Design on Non-Convex Regions: Optimal Experiments for Spatial Process Prediction

by

Matthew Timothy Pratola

B.Sc., 2005.

Brock University

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Summer 2006

APPROVAL

Name: Matthew Timothy Pratola

Degree: Master of Science

Title of project: Design on Non-Convex Regions: Optimal Experiments for Spatial Process Prediction

Examining Committee: Dr. Carl Schwarz
Chair

Dr. Derek Bingham
Senior Supervisor
Simon Fraser University

Dr. Charmaine Dean
Simon Fraser University

Dr. Randy Sitter
External Examiner
Simon Fraser University

Date Approved: August 2, 2006

Abstract

Modeling a response over a non-convex design region is a common problem in diverse areas such as engineering and geophysics. Unfortunately, the tools available to model and design for such responses are limited. Recently, some success has been found by applying the Gaussian Process (GP) model with the so-called water distance metric. However, a difficulty is that transformation of the water distances is required to be able to model a GP over such regions. The specific questions of exactly how to make this transformation, select design points and fit GP models have received little attention. In this thesis, we build on existing results to propose a valid transformation. A new method for selecting design points with the GP model over non-convex regions is then proposed. Optimal designs for prediction are described, and a simulation study is used to demonstrate the improvements that are realized.

Keywords: optimal design; non-convex; spatial; Gaussian; process; ISOMAP

Acknowledgements

First, I would like to thank my parents for their support, and especially the high quality education they afforded me during my childhood. I would likely not be working in academia today without the excellent foundation laid in my early schooling.

I would like to thank Sandy, Crystal and Chun-fang for their friendship. I would also like to thank my non-statistics friends including Mike, Jodie, Kiran, Leon and Ed for the good times over the years.

I would also like to thank Dr. John Kern from the Department of Mathematics and Computer Science at Duquesne University for providing the Florida data-set and patiently answering related questions.

I will always be very thankful and indebted to my undergraduate supervisor and friend Dr. Thomas Wolf. I would not be where I am today without his help and belief in me.

I would like to sincerely thank Pritam Ranjan for his friendship and the valuable discussions over the past three years. There are very few people I hold in such high regard. My most heartfelt thanks go to Natasha Lysenko for her help and understanding during a difficult time. This would not have been possible without her.

Finally, I am most thankful and grateful to my supervisor, Dr. Derek Bingham, for his continuing patience, guidance and support during my graduate education.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivating Applications

Scientists are often interested in determining the relationship between experimental factors and response variables. This is a common endeavour in most areas of scientific investigation (e.g., engineering and geophysics). For instance, an engineer may be trying to maximize the yield of a production process that depends on many machine factor settings. Finding the machine settings that result in the maximum yield is a difficult problem when considering a high-dimensional factor space, a complex response function, the presence of noise in our response measurements and the high cost (in dollars or time) of obtaining experimental data. However, if an approximate functional relationship between the yield (response variable) and machine factors can be found that closely models the true functional behaviour, then one can attempt to meet the experiment's goals. Indeed, the approach known as Response Surface Methodology (Box and Draper, 1987) from the experimental design literature is such a strategy.

To help establish the relationship between the factors and the response, an experimenter needs to obtain values of the response at different factor settings. A fundamental statistical consideration is the choice of model used to summarize this relationship. Assuming a model can be chosen that represents reality well, the design question is to choose settings for the factor variables that best allow the model to

help answer the scientific question of interest.

A common example in the field of designed experiments is that of polynomial regression. Typical designs for such models include fractional factorial designs, which are used to control run size in a manner that minimizes bias in parameter estimates, and optimal designs which are criterion-based approaches, such as selecting designs that minimize the variance of parameter estimates. Once the model is fit to the data, we can go forth and attempt to answer questions of interest under the assumed linear model. If we have selected an appropriate design and our model fits the data well, we can attempt to address the question of interest with reasonable confidence. The selection of a design is therefore integral to modeling the response.

The usual approach to a designed experiment is to select levels of the factors under the assumption that these variables are independent of one another. Geometrically, this can be interpreted as a rectangular design region. For instance, suppose we have the two design factors temperature, in degrees celcius (°C), and pressure, in kiloPascals (kPa). If we are interested in the levels (5°C, 10°C) for temperature and (100 kPa, 104 kPa) for pressure, then we have defined a square 2-dimensional design space $\Omega \in \Re^2$. We could write these levels as $(0, 1)$ for both variables and obtain, say, a 2^2 factorial design for fitting a linear regression model. However, in practice it is not always the case that design variables are independent of one another. That is, the geometric region formed by the design factors may not be rectangular. Design variable dependence can occur quite naturally in many settings, such as:

- inherent physical limitations, also known as bad region avoidance (Taguchi, 1987; Hamada and Wu, 1995)
- modeling of natural phenomena, such as river temperature.

The bad region avoidance case can occur in industrial experiments where a response of interest can only be measured at certain settings of the design variables but not others. An example is shown in Figure 1.1, where design variables x_1 and x_2 have a non-linear relationship. We can imagine a response, such as % failed parts, may exhibit reasonable behaviour within the design region but which quickly approaches 100% failure outside the region due to the interplay between the dependent design

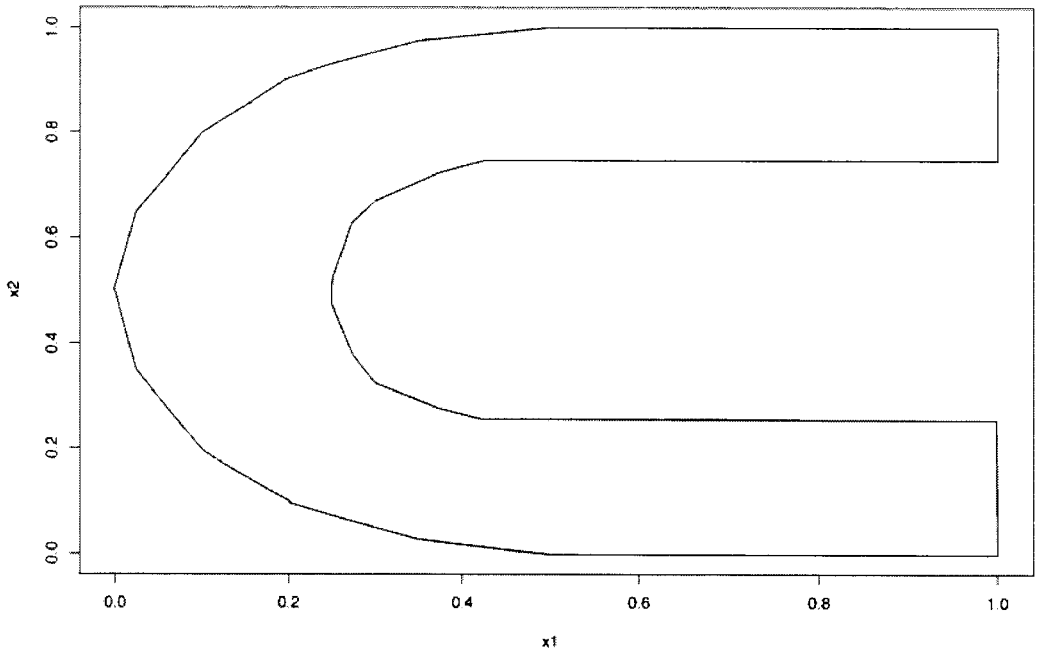


Figure 1.1: Horseshoe design region $\Omega \in \Re^2$

variables and the response in question. Or, it may be that measuring the response outside of this region is simply impossible due to physical laws. This is represented visually by the hypothetical response shown in Figure 1.2. Therefore the bad region avoidance rational is motivated by expert knowledge that suggests investigating the complete $[0,1]^2$ space shown in Figure 1.2 is at best of no interest, or, at worse meaningless. In other words, the behaviour here is much different from fitting a linear model over $[0,1]^2$ and then simply considering the fitted model only over the region, Ω, shown in Figure 1.1. Intuitively, the behaviour of the response should be modeled as a function of the region Ω, thereby suggesting that the coordinate system defined by (x_1, x_2) may not be the natural coordinate system for this problem. Perhaps a better way of modeling a response such as Figure 1.2 is to suggest a contrived coordinate system of some sort that is more descriptive of the horseshoe shape of the region. For example, we might define the coordinate system x_a: the centerline along the horseshoe, and x_b: the distance from this centerline. Under this coordinate system, we *can*

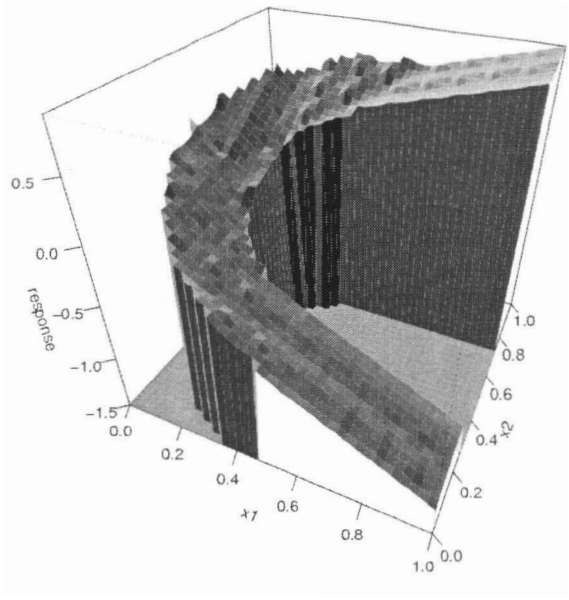


Figure 1.2: Response over region of dependent design variables

extract meaningful information as it would appear that the function increases in x_a but has little change in x_b.

The intuition for this type of modeling problem becomes more clear when we consider the case of modeling a natural phenomena such as river temperature. Suppose the temperature depends on the distance from the shore and the velocity of the water in the direction of flow. Then it is quite intuitive that fitting a model using the (lattitude, longitude) coordinates of points along the river might not make sense.

In the next two sections, real examples are used from the literature to illustrate the problem of modeling over such non-rectangular design regions. The first is an engineering application of spot welding and the second example considers ocean temperatures around a peninsula.

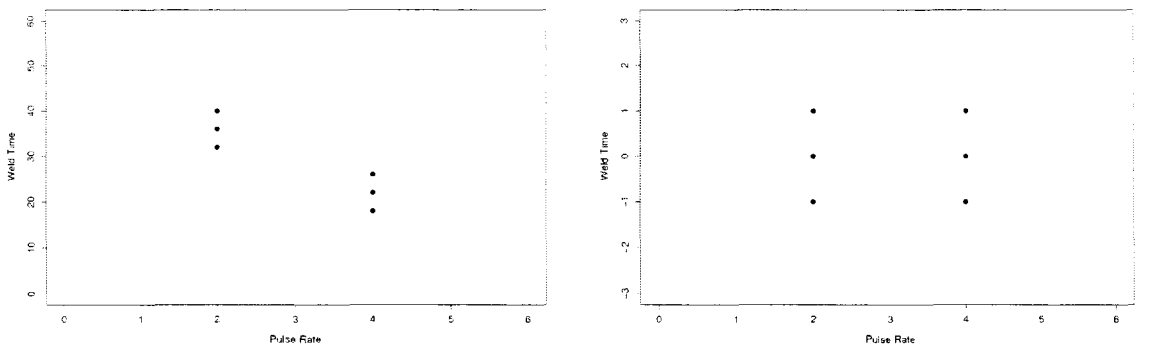
Table 1.1: Design Variable Levels for Welding Experiment

Factors		Levels		
A	Pulse Rate	2	4	-
B	Weld Time	Low	Med	High
C	Cool Time	6	12	18
D	Hold Time	10	18	26
E	Squeeze Time	15	20	25
F	Air Pressure	50	55	60
G	Current %	85	90	95
H	Tip Size	$\frac{3}{8}''$	$\frac{1}{4}''$	$\frac{1}{4}''$

1.1.1 Spot Welding Experiment

In (Chen et al., 1984), the authors outline a welding experiment investigating 8 design varibles to find an optimum weld strength. A manufacturer of engine support brackets was having trouble with weld strength consistency, and the subsequent failure of many brackets in the field. Of the 8 relevant design variables, six were investigated at 3 levels and two at 2 levels, as shown in Table 1.1. In this experiment, the levels of Weld Time are dependent on the levels of variable Pulse Rate as shown in Figure 1.3(a). An orthogonal array (Hedayat et al., 1999) was used to run a designed experiment, and standard ANOVA was done using weld strength as the response. Level 3 for factor H in the 3-level orthogonal array was set to equal level 1 since factor H is actually a 2-level factor.

The dependence shown in Figure 1.3(a) is an example of the bad region avoidance rational. In this example, it is not feasible to have a weld time of 40 with a pulse rate of 4. For instance, a long weld time and high pulse rate may result in the metal reaching too high a temperature and melting to the extent that no weld is formed. In contrast, too short a weld time with a low pulse rate may not heat the metal sufficiently to form a weld of the desired size and strength. The original authors of this study (Chen et al., 1984) do not consider the dependence between weld time (A) and pulse rate (B), instead modeling the data as if the (low, medium, high) levels of B are independent of the level of A. This leads to the question of interpretation for



(a) (b)

Figure 1.3: (a) Dependence Between Design Factors Pulse Rate and Weld Time for Welding Experiment and (b) Recoded Design Factors

parameter estimates of main effect B and the $A \times B$ interaction term considered in their model.

In (Hamada and Wu, 1995), the authors re-analyze this experiment taking into account the dependence between factors A and B. They propose a solution to this problem by viewing it as a nested experiment, and performing design and analysis accordingly. They noted improvements in finding factor levels which lead to an optimal response when re-analyzing the welding experiment. In their paper, the dependence of the two related design variables X_A and X_B must satisfy the following relation:

$$E[Y] = f_1\left(\frac{X_A - c_A}{s_A}\right) + f_2\left(\frac{X_B - c_B(X_A)}{s_B(X_A)}\right), \tag{1.1}$$

where c_A is a centering constant, s_A is a scale constant, $c_B(X_A)$ is a centering parameter dependent on the level of X_A and $s_B(X_A)$ is a scale parameter dependent on the level of X_A. This forms a linear relationship for the dependence between variables X_A and X_B which can be seen by writing $X'_B(X_A) = \frac{X_B - c_B(X_a)}{s_B(X_A)} = s_B(X_A)^{-1}X_B + K_B(X_A)$. In words, if we think of the centering and scaling parameters as normalizing a design variable to say (low, medium, high) settings, then the relationship (1.1) says that the actual values of X_B for $X'_B = (low, medium, high)$ change according to the value of X_A. This can be viewed pictorially as in Figure 1.3(a) where the dependence between factors pulse rate and time is removed by applying the scaling and centering

transformation to arrive at the re-coded factors shown in Figure 1.3(b).

Unfortunately, this approach does not provide a predictive model since the values of $s_B(X_A), K_B(X_A)$ are not themselves modeled. That is, we have no knowledge for how the assumed linear relationship between X_A and X_B varies for arbitrarily chosen X_A, or indeed whether a linear assumption is realistic. Instead, a set of conditional linear models are formed which depend on the level of X_A. Even if we have knowledge of the levels of X_B when $X_A = 1$ and when X_A=2, we still cannot predict the response when, say, $X_A = 1.5$.

