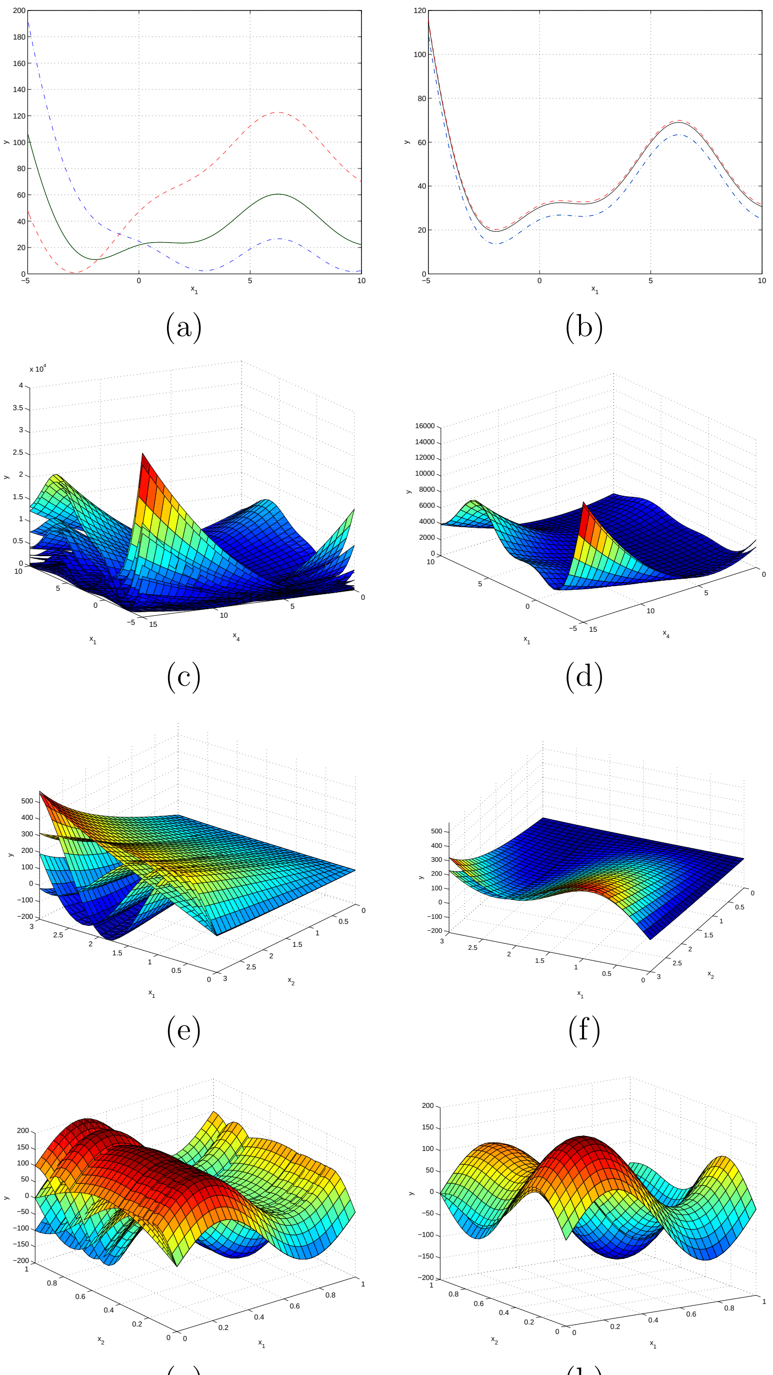
For the moderate and unequal weight distributions, the weighted versions of the criteria improves the predictive performance of the designs in most cases (i.e., a larger number of positive ERMSPE differences among the 30 runs). The extreme left region in plots (a), (b), (d), (f), (g), (i) and (j) displays the ERMSPE differences between the EIGF and wEIGF criteria, while the extreme right region shows the IEIGF and IEIGFw diffferences. There is a clear trend that the improvement increases (i.e., change in the magnitude of the positive ERMSPE differences) as the weights become more unequal.