Let us examine for a moment the linear relationship outlined above for known $s_B(X_A), K_B(X_A)$ at a given level of X_A. Then a second order model for $E[Y]$ is

$$\begin{aligned} E[Y] &= \beta_{A1}X'_A + \beta_{A2}{X'_A}^2 + \beta_{B1}X'_B(X_A) + \beta_{B2}{X'_B}^2(X_A) \\ &= \beta_{A1}X'_A + \beta_{A2}{X'_A}^2 + \beta_{B1}\left(s_B(X_a)^{-1}X_B + K_B(X_A)\right) + \\ &\quad \beta_{B2}\left(s_B(X_A)^{-1}X_B + K_B(X_A)\right)^2. \end{aligned}$$

Ignoring the centering parameter $K_B(X_A)$ for simplicity, we have,

$$\begin{aligned} E[Y] &= \beta_{A1}X'_A + \beta_{A2}{X'_A}^2 + \beta_{B1}s_B(X_a)^{-1}X_B + \beta_{B2}s_B(X_A)^{-2}X_B^2 \\ &= f_1\left(X'_A\right) + \beta_{B1}s_B(X_A)^{-1}X_B + \beta_{B2}s_B(X_A)^{-2}X_B^2, \end{aligned}$$

since terms involving X_B are of interest. The linear model interpretation then follows: a 1-unit increase in X_B leads to a $\beta_{B1}s_B(X_A)^{-1}$-unit increase in Y, and so on. This is in contrast to the usual linear model $E[Y] = \beta_1X_1 + \beta_2X_2$, in which a 1-unit increase in X_2 leads to a β_2-unit increase in Y, *independent* of X_1.

The property that Y increases with X_B dependent on the value of X_A in the proposed model is referred to as interaction elimination. Interaction elimination means we remove the dependence of Y on the joint effect of X_A and X_B by the transformation to X'_A and X'_B, and was proposed by Taguchi (Taguchi, 1987). If the joint effect is exactly due to the linear relationship between X_A and X_B described, then modeling on X'_A and X'_B need only consider the main effects in these transformed design variables. However, as described in (Hamada and Wu, 1995), the assumption of interaction elimination when modeling with the transformed variables may not hold, in which case one still need consider how Y depends on the joint effect of X'_A and X'_B.

We see now that the approach described in (Hamada and Wu, 1995) suffers from some limitations in handling dependent design variables, namely:

- design factors are limited to a linear dependence structure
- the proposed model is not a predictive model.

1.1.2 Geophysics Example

Another real world example of a non-rectangular design region can be easily found when modeling over natural waterways which tend to form complex shapes. Figure 1.4 shows Florida (green) which is bounded by the Gulf of Mexico to the west and the Atlantic ocean to the east. The waters surrounding Florida naturally form a non-rectangular region, and the plot displays surface water temperature. The temperature readings are of a fine resolution taken by satellite, with red indicating cooler water temperatures and yellow warmer water temperatures.

Florida is one of the United States' most populated states, and a majority (80%) of residents live in coastal regions (Merz, 2001). As the area is well known for its tropical storms, predicting the behaviour in this waterway is important for the safety of the residents and the protection of property. In this regard, the Coastal Ocean Monitoring and Prediction System (COMPS) (Merz, 2001; Weisberg et al., 2002) was created with the task of monitoring this waterway. COMPS collects data such as surface water temperature, ocean current and sea level, among others, with a network of buoys. With this data, COMPS aims to predict various critical behaviours, such as storm surges and flooding or industrial accidents such as oil spills to minimize the environmental impact.

One should recognize that modeling such a complex waterway with many dependent variables affecting the state of any particular response (such as the water temperature shown) is very difficult. Treating the Florida region as a rectangular design region clearly does not make any sense, if only for the fact that ground temperature and water temperature are not the same variables, and so shouldn't be treated as one in forming a predictive model for the waterway.

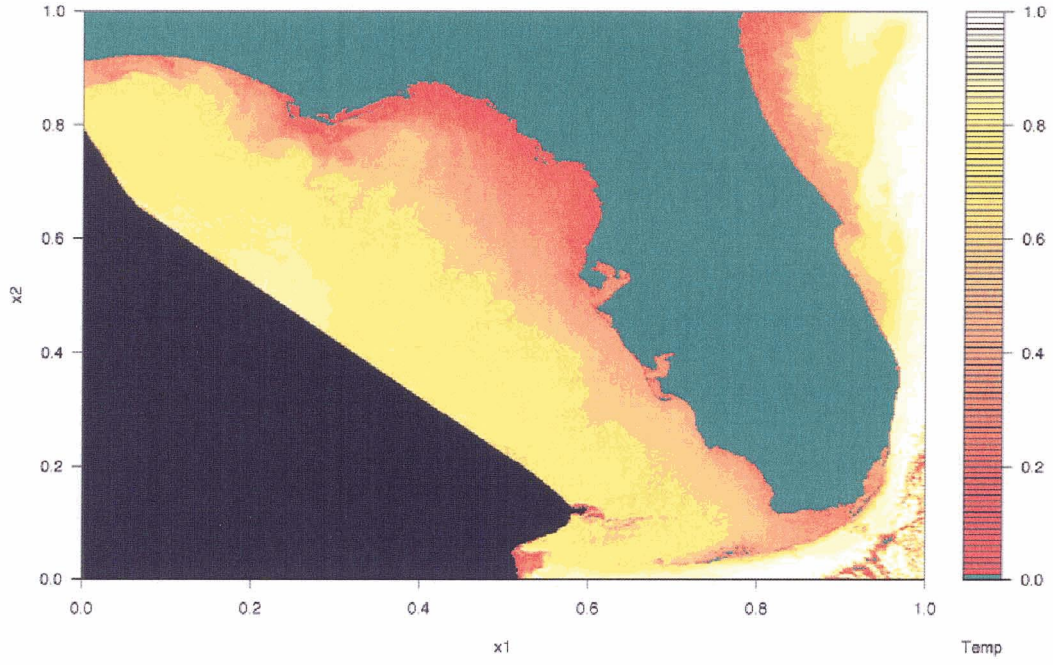


Figure 1.4: Heatmap of Florida water surface temperature

1.2 Geometric Interpretation of Regression

At this point, let us clearly define our non-rectangular design regions of interest. For any set S, we say that S is convex if for any $x_1, x_2 \in S$ and for any constant $\alpha \in [0, 1]$, we have $\alpha x_1 + (1 - \alpha)x_2 \in S$. That is, the set is convex if any point x^* lying on the line segment joining x_1, x_2 is also in S. Otherwise, we say S is non-convex. Then, the region shown in Figure 1.1 is non-convex since it is clear that one can easily select two points in the horseshoe region such that the line segment joining these two points will partially lie outside of the horseshoe region. The same behaviour can be seen if we investigate the Florida region shown in Figure 1.4. The design regions we are investigating in this dissertation are non-convex.

It might be tempting to view the non-convex design region as much ado about nothing. For instance, one might believe it is perfectly acceptable to fit a given

model over a non-convex design region and proceed as normal by only considering the predicted response over the region of interest. However, this is not the case. We consider a more formal argument using the geometric interpretation of regression to illustrate this point.

Recall the definition of a vector space is the set $\mathscr{V}_n = \{v_i = (v_{i1}, \ldots, v_{in}), v_{ij} \in \mathscr{R}, i = 1, \ldots, l\}$, which contains the vector 0 and is closed under addition and scalar multiplication. Then, for instance, R^k is a vector space for any k. Consider our response vector $Y \in \mathscr{R}^n$, design region $X \in \mathscr{R}^p$ and residual vector $\hat{\varepsilon}$. Fitting the usual linear model, $Y = X\beta + \varepsilon$, can be viewed as projecting the vector, Y, from the n-dimensional vector space R^n onto the p-dimensional vector space R^p spanned by the columns of X such that the norm of the residual vector $\hat{\varepsilon}$ is minimized.

Now consider the non-convex subregion of interest, S, such as that shown in Figure 1.1. Suppose this subregion is contained in the vector space $\mathscr{R}^p$, and centered at the origin so that it contains the vector 0. Denote the function defining the subregion as $B(s') = 1$ if $s' \in S$, 0 otherwise. Each point in this subregion is a vector in $\mathscr{R}^p$, hence $S = \{s_i = (s_{i1}, \ldots, s_{ip}), s_{ij} \in \mathscr{R}, B(s_i) = 1, i = 1, \ldots, l\}$. It is clear that the set S is not closed under addition or scalar multiplication, since for bounded S, there exists a constant c and vectors s_1, s_2 such that $cs_1 \notin S$ and/or $cs_1 + s_2 \notin S$. Therefore, the set S does not form a vector space, and in performing regression on the vectors lying in S we are actually regressing onto the subspace $\mathscr{R}^p$ which only contains S.

So, the notion of selecting a non-convex design region and fitting a linear model to this region is inappropriate as we are actually fitting a linear model to the vector space which only contains this region as a subset. Let us refer to this vector space which contains our region only as a subset as the ambient vector space. Since we are actually fitting the model to the ambient vector space, we are completely ignoring the non-convexity of the design region, which is not what we want. Fitting a model to this ambient vector space when the actual response is a function of the non-convex region and not of this vector space (e.g., Figure 1.2) will result in a model that is meaningless with respect to the true response. This is very different from the usual case, where if the region *were* convex, then modeling over this ambient vector space *is* what we want. This is because in the convex case, we fit the model over the ambient vector space but simply restrict the region of space where the model is considered to

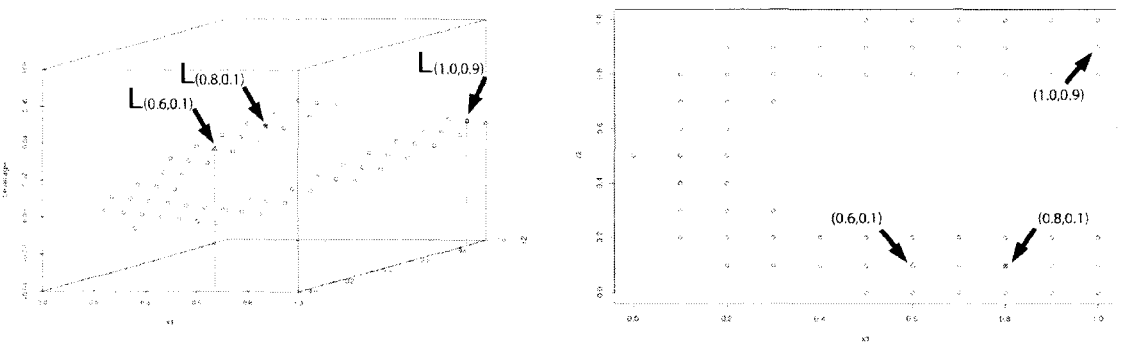
the convex design region of interest.

1.3 Leverages in Regression

Another argument against using the usual regression approach is found when considering leverages. Recall that projecting the vector $Y \in \mathscr{R}^n$ into the p-dimensional design space to obtain a predictive model is accomplished by the so-called hat matrix H. Fitting the linear model yields the estimated regression coefficients $\hat{\beta} = \left(X^T X\right)^{-1} Y$ from which we construct $\hat{Y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T Y = HY$. So, the projection is given by the matrix $H = X\left(X^T X\right)^{-1} X^T$, and we have the further interpretation that each predicted point is the weighted sum of neighbouring points since $\hat{Y}_i = \sum_{j=1}^{N} h_{ij} Y_j$. The weights $\{h_{ij}\}$ are commonly referred to as leverages, and indicate the contribution a point Y_j has on the prediction of Y_i irrespective of the actual response function.

A scatterplot of the leverages for the point $Y_{(0.8,0.1)}$ (denoted by a square) in the horseshoe region is shown in Figure 1.3(a). This plot shows that the leverages for this point are positive at the ends of the horseshoe, and small or negative near the top of the horseshoe. For instance, two points with nearly equal leverage for the prediction of $Y_{(0.8,0.1)}$ occur at $(0.6, 0.1)$ and $(1.0, 0.9)$ respectively with leverages $L_{(0.6,0.1)}$ and $L_{(1.0,0.9)}$. However, while the point $(1.0, 0.9)$ may seem relatively near to $(0.8, 0.1)$ in terms of the straight-line (Euclidean) distance between them (see Figure 1.3(b)), we know from the response function shown in Figure 1.2 that the behaviour of the response at these two points is very different, and it would seem unlikely that positive leverage should be given to $Y_{(1.0,0.9)}$ when predicting $Y_{(0.8,0.1)}$. In fact, the point $(1.0, 0.9)$ could be considered the furthest away from $(0.8, 0.1)$ when viewing the behaviour of the function along the path formed by the horseshoe region in Figure 1.2. In this sense, the horseshoe region is somewhat of a worse-case example since two points which are near in terms of their Euclidean distance are actually far apart.

The importance of distance can be seen by noting the relationship between the leverages and the Mahalanobis distance measure. Recall that the Mahalanobis distance (Ravishanker and Dey, 2002) is essentially a normalized version of Euclidean distance, and can be written as $d_M(x_i) = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}$. In the simplest



(a) (b)

Figure 1.5: (a) Leverages for point $(0.8, 0.1)$ at points $(0.6, 0.1)$ and $(1.0, 0.9)$ for regression over the horseshoe region, (b) the same points in the horseshoe design region

case, one may have $\Sigma \approx I_N$, the NxN identity matrix, in which case the distance measure reduces to usual Euclidean distance. Consider the point x_i in our horseshoe region, and the remaining points $X_{(i)} = X - \{x_i\}$. Let us assume that the existing points $X_{(i)}$ are already mean centered, so that $\bar{X}_{(i)} = 0$. We can estimate the covariance matrix by $\Sigma \approx S = \frac{1}{N-2} X_{(i)}^T X_{(i)}$, where here S is the sample covariance. Then we have $d_M^2(x_i) \approx x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i = h_{ii}$, the leverage for point x_i. This result means that the leverage of a new point x_i is strongly related to the corresponding distance to the centroid. So, the notion that the leverages found do not make sense when fitting a linear model to the horseshoe region can be interpreted as a question of the validity of the distance measure used.

So far, we have seen in the Welding example an approach that did not allow for a predictive model and placed constraints on the way dependent design variables could be related. In our second example, we have further seen a problem where the relationship between dependent design variables may be very complex, and certainly non-linear. However, there are recent approaches that have been used to handle these types of problems. In the next chapter, we present these approaches and consider the related design of experiments problem.

Chapter 2

Methods

In this chapter, we start off by outlining the usual Gaussian Process model over a rectangular design region and the corresponding Integrated Mean Squared Error (IMSE) optimal design criterion. These are then adapted to non-convex design regions. We develop methodology for the GP model in the case of non-convex design regions, and present a new form of the IMSE design criterion for this case. Finally, the problem of actually constructing designs over non-convex regions is addressed.

2.1 Gaussian Process Model

Suppose the response of interest is modeled over the D-dimensional rectangular design region $\Omega \in \mathscr{R}^D$ formed by the independent factors $F = (f_1, \ldots, f_D)$. If $\mathbf{x} = (x_1, \ldots, x_D) \in \Omega$ is any vector in our region of interest, then we are interested in modeling the response as a function of $\mathbf{x}$. For the Gaussian Process (GP) model, we model the response at a particular value of $\mathbf{x}$ (denoted $Y(\mathbf{x})$) as a random function with $Y(\mathbf{x}) = \mu + Z(\mathbf{x}) + \varepsilon(\mathbf{x})$, where μ is the overall mean response, $Z(\mathbf{x})$ is a random function that describes the systematic departure from the mean, and $\varepsilon(\mathbf{x})$ represents measurement error. The random function $Z(\mathbf{x})$ is taken to be normally distributed with constant variance σ_z^2 and unknown correlation $Cor(Z(\mathbf{x}_i), Z(\mathbf{x}_j))$. The measurement error term is also taken to be normally distributed where each $\varepsilon(\mathbf{x})$ is independently distributed with equal variance (σ_ε^2), and are also independent of

$Z(\mathbf{x})$. Since we of course do not know the form of $Cor\,(Z(\mathbf{x}_i), Z(\mathbf{x}_j))$, we make an assumption of its functional form, for instance $Cor\,(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = r(\theta, \mathbf{x}_i, \mathbf{x}_j)$. Then the functional form of the correlation indirectly describes the correlation structure of the random function. So, we model the correlation structure of the response rather than modeling the mean as is done in normal regression.

To utilize the model, the values of the parameters that define the correlation function, as well μ, σ_z^2 and σ_ε^2 must be estimated. This can be accomplished following the usual likelihood framework for finding the parameter values which maximize the likelihood of data observed at N sampled points (i.e., MLE). However, let us first introduce the form of the covariance function we will use in this thesis.