Given that the wEIGF and IEIGFw criteria are the better performers among the four designs, we will compare their performances across the different response surfaces and weight distributions. For the moderate weights situation (see plots (b), (d), (g) and (j) in Figure 6.2), the IEIGFw criterion dominates the wEIGF criterion in terms of the number and magnitude of positive ERMSPE differences (see region second from the right). However, it is not so clear whether the wEIGF criterion or IEIGFw criterion is better when the weights for the environmental variables become more unequal in Figure 6.2(a), (f) and (i). A closer look at the true surfaces in Figure 6.1 shows that if the shape of the averaged surface (in the right column) is significantly different from the shape of the surfaces for the different inputs of the environmental variables, then the IEIGFw criterion tends to perform better.

6.4 Discussion and Conclusion

In this chapter we have considered the case of a single deterministic response $y(\cdot)$ that depends on both control and environmental variables, and we are interested in achieving good global model fit for the integrated response surface over the distribution of the environmental input variables. Refinements to the EIGF criterion, introduced in the previous chapters with appropriate weighting schemes have been shown to perform well. Among the four sequential adaptive designs, the wEIGF and IEIGFw criteria are very competitive in terms of predictive accuracy using the GASP model with the cubic correlation function. Overall, the IEIGFw criterion is our preferred method given that it performs well for equal and moderate weights, and does not perform too badly in situations where the weights are highly unequal.

For the IEIGFw criterion, we have only considered selecting the environmental variables using (6.14) in our examples. Further refinements can be made to this selection approach. Additional criteria can also be proposed and this is open to further research.



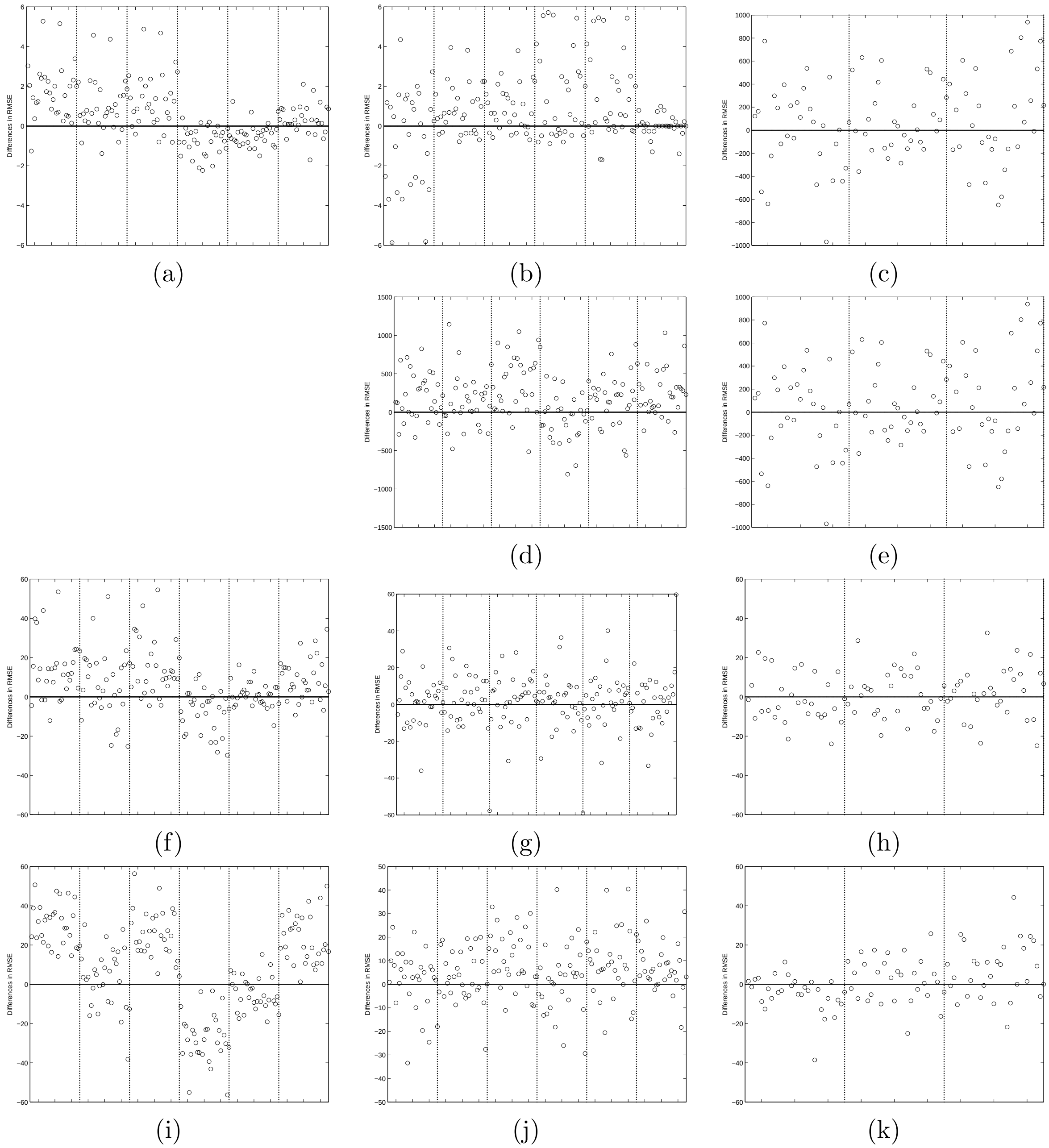
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 6.1: Plots of true surfaces (left column) for various support points of the environmental variable(s) and averaged surfaces (right column) for different probability distributions for the environmental variables.
(a),(b): Function 1, two-dimensional Branin function; (c),(d): Function 2, four-dimensional Branin function; (e),(f): Function 3, three-dimensional exponential-sine function; (g),(h): Function 4, three-dimensional sine function.



(a) (b) (c)

(d) (e)

(f) (g) (h)

(i) (j) (k)

Figure 6.2: Plots of ERMSPE differences for pairwise comparisons of design criteria. (First row) Function 1: Two-dimensional Branin function, (Second row) Function 2: Four-dimensional Branin function, (Third row) Function 3: Three-dimensional exponential-sine function, (Fourth row) Function 4: Three-dimensional sine function. (Left column: a,f,i) unequal weights*, (Middle column: b,d,g,j) moderate weights*, (Right column: c,e,h,k) equal weights$^+$.

*Unequal and moderate weights: [*regions left to right separated by vertical dotted lines*] ERMSPE differences for EIGF-wEIGF, EIGF-IEIGF, EIGF-IEIGFw, wEIGF-IEIGF, wEIGF-IEIGFw, IEIGF-IEIGFw.

$^+$Equal weights: ERMSPE differences for EIGF-wEIGF, EIGF-IEIGF, wEIGF-IEIGF.

CHAPTER 7

CONCLUSION AND FUTURE RESEARCH

Williams (2000) and Lehman (2002) have shown that sequential design strategies are valuable for computer experiments in the context of global optimization. This thesis continues to contribute to this area of research and, for the objective of achieving good global model fit of the response surface, has shown that sequential adaptive designs provide an efficient alternative to fixed-point designs such as the maximin LHD used in our examples. We have shown in Chapters 3 and 4 that the adaptive property of the sequential adaptive design criteria 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. Possibilities for future research include incorporating sequential adaptive designs into the model calibration (i.e. estimation of simulator input parameters) and prediction process. As part of an overall model verification and validation process, a Bayesian approach (see Higdon *et al.*, 2004) can be implemented to combine model simulation output and physical observations to estimate model parameters. These, in turn, can be used for predicting the response at some untried input points. One of the goals would be to reduce prediction uncertainty through more accurate computer outputs.

Most of the research in computer experiments have adopted the approach of assuming a single stationary Gaussian Stochastic Process (GASP) model across the entire input space in the design and analysis stages. 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. Future research, as mentioned in Chapter 4, might explore combining sequential adaptive designs with more complicated stochastic models, such as the Bayesian Treed approach by Gramacy (2005).

Results from our simulation studies have shown that the choice of design is crucial to the success of building an efficient and accurate GASP model. In addition to designs, we have considered the use two types of correlation functions, namely the Gaussian and cubic correlation. Given the instability encountered in fitting GASP model with the Gaussian correlation, Chapter 5 presents the use of a cross validation approach for assessing the fit of the GASP model with the Gaussian correlation and our examples have illustrated improvements in predictive accuracy of the GASP model with the modified design algorithm. Other types of correlation functions and diagnostic tools for model goodness of fit can be explored for further research.