One of the most common correlation functions used in Gaussian Processes is known as the Gaussian correlation function. Suppose we have two points $\mathbf{x}_i$, $\mathbf{x}_j$ in our region Ω, and let $d_{ijk} = |\mathbf{x}_{ik} - \mathbf{x}_{jk}|$ be the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ along the k'th dimension, $k \in 1 \ldots D$. Then the Gaussian correlation between $y\,(\mathbf{x}_i)$ and $y\,(\mathbf{x}_j)$ is written as:

$$r_{ij} = e^{-\sum_{k=1}^{D} \theta_k {d_{ijk}}^2}, \tag{2.1}$$

where $\theta_k \in (0, \infty)$ is a correlation parameter for each dimension. An equivalent and perhaps more easily interpreted expression for the correlation is given by:

$$r_{ij} = \prod_{k=1}^{D} \rho_k^{4{d_{ijk}}^2} \tag{2.2}$$

where now the parameter $\rho_k \in (0, 1)$ has the interpretation that $\rho \approx 0$ implies low correlation while $\rho \approx 1$ implies high correlation. We will make use of (2.1) mainly in presenting formulae in a manner consistent with the literature, while the form (2.2) will be used in discussions due to its interpretability.

The Gaussian correlation function shows that the correlation between two observed values is given by the products of independent correlation functions for each dimension. The correlation along each dimension is given by a parameter which is exponentially decayed according to the squared Euclidean distance between the points in the region Ω along that dimension. Hence, if the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is small, the correlation will be high and so large weight will be given to response $y(\mathbf{x}_j)$

if we were interested in predicting $Y(\mathbf{x}_i)$. If the distance is large, the correlation will be low, and so small weight will be given to response $y(\mathbf{x}_j)$ in predicting $Y(\mathbf{x}_i)$. So, the rate at which the correlation changes is modeled by the parameter, ρ_k (or θ_k), independently in each dimension. The behaviour of the response function can then be thought as being modeled by the correlation structure in the direction of each dimension forming the region Ω.

Some further clarification can be found by examining (2.2) if we momentarily ignore the possibility of measurement error (i.e., $\sigma_\varepsilon^2 = 0$) and consider $D = 1$ and $\Omega = [0, 1]$. Suppose two response values $y(\mathbf{x}_1)$ and $y(\mathbf{x}_2)$ are taken such that $d_{12} = \frac{1}{2}$. Then $r_{12} = \rho^{4{d_{12}}^2} = \rho^{4\left(\frac{1}{2}\right)^2} = \rho$, which means that in this example, a distance of $\frac{1}{2}$ is the distance that gives a correlation of exactly ρ between the corresponding responses. If instead we consider $d_{12} = 1$, then $r_{12} = \rho^{4(1)^2} = \rho^4$, and since this is the maximum distance allowed between two points in this example, it follows that ρ^4 is the minimum correlation between observations. Finally, if $d_{12} = 0$, then $r_{12} = 1$, and our two responses are equal, as we would expect when there is no measurement error.

The overall behaviour of the correlation function is shown in Figure 2.1 for $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. Suppose we wanted to predict the value of the response at a point $\mathbf{x}$ and are interested in how $y(\mathbf{x}_1)$ will affect the predicted response $\hat{Y}(\mathbf{x})$. If the distance between these points is small (say 0.1), then large weight will be given to the prediction $\hat{Y}(\mathbf{x})$ for both $\rho = 0.1$ or $\rho = 0.9$. If instead the distance between these points were large (say 1.0) then large weight will be given to the prediction $\hat{Y}(\mathbf{x})$ for $\rho = 0.9$ as the correlation function decays very slowly in this case, but if $\rho = 0.1$ then almost zero weight will be given to the prediction $\hat{Y}(\mathbf{x})$.

Let us now proceed to show how we can fit the model to data in order to estimate the GP parameters. Suppose we have a design of N points given by $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, where each $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in \Omega$. The corresponding responses are given by $y^T = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N))$, where $y(\mathbf{x}_i) = \mu + Z(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i)$ are realizations of the random function $Y(\mathbf{x})$. Then $Z(X) \sim N(0, \Sigma)$ and $\varepsilon(X) \sim N(0, \sigma_\varepsilon^2 I)$. Let $\mathbf{1}$ be the $N \times 1$ vector of one's, $\Sigma = \sigma_z^2 R$, $\tilde{\Sigma} = \sigma_z^2 R + \sigma_\varepsilon^2 I$, and $\check{\Sigma} = \left(R + \frac{\sigma_\varepsilon^2}{\sigma_z^2} I\right)$, where $R = [r_{ij}]$ is as defined in (2.1). The log-likelihood for the GP model is then given as:

$$l = -\frac{1}{2} log \left|\tilde{\Sigma}\right| - \frac{1}{2} (y - \mathbf{1}\hat{\mu})^T \tilde{\Sigma}^{-1} (y - \mathbf{1}\hat{\mu}). \tag{2.3}$$

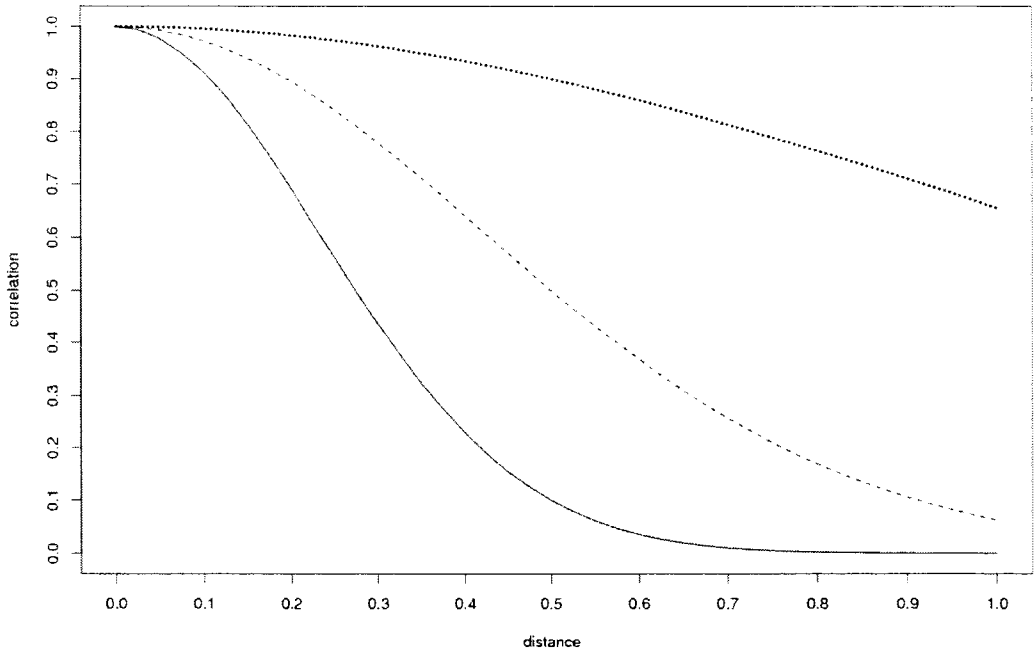


Figure 2.1: Correlation between two responses $y(\mathbf{x}_1)$, $y(\mathbf{x}_2)$ as a function of distance for $\rho = 0.1$ (solid), $\rho = 0.5$ (dashed) and $\rho = 0.9$ (dotted).

The estimate of the mean can be found as:

$$\hat{\mu} = \frac{\mathbf{1}^T \tilde{\Sigma}^{-1} y}{\mathbf{1}^T \tilde{\Sigma}^{-1} \mathbf{1}},$$

and substituting our estimate for the mean back into the log likelihood, the remaining parameters $\hat{\sigma_z}^2$, $\hat{\sigma_\varepsilon}^2$ and the vector $\hat{\theta}$ (or $\hat{\rho}$) can be found numerically (Wolfinger et al., 1994).

Since the GP model predicts the response as a weighted function of the responses observed at the design points, the usual predictor considered is of the linear form $a^T(\mathbf{x})y$ (Sacks et al., 1989a) where $a^T(\mathbf{x})$ are the weights given to the y's for predicting the response at a new location $\mathbf{x}$. The best linear unbiased predictor (BLUP) of this form at the new location $\mathbf{x}$ can be shown to be $\hat{Y}(\mathbf{x}) = r^T(\mathbf{x})\tilde{\Sigma}^{-1}(y - \mathbf{1}\hat{\mu}) + \hat{\mu}$ (see Appendix B), where $r^T(\mathbf{x}) = (r(\mathbf{x}, \mathbf{x}_1), \ldots, r(\mathbf{x}, \mathbf{x}_N))$ is the vector of correlations between the new point $\mathbf{x}$ and all the design points $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The Mean Squared

Error (MSE) of this predictor is:

$$E\left[\left(Y(\mathbf{x})-\hat{Y}(\mathbf{x})\right)^2\right] = \sigma_z^2 - (\sigma_z^2)^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}r(\mathbf{x}) + \left(1-\sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\right)^2\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1} + \sigma_\varepsilon^2$$

(see Appendix B). A more convenient form for our purpose is:

$$E\left[\left(Y(\mathbf{x})-\hat{Y}(\mathbf{x})\right)^2\right] = \sigma_z^2 - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \tilde{\Sigma} \end{pmatrix}^{-1}\begin{pmatrix} 1 & r^T(\mathbf{x}) \\ r(\mathbf{x}) & r(\mathbf{x})r^T(\mathbf{x}) \end{pmatrix}\right],$$

since terms involving $\mathbf{x}$ are separable from the remaining terms. We will show later that this is necessary in formulating our design criterion.

2.1.1 Design for Gaussian Processes

There are numerous approaches and criteria to consider when constructing optimal designs. For instance, a common design-based approach are the"space-filling" designs, such as (Johnson et al., 1990). Model-based designs are typically found by selecting design points that maximize or minimize an appropriate optimality criterion. Our interest lies in prediction, so we will consider the I-Optimal criterion, or *Integrated Mean Squared Error* (IMSE) optimal designs. This criterion minimizes the average squared prediction error by minimizing the integral of the MSE over the design region Ω. This makes sense when we are interested in prediction, as one would like to minimize the average error when predicting the response at any point $\mathbf{x} \in \Omega$.

In the usual case of independent design variables, the selection of I-Optimal design points can be stated formally as:

$$\underset{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N}{\arg\min} \int_\Omega E\left[\left(Y(\mathbf{x})-\hat{Y}(\mathbf{x})\right)^2\right] d\mathbf{x},$$

where $\mathbf{x}_i \in X \subset \Omega$. It is customary in the literature (Sacks et al., 1989a; Sacks et al., 1989b) to in fact minimize the IMSE normalized by σ_z^2. Minimizing the IMSE instead of only the MSE makes sense since integrating the MSE over the region takes into account the correlation structure of the GP model as including a point $\mathbf{x}$ in the

design not only reduces the error at that point, but also reduces the error at nearby neighbouring points. So, the design problem becomes:

$$\underset{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N}{\arg\min}\ J\left(\theta,\sigma_z^2,\sigma_\varepsilon^2\right),$$

where $J\left(\theta,\sigma_z^2,\sigma_\varepsilon^2\right) = \frac{1}{\sigma_z^2}\int_\Omega E\left[\left(Y(\mathbf{x})-\hat{Y}(\mathbf{x})\right)^2\right]d\mathbf{x}$. Let us write $\sigma_\tau^2 = \frac{\sigma_\varepsilon^2}{\sigma_z^2}$. Then, with some algebra (see Appendix C), we can write the σ_z^2-normalized form of the IMSE as:

$$J\left(\theta,\sigma_\tau^2\right) = \left(1+\sigma_\tau^2\right) - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma} \end{pmatrix}^{-1} \int_\Omega \begin{pmatrix} 1 & r(\mathbf{x})^T \\ r(\mathbf{x}) & r(\mathbf{x})r(\mathbf{x})^T \end{pmatrix} d\mathbf{x}\right]. \quad (2.4)$$

This form is different from that presented in the literature, since for instance (Sacks et al., 1989a; Sacks et al., 1989b) assume $\sigma_\varepsilon^2 = 0$ which simplifies the design problem. This simplification is useful in computer experiments where there is no measurement error in the response. Here we consider the effects of both σ_z^2 and σ_ε^2 by considering the ratio of these variabilities in formulating J, which just so happens to be a generalization of the σ_z^2-normalized form given by (Sacks et al., 1989a; Sacks et al., 1989b) if we consider $\sigma_\varepsilon^2 \neq 0$. As mentioned earlier, the terms involving $\mathbf{x}$ are separable which eases the integration. Since the design variables are independent, the integration terms can be solved analytically because the correlation function is also separable, hence

$$\begin{aligned}\int_\Omega r_j d\mathbf{x} &= \prod_{k=1}^{D}\int_{a_k}^{b_k} e^{-\theta_k\left(x_k-x_{jk}\right)^2}dx_k \\ &= \prod_{k=1}^{D}\sqrt{\frac{\pi}{\theta_k}}\left[\Phi\left(\sqrt{2\theta_k}(b_k-x_{jk})\right)-\Phi\left(\sqrt{2\theta_k}(a_k-x_{jk})\right)\right],\end{aligned}$$

where $\Phi(.)$ is the standard normal CDF.

Note that (2.4) is also a function of the parameters θ and σ_τ^2. This is problematic, as our design criterion is dependent on exactly the parameters that we would fit via maximum likelihood given a design! The parameter θ is often handled (Sacks et al., 1989a; Sacks et al., 1989b) by performing a robustness study where designs are constructed at a number of assumed values of θ, and then compared using the

relative efficiencies of these designs. Let us follow this approach for our purposes, and define the relative efficiency of two designs in the following manner:

$$R_{eff}\left(X(\theta_T), X(\theta_A)\right) = \frac{J\left(\theta_T, X(\theta_T), \hat{Y}_{\theta_{MLE}}\left(X(\theta_T)\right)\right)}{J\left(\theta_T, X(\theta_A), \hat{Y}_{\theta_{MLE}}\left(X(\theta_A)\right)\right)}, \tag{2.5}$$

where θ_T is the true value of the parameter, $X(\theta_T)$ is the set of optimal design points found under knowledge of θ_T, and $\hat{Y}_{\theta_{MLE}}\left(X(\theta_T)\right)$ is the fitted response under the optimal design for θ_T, while $X(\theta_A)$ is the set of optimal design points found under an assumed $\theta_A \neq \theta_T$ with corresponding fitted response $\hat{Y}_{\theta_{MLE}}\left(X(\theta_A)\right)$.

So, the numerator in (2.5) is the minimum IMSE that can be expected since we have constructed the design under knowledge of the true value of the parameter θ, while the IMSE denoted in the denominator is likely larger since we have assumed an incorrect value of θ in constructing our design. Hence, the relative efficiency lies in the range [0,1], where a value close to 1 indicates that the design constructed under the assumed θ_A gives performance nearly equivalent to a design constructed under knowledge of the true parameter value, while a value close to 0 indicates that the design constructed under θ_A performs very poorly as compared to the optimal design constructed under knowledge of θ_T.

A robust design is a design constructed under an assumed θ_A which gives good performance under a wide range of θ_T's. Stated formally, a robust θ is found as:

$$\theta_R = \arg\max_{\theta_A}\left(\min_{\theta_T} R_{eff}\left(X(\theta_T), X(\theta_A)\right)\right). \tag{2.6}$$

Selection of the robust θ_R can be done by performing a simulation study using designs constructed with various θ_A's and θ_T's and then selecting the particular θ_A that maximizes the minimum relative efficiency observed under assumption of θ_A. This is exactly what will be done in later sections. However, at this point we need to also consider the parameter σ_τ^2.