In extending the various sequential adaptive designs to more complex problems such as a situation where both control and environmental (noise) variables may be present, we have shown that the designs have been effective in achieving good model fit for integrated response surfaces averaged over the probability distribution of the environmental variables. In cases where the probability distribution of the environmental variables are not uniform, it is important that a design criterion should take into account the weights.

In conclusion, we are encouraged by the positive results in this thesis and optimistic that sequential adaptive designs are potentially more effective and efficient for prediction of responses at unobserved input points than fixed-point designs.

APPENDIX A

DERIVATIONS AND THEOREMS

A.1 Theorems for Probablity Distributions

Theorem A.1. Suppose $\boldsymbol{U}_i$ for $i \in 1, 2$ denote $q_i \times 1$ random vectors having the Gaussian distribution

$$\begin{pmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{pmatrix} \mid \boldsymbol{\beta}, \sigma^2 \sim N_{q_1,q_2}\left[\begin{pmatrix} \boldsymbol{F}_1 \\ \boldsymbol{F}_2 \end{pmatrix}\boldsymbol{\beta}, \sigma^2 \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{pmatrix}\right] \tag{A.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}_k$ and $\sigma^2 > 0$. Assuming that each of the elements of $\boldsymbol{F}_i$ and $\boldsymbol{R}_{ij}$ are known, each $\boldsymbol{F}_i$ has full column rank, and the correlation matrix is positive denite. Then

$$[\boldsymbol{U}_1 | \boldsymbol{U}_2, \boldsymbol{\beta}, \sigma^2] \sim N_{q_1}(\boldsymbol{m}_{1|2}, \sigma^2 \boldsymbol{R}_{1|2}),$$

where $\boldsymbol{m}_{1|2} = \boldsymbol{F}_1\boldsymbol{\beta} + \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}(\boldsymbol{U}_2 - \boldsymbol{F}_2\boldsymbol{\beta})$, and $\boldsymbol{R}_{1|2} = \boldsymbol{R}_{11} + \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{12}$.

Proof. The proof can be found in most standard books on linear models such as Seber(1977).

A.2 Proof of Sequential Maximum MSPE and Maximum Entropy Criteria Are Equivalent

Recall that if the prior distribution for $\boldsymbol{\beta}$ are diffuse (i.e. $\tau^2 \to \infty$), the maximum entropy criterion is given by (2.6),

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

where $\boldsymbol{F} = (\boldsymbol{F}_n, 1)^T$ and $\boldsymbol{F}_n$ is a $n \times 1$ vector of 1's and n is the number of observations.

Suppose we implement the sequential maximum entropy criterion (Subsection 2.3.2) with one point added at each time. The correlation matrix $\boldsymbol{R}$ (which now includes the candidate point, $\boldsymbol{x}_0$) can be partitioned into

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_n & r_n(\boldsymbol{x}_0) \\ r_n(\boldsymbol{x}_0)^T & 1 \end{pmatrix}, \tag{A.3}$$

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 $r_n(\boldsymbol{x}_0)$. As a result, the determinant of $\boldsymbol{R}$ can be written as a product of $\det(\boldsymbol{R}_n) \times \det(1 - r^T(\boldsymbol{x}_0)\boldsymbol{R}_n^{-1} r(\boldsymbol{x}_0))$. The inverse of R can be expressed as

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{R}_n^{-1} + \frac{\boldsymbol{R}_n^{-1} r_n(\boldsymbol{x}_0)\, r(\boldsymbol{x}_0)^T \boldsymbol{R}_n^{-1}}{1 - r_n(\boldsymbol{x}_0)^T \boldsymbol{R}^{-1} r_n(\boldsymbol{x}_0)} & \frac{-\boldsymbol{R}_n^{-1} r_n(\boldsymbol{x}_0)}{1 - r_n(\boldsymbol{x}_0)^T \boldsymbol{R}^{-1} r_n(\boldsymbol{x}_0)} \\ \frac{-r_n(\boldsymbol{x}_0)^T \boldsymbol{R}_n^{-1}}{1 - r_n(\boldsymbol{x}_0)^T \boldsymbol{R}^{-1} r_n(\boldsymbol{x}_0)} & \frac{1}{1 - r_n(\boldsymbol{x}_0)^T \boldsymbol{R}^{-1} r_n(\boldsymbol{x}_0)}. \end{bmatrix} \tag{A.4}$$

See Rao (2001) page 33 for details on taking inverse of matrices.