Since the IMSE is computed normalized to σ_z^2, the parameter σ_τ^2 can be thought of as the ratio of large-scale (ie functional) variation versus small-scale (ie noise) variation. If we assume that the total variation of the response vector y is scaled so that $\sigma_y^2 = 1$, and we expect most variability to be due to the functional pattern

present in the y's, a reasonable assumption is to expect the value of σ_z^2 to be close to 1 and σ_ε^2 close to 0. Accordingly, we can take a Model Averaging approach and place inverse gamma priors centered at 0.9 for σ_z^2 and 0.1 for σ_ε^2 as motivated in (Chipman, 1997; Linkletter et al., 2005) in order to model σ_τ^2. We then average our design criterion over this distribution, ie:

$$J(\theta) = E_{\sigma_\tau^2}\left\{\left(1+\sigma_\tau^2\right) - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma} \end{pmatrix}^{-1} \int_\Omega \begin{pmatrix} 1 & r(\mathbf{x})^T \\ r(\mathbf{x}) & r(\mathbf{x})r(\mathbf{x})^T \end{pmatrix} d\mathbf{x}\right]\right\}, \tag{2.7}$$

and use the form (2.7) in performing the robustness study to select an appropriate θ_R that is optimal in the sense of (2.6). We propose this new approach to extend the construction of I-Optimal designs to the general case of $\sigma_\varepsilon^2 \neq 0$.

2.2 Gaussian Processes and Non-Convex Domains

The GP model can be viewed as performing local fits to form a global estimate of the response. This is very different from our usual linear model, which performs a global fit to the data. The global approach has a notable limitation in the case of dependent design variables as we saw earlier, while the local approach will enable greater flexibility in considering dependent design variables. This is particularly true in practice where we may not know the exact relationship (in functional form) between the design variables.

In order to fit a GP model, the pairwise distances between design points are required. Typically the distance measure used is simply Euclidean distance, however since the model only depends on pairwise distances, one can use alternative distance metrics (subject to some constraints). It is this ability to change the distance metric used in defining our process that will allow easier consideration of dependent design variables.

Using the Euclidean distances between points ignores the relationship that exists between the design variables in non-convex regions. Recently, researchers in geostatistics have utilized the so-called *water* distance, or more generally, *geodesic* distance as

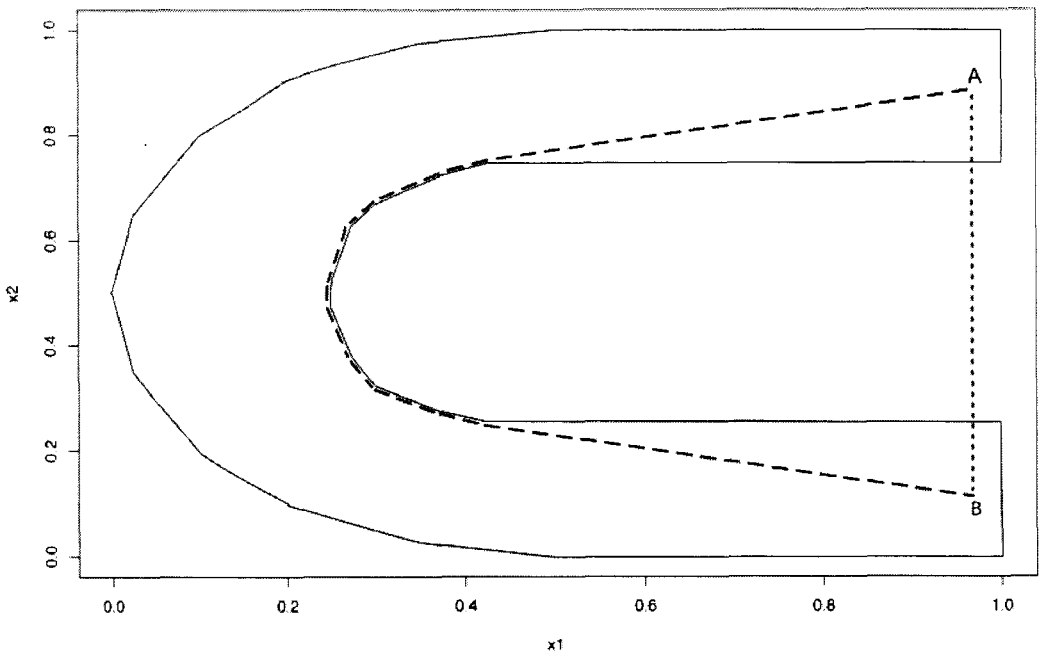


Figure 2.2: Euclidean (dotted) vs. Geodesic (dashed) distances between points A and B in a dependent design variable region

an alternative distance metric (Rathbun, 1998; Loland and Host, 2003). We will refer to the geodesic distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ as $d_g(\mathbf{x}_i, \mathbf{x}_j) = d_{g_{ij}}$. The geodesic distance between two points lying in a connected region of arbitrary shape is simply the length of the shortest path connecting these two points such that the path lies completely in the region of interest. Figure 2.2 shows the geodesic distance between points A and B lying in a design region formed by dependent variables. Recall that such a region is non-convex because the straight line joining two points may not entirely lie within the region of interest, as is the case for the points A and B shown. Formally, suppose now that we consider the new space $\mathscr{G}$ embedded in the ambient Euclidean space $\mathscr{R}^D$ with geodesic distance metric $d_{g_{ij}}$. For instance, the horseshoe region shown in Figure 2.2 is embedded in the ambient Euclidean space $\mathscr{R}^2$. We map our design space from Ω to $\mathscr{G}$ by substituting the geodesic distance $d_{g_{ij}}$ for the Euclidean distance d_{ij}. We make this mapping as a means of handling the non-convexity of Ω.

In the geodesic space, we define the correlation function as

$$r_{ij} = e^{-\theta_0 d_{g_{ij}}^2} = \rho_0^{4 d_{g_{ij}}^2}, \tag{2.8}$$

where we now only have the single parameter θ_0 instead of the vector considered before. This follows since the geodesic distance is essentially dimensionless no matter what the dimension of the ambient Euclidean space. As before, we can fit the model by maximizing the likelihood (2.3) with the appropriate substitutions in the calculation of $\tilde{\Sigma}$. Since the correlation function is now integrated over our non-convex region, we must estimate the integral $\int r d\mathbf{x}$ numerically. Accordingly, we will calculate our design criterion using Monte-Carlo integration (Press et al., 1992):

$$J(\theta) = E_{\sigma_\tau^2}\left\{(1+\sigma_\tau^2) - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma} \end{pmatrix}^{-1} \frac{1}{N_G}\sum_{G\in\mathscr{G}} \begin{pmatrix} 1 & r(\mathbf{x})^T \\ r(\mathbf{x}) & r(\mathbf{x})r(\mathbf{x})^T \end{pmatrix}\right]\right\}, \tag{2.9}$$

where $G = \{G_1, \ldots, G_{N_G}\}$ is a sample of N_G points from $\mathscr{G}$.

Modeling in $\mathscr{G}$-space has been discussed previously (Rathbun, 1998; Loland and Host, 2003), however design for non-convex regions, to the best of our knowledge, is new. Yet despite the simplicity suggested by (2.9), there are numerous difficulties in actually being able to construct designs in non-convex regions. The most significant problem lies with the very assumption of using the geodesic metric itself, since modeling a GP using the geodesic metric can lead to a correlation matrix R that is not positive semi-definite (Rathbun, 1998). We now outline a novel approach that allows us to resolve this issue.

2.2.1 ISOMAP

In (Tenenbaum et al., 2000), the authors propose an approach to project an intrinsically low-dimensional surface located in a high-dimensional ambient space into a low-dimensional embedding space in order to recover the underlying geometry of the surface. We will see in a moment how this method, known as ISOMAP, will be useful to us. First, let us understand this method and the problem it was originally designed to solve. The basic steps are:

- fill the surface lying in high-dimensional space with a grid of points
- construct a neighbourhood graph
- compute the shortest paths between points
- construct a low dimensional embedding of the higher dimensional surface.

The ISOMAP algorithm begins by filling the region $\mathscr{G}$ with a grid of points and then constructing a graph over these points by connecting the ϵ-nearest points to one another. These nearest neighbours are connected according to their Euclidean distance in the ambient space $\mathscr{R}^D$. Based on this graph, the geodesic distances between the grid of points are approximated by computing the shortest paths along the graph defined by the grid of points. The shortest paths are found by solving the all-pairs shortest path problem by applying Floyd's algorithm (Cormen et al., 1990). This is essentially done by adding up the Euclidean distances along short hops between neighbouring points that lie on the path joining two points of interest. Although calculating geodesic distances in this way yield only an approximation dependent on the parameter ϵ, for simplicity we still refer to these calculated distances as the geodesic distance d_g.

Using a grid of 63 points filling the horseshoe region in equally spaced intervals of 0.1 units, the geodesic and Euclidean pairwise distances were calculated. Figure 2.3 shows the scatterplot of pairwise Euclidean distances versus pairwise geodesic distances for the horseshoe region. One can notice that the two distance metrics quickly become unequal for points that are geodesically further apart, as we would expect. Since the correlation structure that we model is directly dependent on these distances, this graph shows that the correlation structure for the horseshoe region will be incorrect if we do not consider the non-convexity of the region.

The pairwise geodesic distances calculated using Floyd's algorithm can be taken as dissimilarites and a new embedding space for these dissimilarites can be found by applying Metric Multidimensional Scaling (Cox and Cox, 2001) to the matrix of geodesic distances computed. In this manner, a Euclidean embedding space can be found such that the distances between points in the embedding space approximate or

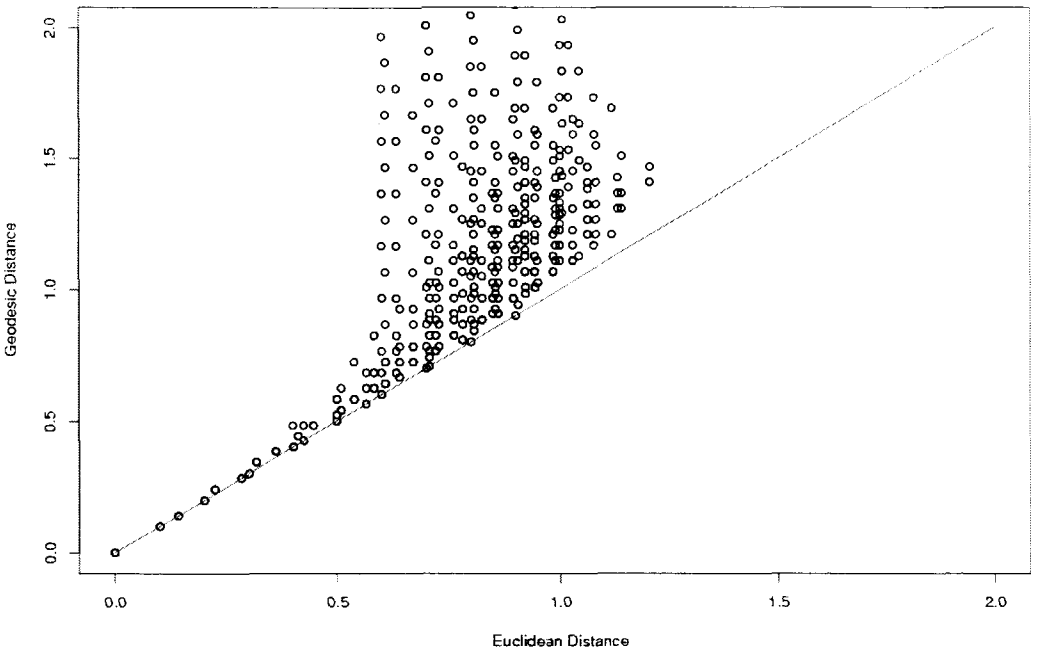
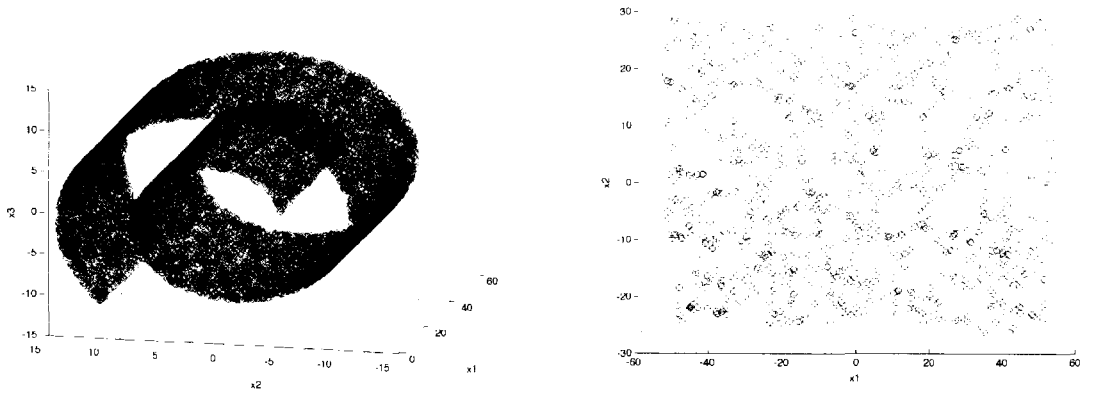


Figure 2.3: Euclidean Distance $d(.,.)$ vs. Geodesic Distance $d_g(.,.)$ for Non-Convex Horseshoe Region

equal the geodesic distances found from the graph constructed over $\mathscr{G}$. Multidimensional Scaling (MDS) works as follows. Suppose we have the matrix of squared pairwise Euclidean distances D^2 with entries $[D^2_{ij}] = d^2_{ij}$. Let $A = -\frac{1}{2}D^2$ and $B = HAH$, where $H = I - \frac{1}{n}1^T1$ is a centering matrix. Applying H to A in this manner is the same as mean-centering our matrix of squared distances. It can then be shown that the resulting B is actually the inner product matrix $B = X^TX$, and by writing it in terms of its spectral decomposition $B = V\Lambda V^T$, then the original coordinates can be recovered as $X = V\Lambda^{\frac{1}{2}}$. Hence MDS provides a way to reconstruct the Euclidean coordinates given only the matrix of Euclidean pairwise distances.

The ISOMAP approach replaces d_{ij} with the geodesic distance calculated over the surface of interest, $d_{g_{ij}}$. If the resulting matrix B is positive semi-definite, then a representation of points in a possibly lower dimensional Euclidean space $\mathscr{E}$ has been found such that the Euclidean distance in $\mathscr{E}$ equals the geodesic distance, as required.



(a) (b)

Figure 2.4: (a) The original swiss roll and (b) ISOMAP's low-dimensional embedding

The classical example discussed in (Tenenbaum et al., 2000) is the swissroll surface shown in Figure 2.4(a). We can see that the swissroll is essentially a 2-dimensional plane that has been curled and placed in a 3-dimensional ambient Euclidean space. Hence, the actual vectors that define the swissroll lie in $\mathscr{R}^3$. In order to recover the true 2-dimensional plane, the authors sample 1000 points from the swissroll, construct an adjacency graph based on this sample and then approximate the geodesic distances between points in the swissroll using the constructed graph. Applying MDS to the geodesic distances then recovers an embedding for the swissroll. The authors note that the eigenvalues of the embedding quickly drop to nearly zero for 2 or more dimensions. Therefore, the appropriate dimensionality of the embedding is 2-dimensions, and the original plane, which we know *is* 2-dimensional, has been recovered as shown in Figure 2.4(b).

Unfortunately, not every non-convex region can be easily embedded. It is quite common for the non-convex regions we are interested in to form a matrix B that is not positive semi-definite. In this case, we can view the reconstruction $\mathbf{x}_\mathscr{E} = V_1\Lambda_1^{\frac{1}{2}}$ and the resulting distance $d_\mathscr{E}$ as a further approximation of the geodesic distance (since d_g is itself the approximate geodesic distance) by using some of the positive (eigenvector,eigenvalue) pairs from the spectral decomposition of B, denoted by (V_1, Λ_1) respectively. Using this approximation, the resulting B_1 will be positive semi-definite,

hence giving us the approximating Euclidean space embedding needed.

Although the authors of ISOMAP were clearly interested in *dimensionality reduction* to uncover simplified forms of surfaces located in a high dimensional space, our insight stems from viewing the ISOMAP approach in the opposite light of projecting a surface into a potentially *higher dimensional* embedding space. We will provide an intuitive argument for this reasoning. First, it should be recognized that if our surface of interest is convex and in an ambient Euclidean space, then the geodesic distance measured between points on this surface is simply the Euclidean distance between these points. In this case we have already shown that the exact low-dimensional embedding can be found. Since the embedding will be exact, the distances in the embedding space will equal the geodesic distances measured in the original space, in which case we say that the embedding is isometric. If instead we have a problem such as the swissroll example shown and an isometric embedding is found, then it must be a lower dimensional embedding by assumption. However, there is a third case that can occur, which is motivated by the horseshoe region shown earlier. In this instance, we have a 2-dimensional surface lying in a 2-dimensional ambient space. Since the inner product matrix B in this case is not positive semi-definite, then the geodesic distances constructed over the horseshoe are not isometric to a Euclidean space. Therefore, we know that no appropriate lower-dimensional Euclidean embedding space exists for the horseshoe. However, we can still embedded our region in a Euclidean embedding space of possibly equal or higher dimensionality than the ambient space. The question is then of selecting the appropriate dimension.

There are two obvious approaches (Cox and Cox, 2001) to selecting the correct embedding dimension when using MDS. In the classical application of MDS where one is usually interested in dimensionality reduction, one would choose the dimension of the embedding space that captures the greatest proportion of variability in the embedded points. This is known as the maximum variability criterion, where the variability of the embedded points along each dimension of the embedding space is given by the eigenvalues from the matrix B:

$$\arg\max_{k} \sum_{i=1}^{k} \lambda_i, \tag{2.10}$$

where $k = 1 \ldots p$.

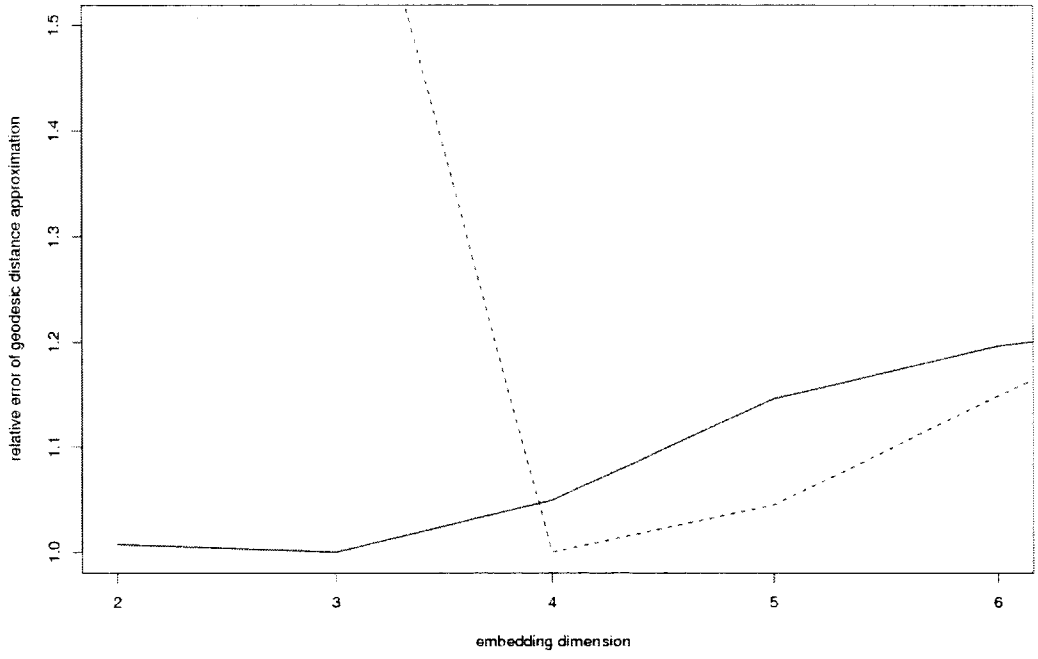


Figure 2.5: Relative Empirical Mean Squared Error of geodesic distance approximation ($EMSE_{\mathscr{E}}$) for 0.10 spaced grid (solid) and 0.05 spaced grid (dashed) in the horseshoe region

Another approach is the mean squared error criterion:

$$EMSE_{\mathscr{E}}(dim(\mathscr{E}) = k) = \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} (d_g(\mathbf{x}_i, \mathbf{x}_j) - d_{\mathscr{E},k}(\mathbf{x}_i, \mathbf{x}_j))^2 \tag{2.11}$$

In fact the criterion of maximizing the variability of the embedding (2.10) is the same as minimizing the reconstruction error of the embedding (2.11) in the Euclidean distance case. Consider the case of Euclidean distances calculated from n p-dimensional points X, and the resulting inner product matrix $B = XX^T$ where the rank $r(B) = r(XX^T) = r(\mathbf{x}) = p$. Since B is positive semi-definite with rank p and hence has p positive eigenvalues and $n - p$ zero eigenvalues, the corresponding spectral decomposition can be written as

$$B = V\Lambda V^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T = \sum_{i=1}^{p} \lambda_i v_i v_i^T = V_1 \Lambda_1 V_1^T,$$

since $(\lambda_{p+1}, \ldots, \lambda_n) = 0$. The criterion of maximum variation would lead to one taking all p (eigenvector,eigenvalue) pairs in the reconstruction, while the reconstruction error criterion would also require one to take all p (eigenvector,eigenvalue) pairs since this reconstructs the original points exactly, hence giving the minimum (zero) error.

In contrast, for the geodesic distance case of interest, we typically do not have a positive semi-definite matrix B, and hence an exact distance preserving embedding in p-dimensional Euclidean space cannot be found. In this case, the two criteria may not agree exactly. However, applying the reconstruction error criterion would seem more appropriate since our model of interest directly depends on the accuracy of our distance measure. Applying this criterion will approximate the geodesic distances with the greatest accuracy by finding a Euclidean space $\mathscr{E}$ that is as close to being isometric to $\mathscr{G}$ as possible.

We can compute the empirical mean squared error of the projection for varying values of $k = 1..K$, with K being the number of positive eigenvalues of our inner product matrix B. The value of k that minimizes $EMSE_{\mathscr{E}}$ is then used in constructing the projection of our points into the embedding space $\mathscr{E}$. Figure 2.5 shows how the values of $EMSE_{\mathscr{E}}$ change for varying k and grid densities. This result suggests an embedding space of 3 or 4 dimensions is appropriate for the horseshoe region.

In terms of modeling and design for the Gaussian Process, if the Euclidean distances in the embedding space equal the geodesic distances, then the embedding space $\mathscr{E}$ is isometric to the original space $\mathscr{G}$ and use of the raw geodesics in forming the correlation matrix R is valid. Otherwise, we use the approximation which will also lead to a valid correlation matrix. We can investigate the error introduced by this approximation by looking at the difference between d_g and $d_{\mathscr{E}}$ in more detail.

It is natural to ask how many points need to be taken in our region of interest to form good estimates of the geodesic distance. Unfortunately, the calculation of geodesic distances is limited by computational complexity. Solving the all-pairs shortest path problem for n_g points using Floyd's algorithm requires $O(n_g^3)$ computations, and is therefore not amenable to extremely large n_g. So in practice, one would estimate d_g and hence $d_{\mathscr{E}}$ using a reasonably small subset of points. We can nonetheless get an idea of how sensitive this procedure is by comparing the true geodesic

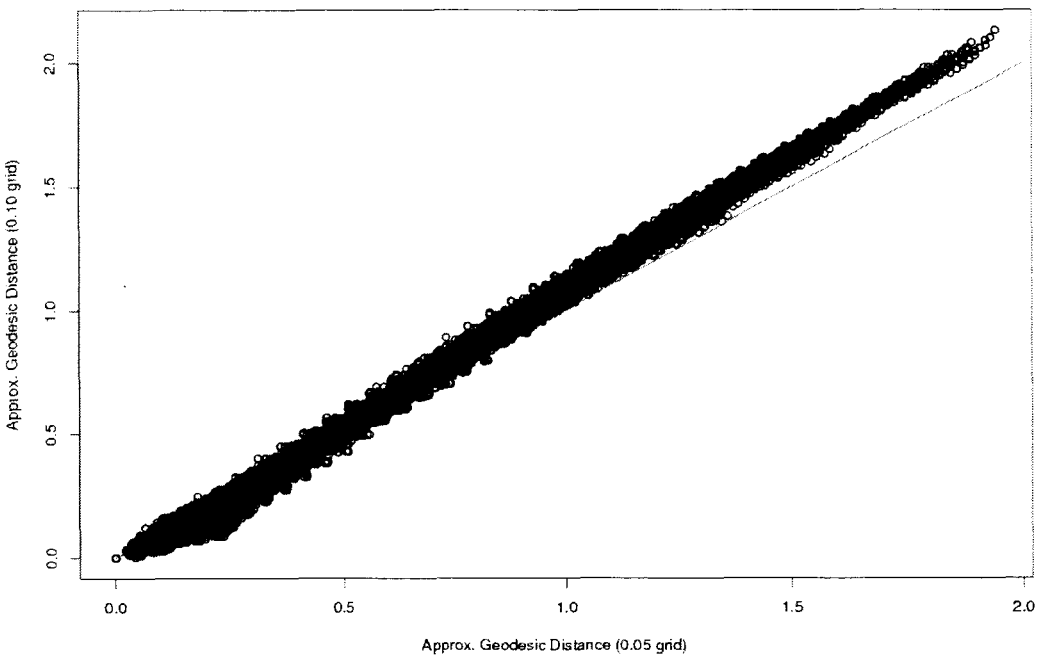


Figure 2.6: Approximated Geodesic Distance $d_{\mathscr{E}}(.,.)$ using 0.05 grid vs. Approximated Geodesic Distance $d_{\mathscr{E}}(.,.)$ using 0.10 grid for the Non-Convex Horseshoe Region

distances and those computed with a grid density 0.10 and 0.05.

The scatterplot in Figure 2.6 was generated by computing $d_{\mathscr{E}}^{(0.05)}(.,.)$ for 247 points using a 0.05 grid spacing and comparing these distances to those computed using the distances $d_{\mathscr{E}}^{(0.10)}(.,.)$ found using 63 points with 0.10 grid spacing. In order to compare the approximated geodesic distances using these two embeddings based on 0.05 and 0.10 grid spacing, we need to further approximate 247-63=184 distances for the 0.10 grid spacing case, which we do using the method described in the next section. For now though, it is sufficient to see that the distances $d_{\mathscr{E}}^{(0.10)}(.,.)$ and $d_{\mathscr{E}}^{(0.05)}(.,.)$ roughly coincide, although there is a noticeable overestimation of large distances when using the 0.10 grid spacing. This is somewhat expected since points which are nearby in $\mathscr{G}$ have distances which are nearly Euclidean and hence will not change greatly with increasing point density. However, as we look at points further apart, their distances

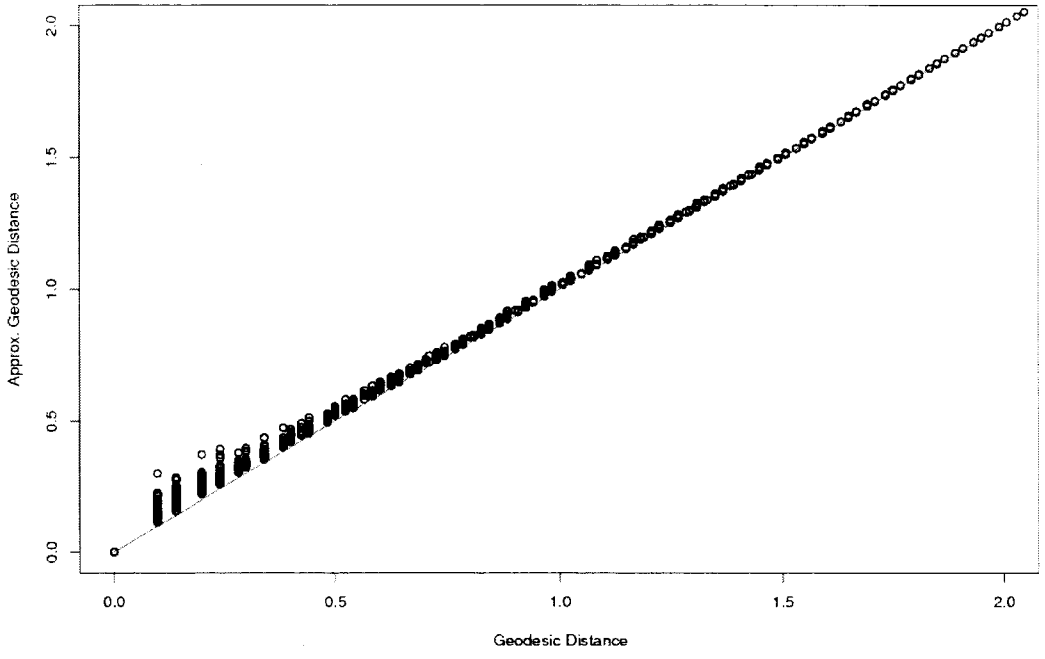


Figure 2.7: Geodesic Distance $d_g(.,.)$ vs. Approximated Geodesic Distance $d_\epsilon(.,.)$ for Non-Convex Horseshoe Region

are reflected by the shortest path length connecting them, which we approximate as the sum of piecewise line segments on the approximate shortest path given the resolution of our grid. It follows that the longer the path length, the greater the number of perturbations away from the minimum path we will have and hence the decreased accuracy for distant points. Increasing the grid density allows the true path to be better approximated, thereby reducing this error.

The scatterplot shown in Figure 2.7 plots the true geodesic distances calculated via Floyd's algorithm to the approximation found with ISOMAP. We see that the large distances are identical, but the approximation has some error for smaller distances, which tend to be overestimated.

From this discussion of ISOMAP, we can note that

- We can project a low-dimensional non-convex region into a new Euclidean space of typically higher dimension by conserving *approximate* distance measure equality
- Although the appropriate MDS criterion for finding this embedding is the MSE of the approximated distance, it is clear that there is some bias in the approximation
- The bias in distance approximation will decrease at the expense of denser sampling and higher computational cost

The computational cost can be a significant barrier to the calculation of geodesic distances. For instance, the ISOMAP algorithm implemented in (Tenenbaum et al., 2000) comfortably handles $1,000$ points in estimating the geodesic distances for the swissroll example. However, in a much higher dimensional space, the computational problem would soon become the limiting factor. Nonetheless, this approach allows modeling the GP over a non-convex design region without any problems by transforming the region to an approximating Euclidean space. The model should be markedly better than ignoring the non-convexity of the design region by modeling over the original Euclidean space.

2.2.2 Out of Sample Extension

Given that we are computationally limited to n_g points in calculating our geodesic distances $D_g = [d_g(.,.)]$, and the resulting approximation of these distances $D_{\mathscr{E}} = [d_{\mathscr{E}}(.,.)]$ found with the p-dimensional embedding of points X, we may be satisfied with the approximation found as measured by $EMSE_{\mathscr{E}}$. However, in subsequently applying our design criterion, we will need to estimate the integral (2.9) using Monte Carlo integration as the sum of the MSE over N_G points sampled in our region of interest. In typical cases, we will have $n_g < 200$, while it will almost always be the case that we will require $N_G >> 200$ to approximate the integral. So, using only $N_g = n_g$ points in estimating (2.9) will be insufficient to get a good estimate of the IMSE.

Additionally, calculating the IMSE requires finding $r(\mathbf{x}), \forall \mathbf{x} \in \{G\}$, where $r(\mathbf{x})$ is the vector of correlations between a point $\mathbf{x}$ and all the points in the design $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. This requires knowledge of the geodesic distances $d_g(\mathbf{x}, \mathbf{x}_1), \ldots, d_g(\mathbf{x}, \mathbf{x}_N)$, which are unknown. A natural approximation one might make is:

$$d_g(\mathbf{x}, \mathbf{x}_j) \approx \min_{i=1}^{n_g} d_g(\mathbf{x}, \mathbf{x}_i) + d_{\mathscr{E}}(\mathbf{x}_i, \mathbf{x}_j),$$

which is nothing but the geodesic distance between $\mathbf{x}$ and $\mathbf{x}_i$ plus the approximated geodesic distance between $e(\mathbf{x}_i)$ and $e(\mathbf{x}_j)$ calculated in the embedding space. However, it was observed that using this approximation can result in negative MSE values. It also does not make sense to calculate $r(\mathbf{x})$ in this way as the distances would not be consistent with those used in calculating R since for that purpose we use the embedding space approximate distances.