Using (A.4), the determinant of $\boldsymbol{R}$ (suppressing the notation $r_n(\boldsymbol{x}_0)$ as r_n) can be re-expressed as

$$\begin{aligned} &= \det(\boldsymbol{R}_n) \times \det(1 - r_n^T \boldsymbol{R}_n^{-1} r_n) \\ &\times \det\Big(\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n + \frac{\boldsymbol{F}_n^T \boldsymbol{R}^{-1} r_n r_n^T \boldsymbol{F}_n}{1 - r_n^T \boldsymbol{R}_n^{-1} r_n} - \frac{r^T \boldsymbol{R}^{-1} \boldsymbol{F}_n}{1 - r_n^T \boldsymbol{R}_n^{-1} r_n} - \frac{\boldsymbol{F}_n^T \boldsymbol{R}^{-1} r}{1 - r_n^T \boldsymbol{R}_n^{-1} r_n} + \frac{1}{1 - r_n^T \boldsymbol{R}_n^{-1} r_n}\Big) \\ &= \det(\boldsymbol{R}_n) \times \; \boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n \\ &\times \Big[(1 - r_n^T \boldsymbol{R}_n^{-1} r_n) + \frac{\boldsymbol{F}_n^T \boldsymbol{R}^{-1} r_n r_n^T \boldsymbol{F}_n}{\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n} - \frac{r^T \boldsymbol{R}^{-1} \boldsymbol{F}_n}{\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n} - \frac{\boldsymbol{F}_n^T \boldsymbol{R}^{-1} r_n}{\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n} + \frac{1}{\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n}\Big] \end{aligned} \tag{A.5}$$

After completing the squares, we get

$$= \det(\boldsymbol{R}_n) \times \boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n \times \Big[(1 - r_n^T \boldsymbol{R}_n^{-1} r_n) + \frac{(1 - \boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} r_n)^2}{\boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n}\Big] \tag{A.6}$$

It is clear from this expression the sequential maximum entropy criterion (assuming a diffuse prior for $\boldsymbol{\beta}$) is the sequential maximum MSPE criterion (2.7) multiplied by a constant factor $\det(\boldsymbol{R}_n) \times \boldsymbol{F}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{F}_n$ which are the first two terms in (A.6) . Hence, the two criteria are equivalent.

A.3 Derivation for Expected Improvement for Global Fit Criterion

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

$$I(\boldsymbol{x}) = (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 \tag{A.7}$$

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

$$E(I(\boldsymbol{x})) = (\hat{Y}(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 + var(\hat{Y}(\boldsymbol{x})). \tag{A.8}$$

Proof: Taking the expected value of (A.7) yields

$$\begin{aligned} \mathrm{E}(Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^2 &= \mathrm{E}[Y^2(\boldsymbol{x}) - 2y(\boldsymbol{x}_{j^*})\mathrm{E}(Y(\boldsymbol{x})) + y^2(\boldsymbol{x}_{j^*})] \\ &= \mathrm{var}(\hat{Y}(\boldsymbol{x})) + \hat{Y}^2(\boldsymbol{x}) - 2y(\boldsymbol{x}_{j^*})\hat{Y}(\boldsymbol{x}) + y^2(\boldsymbol{x}_{j^*}) \\ &= \mathrm{var}(\hat{Y}(\boldsymbol{x})) + (\hat{Y}^2 - y(\boldsymbol{x}_{j^*}))^2 \end{aligned} \tag{A.9}$$

where we recall that $Y(\boldsymbol{x}) \sim \mathrm{N}(\hat{Y}(\boldsymbol{x}), \mathrm{var}(\hat{Y}(\boldsymbol{x})))$ where $\hat{Y}(\boldsymbol{x})$ is the BLUP (1.7) and $\mathrm{var}(\hat{Y}(\boldsymbol{x}))$ is the MSPE in (1.8).