Both of these issues can be resolved by projecting the additional $(N_g - n_g)$ points lying in $\mathscr{G}$ – *space* into the embedding space $\mathscr{E}$ in a manner *consistent with the projection already found* using the sample of n_g points. The ability to project additional out-of-sample points in this manner is shown in (Bengio et al., 2003) as:

$$e_k(\mathbf{x}) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n_g} v_{ik} \tilde{d}_g(\mathbf{x}, \mathbf{x}_i) \tag{2.12}$$

Therefore, we can deal with out of sample points by projecting them into the embedding space in a manner consistent with the original embedding found using n_g points.

2.3 Constructing Designs on Non-Convex Regions

We now look at how I-Optimal designs can actually be constructed. The basic steps one need take are:

- Construct ϵ-grid of size n_g over design region Ω of interest and calculate geodesic distances
- Construct family of $d = 1 \dots D$ dimensional embeddings $\mathscr{E}^d$
- Apply (2.11) to find the best distance preserving embedding $\mathscr{E}^k$
- Project N_G points from $G \in \mathscr{G}$ to the embedding space $\mathscr{E}^k$ using (2.12)
- Draw $N_{\sigma_\tau^2}$ values from the distribution for σ_τ^2
- Choose the assumed value(s) of ρ (or θ) and the number of desired design points n_X
- Search for I-Optimal design

At this point we have shown how to calculate the geodesic distances and select a distance preserving embedding as well as using the out of sample formula for embedding additional points for the purpose of estimating the integral over the design region. We then can construct our designs using the IMSE criterion by performing the Monte Carlo integral over the design region and taking the expectation with respect to a specified distribution for σ_τ^2. However, before making use of the criterion in actually constructing designs, we must make the important distinction of which representation, in terms of space, of the design region are we taking the integral with respect to. That is, since we now know that the distances calculated in $\mathscr{G}$ are approximated by the usual Euclidean distances in $\mathscr{E}^k$, we must consider which space we will use in constructing our designs.

2.3.1 Model 1: Distance Approximation Approach

If we simply consider the embedding space $\mathscr{E}$ as providing a valid distance approximation for $\mathscr{G}$ by making the simple substitution $d_{\mathscr{G}}(.,.) \approx d_{\mathscr{E}}(.,.)$, then our criterion becomes:

$$J_{\mathscr{G}}(\theta) = \left(1 + \bar{\sigma_\tau^2}\right) - trace\left[\frac{1}{N_{\sigma_\tau^2}} \sum_{i=1}^{N_{\sigma_\tau^2}} \begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma}_i \end{pmatrix}^{-1} \frac{1}{N_G} \sum_{G \in \mathscr{G}} \begin{pmatrix} 1 & \hat{r}_{\mathscr{G}}(\mathbf{x})^T \\ \hat{r}_{\mathscr{G}}(\mathbf{x}) & \hat{r}_{\mathscr{G}}(\mathbf{x}) \hat{r}_{\mathscr{G}}(\mathbf{x})^T \end{pmatrix}\right] \tag{2.13}$$

where $G = \{G_1, \ldots, G_{N_G}\}$ is a sample of N_G points from $\mathscr{G}$, $\sigma_{\tau i}^2$ is the i'th draw from the distribution for σ_τ^2 giving

$$\check{\Sigma}_i = R + \sigma_{\tau i}^2 I$$

and

$$r_{\mathscr{G}j}(\mathbf{x}) = e^{-\theta_0 d_{\mathscr{G}}^2(\mathbf{x}, \mathbf{x}_j)} \approx e^{-\theta_0 d_{\mathscr{E}}^2(\mathbf{x}, \mathbf{x}_j)} = \hat{r}_{\mathscr{G}j}(\mathbf{x}).$$

In other words, when we view the problem as finding a k-dimensional Euclidean embedding space $\mathscr{E}^k$ that best approximates the geodesic distances $\mathscr{G}$ then we still model over the space $\mathscr{G}$ only we will have the single parameter θ_0 (or ρ_0) since we only have one modeling dimension in $\mathscr{G}$. This approach has not been considered in the literature, and means that we model the correlation structure as if it decays at the same rate in all directions. This allows the nice property that when we view the problem as a distance approximation, we don't need to make any assumptions on the true dimensionality of the response. However, if instead the true response does lie in a multidimensional space, then modeling the response in this way is not the right thing to do.

2.3.2 Model 2: Embedding Space Approach

An alternative view is to suppose that the distance preserving space $\mathscr{E}^k$ is the true underlying space that the response is a function of, in which case we view the response as being parameterized by a $k - dimensional$ independent variable. In this case, our

criterion will be:

$$J_{\mathscr{E}}(\theta) = (1+\bar{\sigma_\tau^2}) - trace\left[\frac{1}{N_{\sigma_\tau^2}}\sum_{i=1}^{N_{\sigma_\tau^2}}\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma}_i \end{pmatrix}^{-1}\frac{1}{N_G}\sum_{E\in\mathscr{E}}\begin{pmatrix} 1 & r_{\mathscr{E}}(\mathbf{x})^T \\ r_{\mathscr{E}}(\mathbf{x}) & r_{\mathscr{E}}(\mathbf{x})r_{\mathscr{E}}(\mathbf{x})^T \end{pmatrix}\right] \tag{2.14}$$

where $E = \{E_1, \ldots, E_{N_G}\}$ is a sample of N_G points from $\mathscr{E}$, $\sigma_{\tau i}^2$ is the i'th draw from the distribution for σ_τ^2 giving

$$\check{\Sigma}_i = R + \sigma_{\tau i}^2 I$$

and

$$r_{\mathscr{E}j}(\mathbf{x}) = e^{-\sum_{i=1}^{k}\theta_i d_{\mathscr{E}i}^2(\mathbf{x},\mathbf{x}_j)}.$$

In this approach, the embedding $\mathscr{E}$ preserves both the distance and true dimensionality of the design region for the response. This allows greater flexibility in the types of functions we can fit due to the k-vector θ (or ρ) parameterizing the correlation function, at the cost of making an assumption on the true dimensionality of the design region. This has a number of consequences, namely:

1. Assessing the true dimensionality of the design region may not be easy, as suggested by Figure 2.5.

2. We must take the points E as being the embedding of points G, hence $E = \{e(G_1), \ldots, e(G_{N_G})\}$. However, while the set of points G may be an equally spaced grid in $\mathscr{G}$, it is unlikely that the resulting points E are equally spaced in $\mathscr{E}$.

3. In the case that $\theta_1 = \theta_2 = \ldots = \theta_k$, then this model degenerates back to the distance approximation approach (Model 1) described above, as one might expect.

In fact, the assumed ISOMAP mapping will certainly give us an embedding space of some sort, but it is still an assumption.

Figure 2.8(a) shows an equally spaced grid of points projected from $\mathscr{G}$-space to $\mathscr{E}^3$-space. One can notice that the projected points in Figure 2.8(b) no longer appear to be equally spaced, particularly in the four "tail" sections which have a much denser

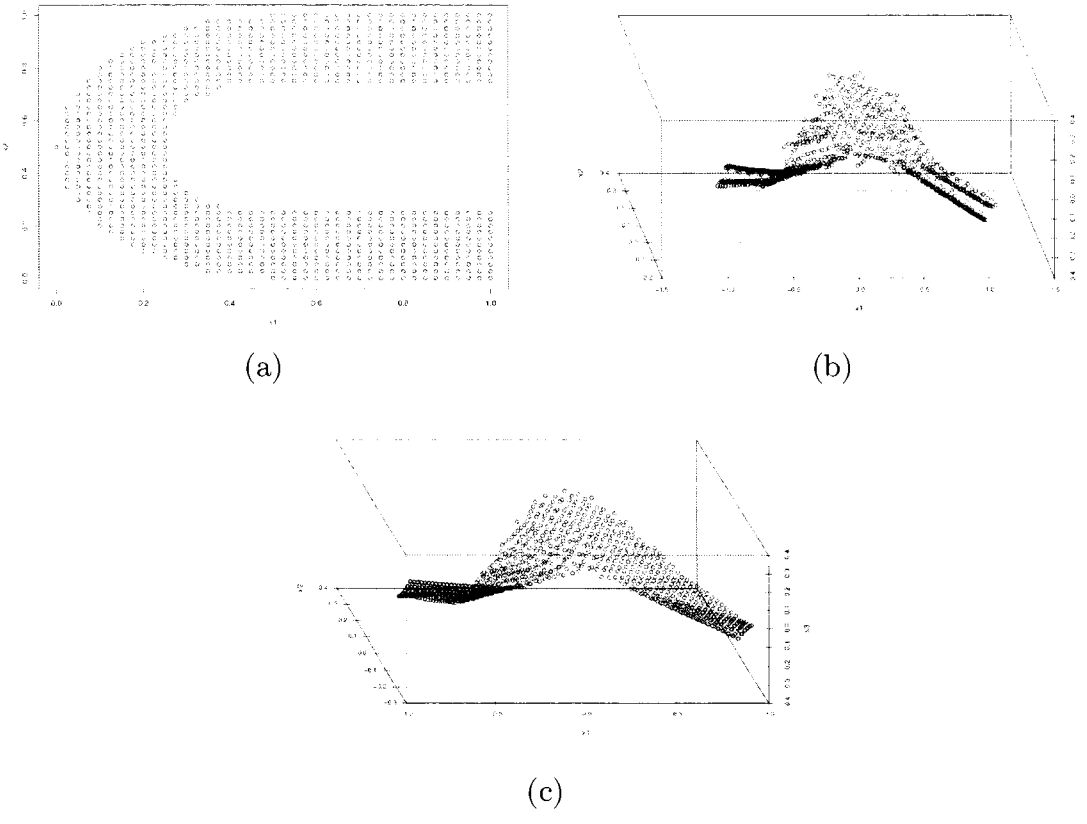


Figure 2.8: (a) Equally spaced points in the Horseshoe region, (b) their embedding in $\mathscr{E}^3$ when geodesics are calculated on a 0.10 grid and (c) their embedding in $\mathscr{E}^3$ when geodesics are calculated on a 0.05 grid

packing of points than the rest of the embedded region. This is not unexpected since the embedding of the horseshoe region only conserves approximate distances. The situation can be improved somewhat by doubling the density of points used in calculating the geodesic distances in $\mathscr{G}$-space which result in a better approximation and lead to the improved embedding shown in Figure 2.8(c). Nonetheless, there is still a noticeably higher density of points in the "tails" in this embedding.

When we are designing over the embedding space $\mathscr{E}$ (e.g., again we are now assuming that $\mathscr{E}$ is the true space rather than $\mathscr{G}$), we would like to have a set of equally spaced candidate points in $\mathscr{E}$-space. However since the embedded points may not retain the equal spacing, we should select a new set of candidates which are

equally spaced in $\mathscr{E}$-space in order to find our design. We handle this issue with a simple approach:

- Project a much denser grid of equally spaced points from $\mathscr{G} \rightarrow \mathscr{E}$
- Use a space-filling algorithm such as (Johnson et al., 1990) to select a smaller candidate set of points from those projected to $\mathscr{E}$. This candidate set is now an equally-spaced set of points covering $\mathscr{E}$.

Finally, one can note that when $\theta_1 = \ldots = \theta_k$, then the correlation function is

$$r_{\mathscr{E}j}(\mathbf{x}) = e^{-\sum_{i=1}^{k} \theta_i d^2_{\mathscr{E}i}(\mathbf{x},\mathbf{x}_j)} = e^{-\theta_0 d^2_{\mathscr{E}}(\mathbf{x},\mathbf{x}_j)} = \hat{r}_{\mathscr{G}}(\mathbf{x}),$$

in which case the problem degenerates to the distance approximation approach.

2.4 Exchange Algorithm

The actual construction of I-Optimal designs is no trivial matter. We make use of the popular exchange algorithm (Silvey, 1980; Muller, 2001) to construct approximate optimal designs. This is a computational method that searches for the best subset of points over a finite grid covering the design region of interest. Suppose we start with a set $g \in \mathscr{G}$ of n_g points covering our design region $\mathscr{G}$ and we want a design $X \in g$ of size n_X. The basic steps of the exchange algorithm are then as follows:

1. Split the set g into $g = g_X \cup \bar{g}_X$ where g_X is a candidate design of size n_X.
2. Calculate the optimality criterion $J(g_X)$.
3. Construct the set of candidate designs $\{g^*_X\}$ that result from all possible $n_X \times (n_g - n_X)$ one-point exchanges between g_X and $\bar{g}_X$.
4. Calculate the optimality criterion for each candidate design in $\{g^*_X\}$.
5. Replace g_X with the ith design from $\{g^*_X\}$ whose criterion $J(\{g^*_X\}_i) < J(g_X)$.
6. Repeat steps 2-5 until no further reduction in the criterion is possible.

The exchange algorithm is then of order $O(kn_X(n_g - n_X))$ where k is the number of repetitions required until convergence. Typically, the number of repetitions required for the types of designs constructed in this thesis ranged from $k = 3$ to $k = 5$. So, while convergence seems reasonably fast, it becomes increasingly burdensome for larger designs. As well, one needs to keep in mind that at each stage of the algorithm, the calculation of the optimal criterion requires the Monte Carlo integration over $\mathscr{G}$ and σ_τ^2. In the types of designs constructed in this thesis, runtimes were generally in the range of a few hours for smaller designs (say $n_X < 20$) up to a few days for larger designs ($n_X \approx 50$).

Typically the initial design g_X in step 1 is found by just a random selection of n_X points from g. It is important to note that we use the same draw of $N_{\sigma_\tau^2}$ points from the distribution for σ_τ^2 whenever calculating the criterion in the exchange algorithm. If instead we draw a new sample from the distribution for σ_τ^2 at each repetition of steps 2-5, then the criterion value for a given design g_X will not be the same between repetitions and therefore the algorithm will not be able to converge. In addition, even upon convergence of the algorithm, we are unlikely to have achieved the true optimal design allowed by our candidate points because of the random starting design or the possibility that the simultaneous exchange of two or more points would allow us to minimize the criterion further (Silvey, 1980). Therefore, the exchange algorithm can be independently replicated a number of times with a different random starting point and draw of $N_{\sigma_\tau^2}$ points from the distribution of σ_τ^2 to attempt to overcome any local optimality. We can then take the design with the minimum criterion over all these replicates as our best I-Optimal design.

2.5 Summary of Methods

Let us now summarize the approach presented for both modeling and performing optimal design for non-convex regions. The steps to model our response using the distance approximation approach are:

1. Fill the design region Ω with n_g points and calculate the $n_g \times n_g$ matrix of geodesic distances. These points and their geodesic distance now form a sample

of points $g \in \mathscr{G}$ from $\mathscr{G}$-space.

2. Project the g points from $\mathscr{G}$-space into $\mathscr{E}$-space using ISOMAP and find the dimension k such that projection into $\mathscr{E}^k$ minimizes (2.11).

3. Replace the geodesic distances computed in step 1 with the approximation found using the embedding space from step 2.

4. Construct a grid $G \in \mathscr{G}$ of size N_G points and project these points into $\mathscr{E}^k$-space using (2.12). These will be our prediction points.

5. Assume we have a set $X \in \mathscr{G}$ of n_X observation points with corresponding response vector y.

6. Project the points X into $\mathscr{E}^k$-space using (2.12) and denote the embedding $e(X)$.

7. Denote the matrix $D_{\mathscr{E}^k}$ as the $n_X \times n_X$ matrix of approximate geodesic distances between observation sites found from the embedding $e(X)$.

8. Denote the matrix $d_{\mathscr{E}^k}$ as the $N_G \times n_X$ matrix of approximate geodesic distances between prediction sites G and observation sites X.

9. Given the response vector y observed at $X \in \mathscr{G}$ and the approximate distances from step 7, find the MLE's of σ_z^2, σ_ε^2 and θ_0 by maximizing (2.3).