A.4 Derivation for Generalized Expected Improvement for Global Fit Criterion

Following Section A.3, the "generalized" expected improvement, for each candidate input point $\boldsymbol{x}$, is derived as follows. First, we let the generalized expected improvement, $E(I^g(\boldsymbol{x})) = EI^q(\boldsymbol{x})$ where

$$\begin{aligned} EI^q(\boldsymbol{x}) &= \int_{-\infty}^{+\infty} (Y(\boldsymbol{x}) - y(\boldsymbol{x}_{j^*}))^q \ \phi(y)dy \\ &= \int_{-\infty}^{+\infty} \sum_{k=0}^{q} (-1)^k \left(\begin{array}{c} q \\ k \end{array} \right) Y^k(\boldsymbol{x}) \ (y(\boldsymbol{x}_{j^*}))^{q-k} \ \phi(y)dy \\ &= \sum_{k=0}^{q} (-1)^k \left(\begin{array}{c} q \\ k \end{array} \right) (y(\boldsymbol{x}_{j^*}))^{q-k} \int_{-\infty}^{+\infty} Y^k(\boldsymbol{x}) \ \phi(y)dy \\ &= \sum_{k=0}^{q} (-1)^k \left(\begin{array}{c} q \\ k \end{array} \right) (y(\boldsymbol{x}_{j^*}))^{q-k} E(y^q(\boldsymbol{x})) \end{aligned} \tag{A.10}$$

where $q = 1, 2, ...$ and $E(Y^q)$ denotes q^{th} moments of the Normal distribution. For notational simplicity, we denote $\hat{Y}(\boldsymbol{x})$ and $Y(\boldsymbol{x})$ as $\hat{Y}$ and Y respectively. Recall that $Y \sim \mathrm{N}(\hat{Y}, \mathrm{var}(\hat{Y}))$ where $\hat{Y}$ is the EBLUP version of (1.7) and $\mathrm{var}(\hat{Y})$ is the MSPE given in (1.8). The first four moments are given by

$$\begin{array}{l} E(Y^0) = 1 \\ E(Y^1) = \hat{Y} \\ E(Y^2) = \hat{Y}^2 + \mathrm{var}(\hat{Y}) \\ E(Y^3) = \hat{Y}^3 + 3\hat{Y}\mathrm{var}(\hat{Y}) \\ E(Y^4) = \hat{Y}^4 + 6\hat{Y}^2\mathrm{var}(\hat{Y}) + 3(\mathrm{var}(\hat{Y}))^2 \end{array} \tag{A.11}$$

Higher moments can be easily derived using a recursive formula. If $X \sim N(\mu, 1)$, then the $(n+1)^{th}$ moment is given by

$$EX^{n+1} \ = \ \mu EX^n + \frac{d}{d\mu} EX^n.$$

See Casella and Berger (2001) for more details.

Using (A.10) and (A.11) for the special case of $q = 4$, we obtain the *generalized expected improvement for global fit* criterion, $E(I^g(\boldsymbol{x}))$, for $g = q/2 = 2$ as

$$\begin{aligned} E(I^2(\boldsymbol{x})) &= y^4(\boldsymbol{x}_{j^*})E(Y^0) - 4y^3(\boldsymbol{x}_{j^*})E(Y^1) + 6y^2(\boldsymbol{x}_{j^*})E(Y^2) - 4y(\boldsymbol{x}_{j^*})E(Y^3) + E(Y^4) \\ &= y^4(\boldsymbol{x}_{j^*}) - 4y^3(\boldsymbol{x}_{j^*})\hat{Y} + 6y^2(\boldsymbol{x}_{j^*})\hat{Y}^2 - 4y(\boldsymbol{x}_{j^*})\hat{Y}^3 + \hat{Y}^4 + \cdots \\ &\quad [\ 6y^2(\boldsymbol{x}_{j^*}) - 12y(\boldsymbol{x}_{j^*})\hat{Y} + 6\hat{Y}\ ]\mathrm{var}(\hat{Y}) + 3(\mathrm{var}(\hat{Y}))^2 \\ &= (y(\boldsymbol{x}_{j^*}) - \hat{Y})^4 + \ 6(y(\boldsymbol{x}_{j^*}) - \hat{Y})^2 \ \mathrm{var}(\hat{Y}) \ + \ 3(\mathrm{var}(\hat{Y}))^2. \end{aligned} \tag{A.12}$$

Taking $g = 1$ (i.e., $q = 2$), we get the original EIGF criterion, $E(I(\boldsymbol{x}))$, in (A.8) and (3.10).