10. Find the BLUP $\hat{Y}(\mathbf{x})$, $\forall \mathbf{x} \in G$, where $r(\mathbf{x})$ is approximated using $d_{\mathscr{E}^k}$ from step 8.

If instead we are interested in predicting the response in the embedding space, the steps required become:

1. Fill the design region Ω with n_g points and calculate the $n_g \times n_g$ matrix of geodesic distances. These points and their geodesic distance now form a sample of points $g \in \mathscr{G}$ from $\mathscr{G}$-space.

2. Project the g points from $\mathscr{G}$-space into $\mathscr{E}$-space using ISOMAP and find the dimension k such that projection into $\mathscr{E}^k$ minimizes (2.11). Alternatively, one could specify a particular dimension k.

3. Construct a grid $G \in \mathscr{G}$ of size N_G points and project these points into $\mathscr{E}^k$-space using (2.12). Denote the embedded set as E. These will be our prediction points.

4. Assume we have a set $X \in \mathscr{G}$ of n_X observation points with corresponding response vector y.

5. Project the points X into $\mathscr{E}^k$-space using (2.12) and denote the embedding $e(X)$.

6. Given the response vector y observed at $e(X) \in \mathscr{E}$, find the MLE's of σ_z^2, σ_ε^2 and $\theta = (\theta_1, \ldots, \theta_k)$ by maximizing (2.3) *in* $\mathscr{E}^k$*-space.*

7. Find the BLUP $\hat{Y}(\mathbf{e})$, $\forall \mathbf{e} \in E$, where $r(\mathbf{e})$ is of course computed in $\mathscr{E}^k$-space.

8. From the mapping of points $G \rightarrow E$, use the reverse mapping to reconstruct the predicted response $\hat{Y}(\mathbf{x})$ in $\mathscr{G}$-space.

The overall procedure to search for the optimal design with the distance approximation approach is:

1. Fill the design region Ω with n_g points and calculate the $n_g \times n_g$ matrix of geodesic distances. These points and their geodesic distance now form a sample $g \in \mathscr{G}$ from $\mathscr{G}$-space. These will be our candidate points.

2. Project the g points from $\mathscr{G}$-space into $\mathscr{E}$-space using ISOMAP and find the dimension k such that projection into $\mathscr{E}^k$ minimizes (2.11).

3. Replace the geodesic distances computed in step 1 with the approximation found using the embedding space from step 2.

4. Construct a grid $G \in \mathscr{G}$ of size N_G points and project these points into $\mathscr{E}^k$-space using (2.12). These will be our integration points.

5. Set the number of N_R replicate designs to find (e.g., $N_R = 20$) and construct N_R draws of size $N_{\sigma_\tau^2}$ from the prior distribution of σ_τ^2 and N_R random initial designs of size n_X.

6. For each $i = 1 \ldots N_R$, run the exchange algorithm described above with the set of candidate points g, the ith random starting design drawn from g, the ith draw of $N_{\sigma_\tau^2}$ points from the prior distribution of σ_τ^2 to approximate the integral of σ_τ^2, the set G to approximate the integral over $\mathscr{G}$ and the correlation matrix R and vector $r(\mathbf{x})$ calculated using the distance approximation from the embedding space.

7. Of the N_R resulting designs, take the design with minimum IMSE as the optimal design.

If instead we are interested in finding the optimal design using the embedding space, the steps required become:

1. Fill the design region Ω with n_g points and calculate the $n_g \times n_g$ matrix of geodesic distances. These points and their geodesic distance now form a sample of points $g \in \mathscr{G}$ from $\mathscr{G}$-space.

2. Project the g points from $\mathscr{G}$-space into $\mathscr{E}$-space using ISOMAP and find the dimension k such that projection into $\mathscr{E}^k$ minimizes (2.11). Alternatively, one could specify a particular dimension k.

3. Construct a grid $G \in \mathscr{G}$ of size N_G points and project these points into $\mathscr{E}^k$-space using (2.12). Denote the embedded set as E. These will be our integration points.

4. Select n_g points in E that are space filling in $\mathscr{E}^k$, and denote this set as e. These will be our candidate points.

5. Set the number of N_R replicate designs to find (e.g., $N_R = 20$) and construct N_R draws of size $N_{\sigma_\tau^2}$ from the prior distribution of σ_τ^2 and N_R random initial designs of size n_X.

6. For each $i = 1 \ldots N_R$, run the exchange algorithm described above with the set of candidate points e, the ith random starting design drawn from e, the ith draw of $N_{\sigma_\tau^2}$ points from the prior distribution of σ_τ^2 to approximate the integral of σ_τ^2, and the set E to approximate the integral over $\mathscr{E}^k$.

7. Of the N_R resulting designs, take the design with minimum IMSE as the optimal design.

8. From the mapping of points $G \rightarrow E$, use the reverse mapping to reconstruct the optimal design in $\mathscr{G}$-space.

With the basic algorithmic steps outlined, we proceed in Chapter 3 to investigate I-Optimal designs on non-convex regions with a simulation study and an applied example of these procedures.

Chapter 3

Results

In order to evaluate the effectiveness of the proposed design methods on prediction error, a number of examples are investigated. We first perform a number of simulation studies for functions over the horseshoe region. The horseshoe region makes for a good example since it is somewhat of a worse-case scenario. This is because points at the end of the horseshoe are nearby in terms of their Euclidean distance, but actually very far apart in terms of the geodesic distance. We consider random functions that are drawn from the GP under Model 1 ($\mathscr{G}$-space) or Model 2 ($\mathscr{E}$-space). A subsequent study considers functions constructed from a second-order linear model in $\mathscr{E}$-space. Finally, we investigate the real-world example of the Florida region and temperature data. In this case, we consider a large design constructed under Model 1 and compare its performance to a reasonable alternative.

3.1 Simulation Study in $\mathscr{G}$-Space

We investigate the performance of I-Optimal designs of the form (2.13) and (2.14) by drawing random functions $f_{\mathscr{G}}$ from Model 1 (i.e., GP in $\mathscr{G}$-space) and comparing the performance of I-Optimal designs to randomly selected points and space filling designs (Johnson et al., 1990). That is, our response in this case is drawn from Model 1, but we consider designs constructed under Models 1 and 2. The basic outline for the study is then:

- Construct designs of size n_X points using criteria (2.13) and (2.14)
- Draw 3,000 realizations from the Gaussian Process and fit the model by maximum likelihood
- Compute the empirical mean square error for each fitted model and report the average.

We consider designs of size $n_X = 16$ that are selected from a candidate set of $n_G = 63$ points that fill the horseshoe region in a (0.10×0.10)-spaced grid. The candidate set was constructed by first filling the $[0,1]^2$ region with the (0.10×0.10) grid and then from this grid selecting only those points that lie in the horseshoe region. The geodesic distances were also computed using a (0.10×0.10)-spaced grid, and based on the distance approximation errors shown in Figure 2.5, the distances were approximated in the embedding space $\mathscr{E}^3$. Since in this case we are drawing random functions in $\mathscr{G}$-space, we will only have one parameter ρ_0 (we refer to the correlation parameter ρ rather than θ from this point onwards due to the easier interpretation of ρ as discussed in Section 2.1). The random functions were constructed with mean $\mu = 0$ and covariance matrix Σ parameterized by ρ_0 and $\sigma_z^2 = 0.9$ and defined on the n_G points. We use the predicted response given this correlation matrix as our "true" response surface and added the normally distributed measurement error with mean 0 and variance $\sigma_\varepsilon^2 = 0.1$ to obtain the observed response values.

A grid of $N_G = 2,239$ points were selected for computing the Monte Carlo integral in evaluating the optimality criterions. This was done by filling the $[0,1]^2$ region with a grid of 4,096 points, and then selecting those points which lie in the horseshoe region. A sample of size $N_{\sigma_\tau^2} = 2,048$ was drawn from the prior distribution of σ_τ^2. We assumed that the variance of the response $var(Y) = 1$ (we can scale our response accordingly) and that 90% of this variability is systemic while 10% is measurement error, which is what would be expected for the generation of the random functions described above. Under this assumption, we placed inverse gamma prior distributions on σ_z^2 and σ_ε^2 centered at 0.9 and 0.1 respectively (the latter being truncated at 0.15) as motivated by (Chipman, 1997; Linkletter et al., 2005). The prior on σ_τ^2 was constructed by taking the ratio $\frac{\sigma_\varepsilon^2}{\sigma_z^2}$, and a histogram of this prior is shown in Figure 3.1.

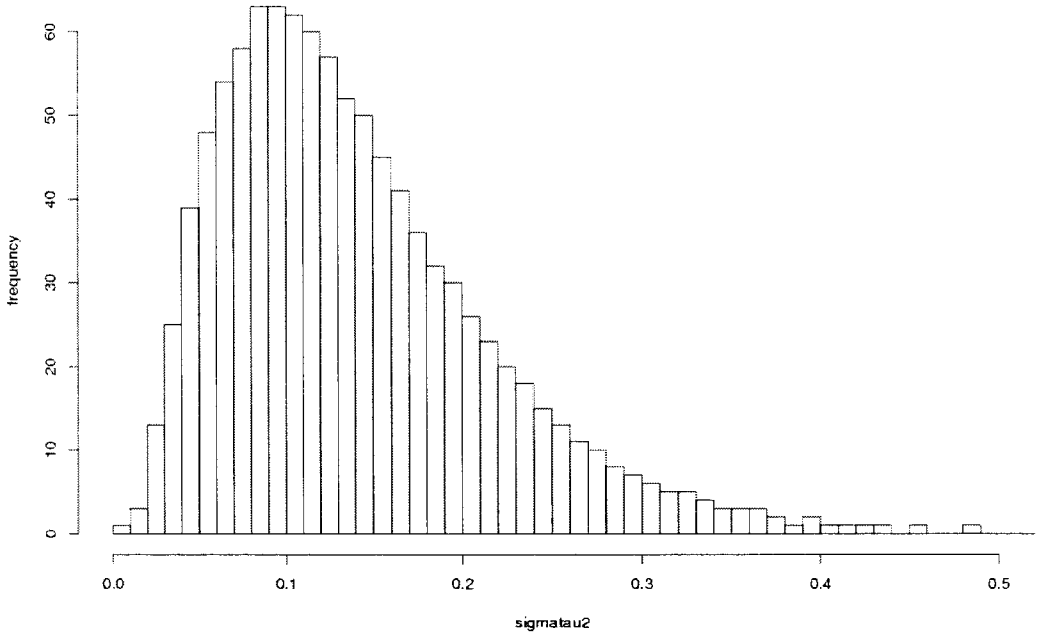
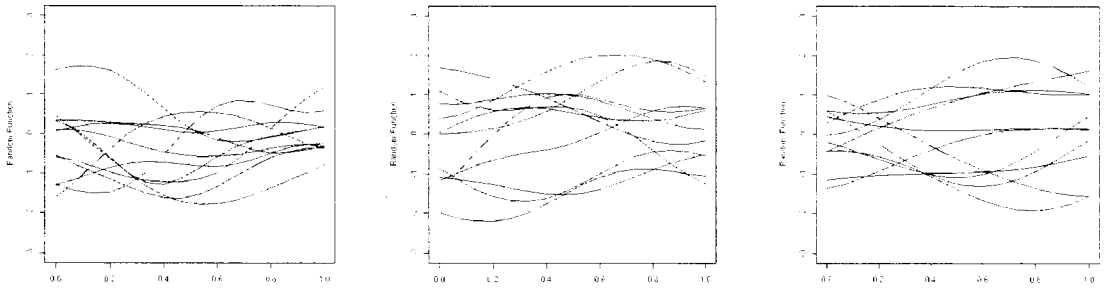


Figure 3.1: Prior distribution for σ_τ^2

We then consider five possible design constructions:

- Designs constructed in $\mathscr{G}$-space using Model 1 (distance approximation approach)
- Designs constructed in $\mathscr{E}^2$-space using Model 2 (embedding space approach)
- Designs constructed in $\mathscr{E}^3$-space using Model 2 (embedding space approach)
- A space filling design in $\mathscr{G}$-space
- A random design in $\mathscr{G}$-space

Designs constructed under Model 1 ($\mathscr{G}$-space) use the original 63 regularly spaced grid points as a candidate set, from which 16 design points will be selected. In contrast, designs constructed under Model 2 ($\mathscr{E}^2$-space and $\mathscr{E}^3$-space) use 63 points



(a) (b) (c)

Figure 3.2: 10 Realizations of the Gaussian Process with (a) $\rho = 0.2$, (b) $\rho = 0.4$ and (c) $\rho = 0.6$

from their respective embedding spaces as candidate sets for selecting the 16 design points. Note that designs constructed in $\mathscr{E}^2$-space are done under a more erroneous distance approximation.

In order to compare the effectiveness of these I-Optimal designs, two additional procedures for selecting design points were considered as a baseline: random designs and a space-filling design. The random designs are simply 16 randomly selected points from $\mathscr{G}$ which are selected independently for each replicate of the study. Essentially, this gives the worse behaviour one would expect to do when selecting 16 points. The space-filling design selects 16 points in $\mathscr{G}$ which minimizes the geodesic distance between these points and the unselected points remaining in $\mathscr{G}$ (Johnson et al., 1990). In other words, this design is space-filling in $\mathscr{G}$-space. Comparing to the space-filling design shows the improvement that can be realized by an I-Optimal design over a simpler method likely to be used by a practitioner.

We consider 3 settings for the correlation parameter in drawing our responses: $\rho_0 = 0.2$, $\rho_0 = 0.4$ and $\rho_0 = 0.6$. To better understand what these functions might look like, Figure 3.2 shows 10 realizations at each of these values of ρ for the 1-dimensional Gaussian Process. The general behaviour one might notice is that functions with $\rho = 0.2$ are fairly complex, while as ρ increases, the realizations become less so.

The results of the simulation study responses drawn from Model 1 (referred to as $f_{\mathscr{G}}$) are shown in Tables 3.1, 3.2, 3.3 and 3.4. Table 3.1 contains the averaged empirical mean squared prediction errors for all three I-Optimal designs as well as the random designs and the space-filling design. As expected, the EMSE for the random designs was noticeably higher than all other approaches. Space-filling designs appear to do well, typically performing only slightly worse than the optimal designs.

Table 3.2 contains the relative efficiencies of the optimal designs for each level of ρ_0 investigated. Reading column-wise for each design model, a relative efficiency of 1.0 will occur in the column $f_{\mathscr{G},\rho=\rho_0}$ when $\rho_{design} = \rho_0$, since this is the best we should be able to do. By comparing the efficiency of designs in each column (and for each design model separately) we can see how the efficiency of designs change when $\rho_{design} \neq \rho_0$. Then, by comparing across columns, we can see which design is robust in the sense of maximizing the minimum efficiency. For the I-Optimal designs we see that $\rho_{design} = 0.4$ gave the greatest robustness while for the 2-D I-Optimal designs $\rho_{design} = 0.2$ was best, and finally for the 3-D I-Optimal designs, $\rho_{design} = 0.2$ was again robust.

We compare the random and space-filling designs to the best optimal designs for the three types of response functions. The efficiency of random designs relative to optimal designs is shown in Table 3.3, and confirms the uniformly poor predictive ability of randomly selected points, ranging from a low of 60% efficiency for more complex responses (small ρ_0), to as high as 79% efficiency for simpler responses (large ρ_0). It is clear that the more complex the response (e.g., as ρ_0 decreases), the lower the efficiency of randomly selected points.