The expected improvement in (A.12) 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 (1.8), i.e., points about which there is large uncertainty and, as mentioned in Subsection 2.3.1, these tend to be far from existing sampled points. The key difference is that there is now interaction between $(y(\boldsymbol{x}_{j^*}) - \hat{Y})^2$ and $\mathrm{var}(\hat{Y})$.

BIBLIOGRAPHY

Bates, R. A., Buck, R. J., Riccomagno, E., and Wynn, H. P. (1996). Experimental design and observation for large systems (Disc: P95-111). *Journal of the Royal Statistical Society, Series B: Methodological*, 58:77–94.

Beers, van WCM and Kleijnen, JPC (2004). Kriging interpolation in simulation: a survey. In R.G. Ingralls, M.D. Rossetti, J. S. and Peters, B., editors, *Proceedings of the 2004 Winter Simulation Conference*.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96:1361–1374.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association*, 54:622–654.

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

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury Press, 2nd edition.

Cox, D. D. and John, S. (1997). SDO: a statistical method for global optimization. In *Multidisciplinary design optimization (Hampton, VA, 1995)*, pages 315–329. SIAM, Philadelphia, PA.

Cressie, N. C. (1993). *Statistics for Spatial Data*. Wiley, New York.

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

Drignei, D. (2006). Empirical Bayesian analysis for high-dimensional computer output. *Technometrics*, 48(2):230–240.

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

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.

Gramacy, R. B. and Lee, H. K. H. (2006). Bayesian treed Gaussian process models. Technical report, Dept. of Applied Math & Statistics, University of California, Santa Cruz.

Halton, J. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.*, 2:84–90.

Handcock, W. J. (1991). On cascading latin hypercube designs and additive models for experiments. *Communications Statistics—Theory Methods*, 20:417–439.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466.

Jin, R., Chen, W., and Sudjianto, A. (2002). On sequential sampling for global metamodeling in engineering design. In *Proceedings of DETC 2002. ASME 2002 Design Engineering Technical Conferences And Computers and Information in Engineering Conference.*

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

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

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(3):425–464.

Kleijnen, JPC and Beers, van WCM (2004). Application-driven sequential designs for simulation experiments: Kriging metamodelling. *Journal of the Operational Research Society*, 55(8):876–883.

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.

Lehman, J. S. (2002). *Sequential Designs of Computer Experiments for Robust Parameter Design*. PhD thesis, The Ohio State University.

Lim, Y. B., Sacks, J., Studden, W. J., and Welch, W. J. (2002). Design and analysis of computer experiments when the output is highly correlated over the input space. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(1):109–126.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Annuals of Mathemtical Statistics*, 27:986–1005.

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

Matérn, B. (1960). *Spatial Variation*. PhD thesis, Meddelanden fran Statens Skogsforskningsinstitut.

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.

Mitchell, T. J. and Morris, M. D. (1992). Bayesian design and analysis of computer experiments: Two examples. *Statistica Sinica*, 2:359–379.

Mitchell, T. J. and Scott, D. S. (1987). A computer program for the design of group testing experiments. *Communications in Statistics: Theory and Methods*, 16:2943–2955.

Morris, M. D. and Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43:381–402.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 66(3):751–769.

Park, J.-S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39:95–111.

Ranjan, P., Binghan, D., and Michailidis, G. (2007). Sequential experiment design for contour estimation from complex computer codes. *Technometrics, to appear.*

Rao, C. R. (1993). *Linear Statistical Inference and its Applications.* Wiley-Interscience, 2nd edition.

Sacks, J. and Schiller, S. (1988). Spatial designs. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics IV, in two volumes, Volume 2*, pages 385–395. Springer-Verlag Inc.

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

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989b). "Design and Analysis of Computer Experiments" (with comments, p423-435). *Statistical Science*, 4:409–423.

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

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

Seber, G. A. F. (1973). *Linear Regression Analysis.* John Wiley and Sons.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 62–656.

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

Sobol´, I. M. (1993). Sensitivity analysis for non linear mathematical models. *Mathematical Model. Comput. Exp.*, 1:407–414.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging.* New York, Springer.

Welch, W. J. (1985). Aced: algorithms for the construction of experimental designs. *American Statistician*, 39:146.

Williams, B. J., Santner, T. J., and Notz, W. I. (2000). Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, 10(4):1133–1152.