The efficiency of space-filling designs relative to optimal designs is shown in Table 3.4, and indicates that space-filling designs retain good efficiency, ranging from 88% to 100%. Nonetheless, we again observe the trend that as the complexity of the response function increases, the efficiency of space-filling designs relative to optimal designs decreases. In terms of model comparison, we see that designs constructed under Model 1 gave better performance in terms of both relative efficiency and empirical prediction errors. This was to be expected since our response model was also Model 1. Optimal designs constructed with the mis-specified distance under Model 2 in $\mathscr{E}^2$-space demonstrated the worse performance of the optimal designs considered in

Table 3.1: Average Empirical Mean Squared Error over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{G}}$ in the Horseshoe Region

Design	ρ_{design}	$f_{\mathscr{G},\rho=0.2}$	$f_{\mathscr{G},\rho=0.4}$	$f_{\mathscr{G},\rho=0.6}$
Random	-	0.151	0.107	0.0781
Space Filling	-	0.103	0.0758	0.0595
Model 1 (I-Optimal)	0.2	0.0923	0.0690	0.0557
	0.4	0.0911	0.0675	0.0540
	0.6	0.0975	0.0698	0.0549
Model 2 (2-D I-Optimal)	0.2	0.101	0.0751	0.0611
	0.4	0.105	0.0762	0.0600
	0.6	0.126	0.106	0.0917
Model 2 (3-D I-Optimal)	0.2	0.0965	0.0773	0.0632
	0.4	0.102	0.0760	0.0625
	0.6	0.110	0.0765	0.0612

terms of both relative efficiency and prediction errors.

Finally, there is some error present that results from performing Monte Carlo integration in constructing our designs as well as fitting the models using maximum likelihood. This can be seen by the relative efficiencies that were found to be marginally above 100%, which cannot actually occur. Nonetheless, the behaviour of our different design models is still evident from these results. An unexpected result occured for designs constructed with $\rho_0 = 0.6$ under Model 2 in $\mathscr{E}^2$-space. Here we see relative efficiencies of the designs for $\rho_0 = 0.2, 0.4$ being much higher than would be expected. In turns out that the design for $\rho_0 = 0.6$ under Model 2 was the only design constructed that had a replicate point, which lead to a much higher prediction error resulting in high relative efficiency seen for the alternative designs constructed with $\rho_0 = 0.2, 0.4$. In this case, it would seem that constructing a design with mis-specified distance lead to the introduction of a replicate design point when likely none should have been present.

Table 3.2: Relative Efficiencies of Optimal Designs over 3000 Simulated Response Surfaces drawn from $f_{\mathcal{G}}$ in the Horseshoe Region

Design	ρ_{design}	$f_{\mathcal{G},\rho=0.2}$	$f_{\mathcal{G},\rho=0.4}$	$f_{\mathcal{G},\rho=0.6}$
Model 1	0.2	1.00	0.98	0.98
I-Optimal	0.4	1.01	1.00	1.02
	0.6	0.95	0.97	1.00
Model 2	0.2	1.00	1.01	1.50
(2-D I-Optimal)	0.4	0.97	1.00	1.53
	0.6	0.80	0.72	1.00
Model 2	0.2	1.00	0.98	0.97
(3-D I-Optimal)	0.4	0.95	1.00	0.98
	0.6	0.88	0.99	1.00

Table 3.3: Efficiency of Random Designs Relative to the Best I-Optimal Design over 3000 Simulated Response Surfaces drawn from $f_{\mathcal{G}}$ in the Horseshoe Region

Design	$f_{\mathcal{G},\rho=0.2}$	$f_{\mathcal{G},\rho=0.4}$	$f_{\mathcal{G},\rho=0.6}$
Model 1 (I-Optimal)	0.60	0.63	0.69
Model 2 (2-D I-Optimal)	0.67	0.70	0.77
Model 2 (3-D I-Optimal)	0.64	0.71	0.78

Table 3.4: Efficiency of Space-Filling Designs Relative to the Best I-Optimal Design over 3000 Simulated Response Surfaces drawn from $f_{\mathcal{G}}$ in the Horseshoe Region

Design	$f_{\mathcal{G},\rho=0.2}$	$f_{\mathcal{G},\rho=0.4}$	$f_{\mathcal{G},\rho=0.6}$
Model 1 (I-Optimal)	0.88	0.89	0.91
Model 2 (2-D I-Optimal)	0.98	0.99	1.00
Model 2 (3-D I-Optimal)	0.93	1.00	1.03

3.2 Simulation Study in $\mathscr{E}^2$-Space

Here we investigate the performance of I-Optimal designs constructed under Models 1 and 2 by drawing random functions $f_{\mathscr{E}^2}$ from the Gaussian Process under Model 2 and again comparing the performance of I-Optimal designs to randomly selected points and space filling designs. We construct designs of size $n_X = 16$ that are selected from a candidate set of $n_G = 63$ points that fill the embedding in $\mathscr{E}^2$-space. The geodesic distances were computed using a 0.10-spaced grid. Since in this case we are drawing random functions and constructing designs in $\mathscr{E}^2$-space, we have two parameters ρ_1, ρ_2. However, since we are assuming $\rho_1 = \rho_2$ in this study, we will simply refer to these correlation parameters as ρ. Responses drawn from this model consider the case where the response is a function of the embedding space, but not the one we would expect (i.e., not the distance-preserving embedding $\mathscr{E}^3$). The random functions were again constructed with mean $\mu = 0$ and covariance matrix Σ parameterized by ρ and $\sigma_z^2 = 0.9$ defined on the n_G points. We use the predicted response given this correlation matrix as our "true" response surface and added the $\sigma_\varepsilon^2 = 0.1$ measurement error to obtain the observed response values. All other parameters remained as above, and we again consider the same five possible design constructions.

The results of the simulation study for the functions of type $f_{\mathscr{E}^2}$ are shown in Tables 3.5, 3.6, 3.7 and 3.8. Table 3.5 contains the averaged empirical mean squared prediction errors for all three I-Optimal designs as well as the random designs and the space-filling design. The EMSE for the random designs was again noticeably higher than all other approaches while space-filling designs continue to perform only a little worse than the optimal designs, particularly at larger values of ρ.

Table 3.6 contains the relative efficiencies of the optimal designs for each level of ρ investigated. For the Model 1 I-Optimal designs we see that $\rho_{design} = 0.4$ gave the greatest robustness. For the Model 2, 2-D I-Optimal designs, $\rho_{design} = 0.4$ was also best, and finally for the Model 2, 3-D I-Optimal designs, $\rho_{design} = 0.2$ was robust.

We compare the random and space-filling designs to the best optimal designs for the three types of response functions. The efficiency of random designs relative to optimal designs is shown in Table 3.7, and confirms the uniformly poor predictive ability of randomly selected points, ranging from a low of 63% efficiency to 82%

Table 3.5: Average Empirical Mean Squared Error over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{E}^2}$ in the Horseshoe Region

Design	ρ_{design}	$f_{\mathscr{E}^2,\rho=0.2}$	$f_{\mathscr{E}^2,\rho=0.4}$	$f_{\mathscr{E}^2,\rho=0.6}$
Random	-	0.128	0.0834	0.0698
Space Filling	-	0.0934	0.0659	0.0528
Model 1	0.2	0.0807	0.0636	0.0508
(I-Optimal)	0.4	0.0827	0.0607	0.0518
	0.6	0.0891	0.0638	0.0519
Model 2	0.2	0.0807	0.0636	0.0536
(2-D I-Optimal)	0.4	0.0836	0.0609	0.0562
	0.6	0.109	0.0951	0.0823
Model 2	0.2	0.0844	0.0681	0.0565
(3-D I-Optimal)	0.4	0.0875	0.0699	0.0561
	0.6	0.0850	0.0701	0.0549

efficiency. It is again clear that as ρ decreases, the lower the efficiency of randomly selected points.

The efficiency of space-filling designs relative to optimal designs is shown in Table 3.8. These designs again demonstrate reasonably good efficiency with values ranging from about 86% to 100%. As before, as ρ decreases, the efficiency of the space-filling design degrades. We again saw that optimal designs constructed under Model 1 gave comparable performance to designs constructed under the true model, which in this case was Model 2 in $\mathscr{E}^2$-space. Designs constructed under Model 2 in $\mathscr{E}^3$-space gave the worse performance of the optimal designs.

3.3 Simulation Study in $\mathscr{E}^3$-Space

Here we investigate the performance of I-Optimal designs constructed under Model 1 and Model 2 by drawing random linear functions modeled in $\mathscr{E}^3$-space. This is motivated by two reasons. The first is to consider a response model similar in spirit to the assumed linear model from the Welding example discussed earlier. Second, it makes no sense to consider a GP response modeled in $\mathscr{E}^3$ as this will construct the

Table 3.6: Relative Efficiencies of Optimal Designs over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{E}^2}$ in the Horseshoe Region

Design	ρ_{design}	$f_{\mathscr{E}^2,\rho=0.2}$	$f_{\mathscr{E}^2,\rho=0.4}$	$f_{\mathscr{E}^2,\rho=0.6}$
Model 1	0.2	1.00	0.95	1.02
(I-Optimal)	0.4	0.98	1.00	1.00
	0.6	0.91	0.95	1.00
Model 2	0.2	1.00	0.96	1.53
(2-D I-Optimal)	0.4	0.97	1.00	1.47
	0.6	0.74	0.64	1.00
Model 2	0.2	1.00	1.03	0.97
(3-D I-Optimal)	0.4	0.96	1.00	0.98
	0.6	0.99	1.00	1.00

Table 3.7: Efficiency of Random Designs Relative to the Best I-Optimal Design over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{E}^2}$ in the Horseshoe Region

Design	$f_{\mathscr{E}^2,\rho=0.2}$	$f_{\mathscr{E}^2,\rho=0.4}$	$f_{\mathscr{E}^2,\rho=0.6}$
Model 1 (I-Optimal)	0.63	0.73	0.73
Model 2 (2-D I-Optimal)	0.63	0.73	0.77
Model 2 (3-D I-Optimal)	0.66	0.82	0.79

Table 3.8: Efficiency of Space-Filling Designs Relative to the Best I-Optimal Design over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{E}^2}$ in the Horseshoe Region

Design	$f_{\mathscr{E}^2,\rho=0.2}$	$f_{\mathscr{E}^2,\rho=0.4}$	$f_{\mathscr{E}^2,\rho=0.6}$
Model 1 (I-Optimal)	0.86	0.92	0.96
Model 2 (2-D I-Optimal)	0.86	0.92	1.02
Model 2 (3-D I-Optimal)	0.90	1.03	1.04

same responses from the study of Section 3.1 because $\mathscr{E}^3$ is the embedding space used in approximating the geodesic distances.

We consider a second order model where the true response is given by

$$Y_{true} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_2 x_3 + \beta_9 x_1 x_3 + \beta_{10} x_1 x_2 x_3,$$

and the observed response will be

$$Y = Y_{true} + error$$

where the error term is simply $N(0, \sigma_\varepsilon^2)$. We make Y_{true} a random function by drawing the β's from normal distributions and requiring the interaction terms to be bounded below the main effect terms with high probability in order to preserve the effect ordering principle.

We will again compare our I-Optimal, random and space-filling designs on this model. Designs of size $n_X = 16$ that are selected from a candidate set of $n_G = 63$ points that fill the embedding in $\mathscr{E}^3$-space. The geodesic distances were computed using a 0.10-spaced grid. In this study, we are using the distance-preserving embedding, but our true response lies in $\mathscr{E}^3$-space rather than $\mathscr{G}$-space. All other parameters remained as in the first two studies, and we again consider the same five possible design constructions.

The results of the simulation study for the linear functions $f_{\mathscr{E}^3,LM}$ are shown in Tables 3.9. This table contains the averaged empirical mean squared prediction errors for all three I-Optimal designs as well as the random designs and the space-filling design. The EMSE for the random designs was again noticeably higher than all other approaches while in this example Space-filling did not offer great performance.

Among the optimal designs, both the Model 1 I-Optimal and Model 2, 3-D I-Optimal designs gave better performance than the Model 2, 2-D I-Optimal design. The best overall design was the Model 2, 3-D I-Optimal design with $\rho = 0.2$, which is expected given the true model of our random responses. The relative efficiency of the space-filling design to this best optimal design was quite low at 58%. The relative efficiency of the best Model 1, I-Optimal design to the best overall design was about 81% while the relative efficiency of the best Model 2, 2-D I-Optimal design to the best overall design was about 55%.

Table 3.9: Average Empirical Mean Squared Error over 1000 Simulated Response Surfaces drawn from $f_{\mathscr{G}^3,LM}$ in the Horseshoe Region

Design	ρ_{design}	$f_{\mathscr{G}^3,LM}$
Random	-	0.275
Space Filling	-	0.162
Model 1	0.2	0.121
I-Optimal	0.4	0.116
	0.6	0.125
Model 2	0.2	0.181
(2-D I-Optimal)	0.4	0.170
	0.6	0.247
Model 2	0.2	0.0939
(3-D I-Optimal)	0.4	0.129
	0.6	0.111

3.4 Florida Study

To demonstrate our method on a real-world example, we return to the Florida water temperature shown in Figure 1.4. The well recognized Florida state is shown in green on the right side of the figure, while the black area to the left indicates an area that we aren't interested in modeling as the task in mind is modeling the behaviour of this waterway along the Florida coastline.

Using a grid of 0.05-spaced points, we constructed the geodesic distances and searched for a design of size $n = 50$ points. We considered an optimal design constructed under Model 1 with correlation parameter $\rho = 0.4$, and a space-filling design constructed in $\mathscr{G}$-space. We assumed the value of $\rho = 0.4$ for constructing the optimal design based on the results of our simulation study.

The space-filling design was taken after 100 iterations of the software, and is shown in Figure 3.3. We took the best design of 20 replicates as the I-Optimal design, which is shown in Figure 3.4. The differences between the two designs are visually quite recognizable. The pattern of design points for the space filling design is fairly equally spaced as would be expected. The pattern for the optimal design instead has variable spacing between points, with some being closer together forming small groups, with

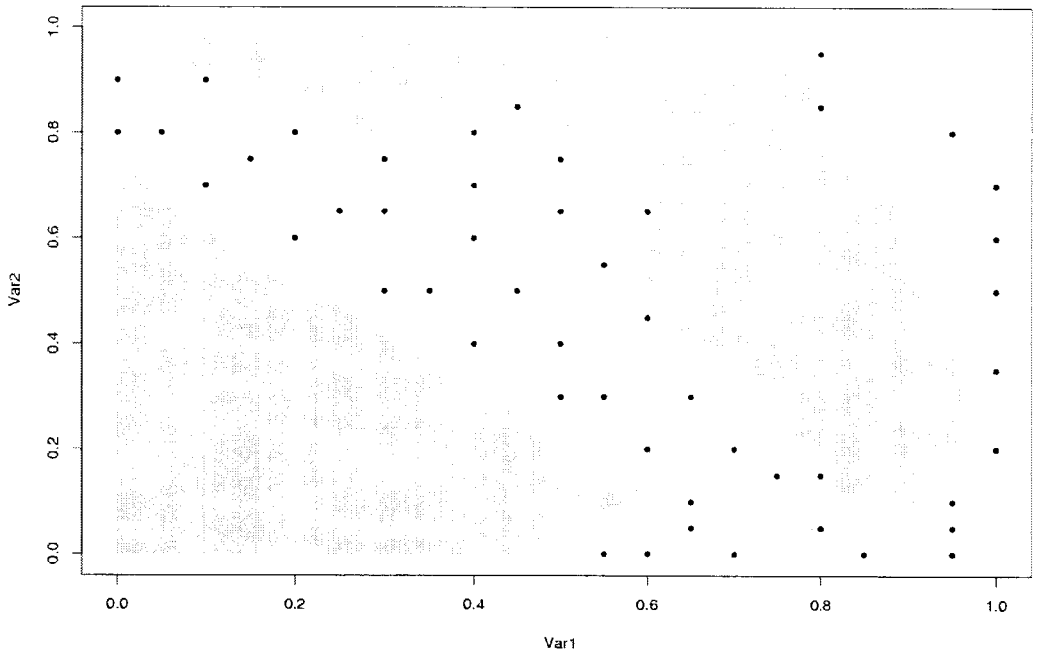


Figure 3.3: A 50 point space-filling design for the Florida coastal waterway

larger spaces separating these groups of points. There are also a number of replicated points, denoted by stars. This behaviour is not unexpected as having different spacing between points allows the optimal design to detect components of the response surface that have differing frequency, as well as being able to better estimate the measurement error with the introduction of some replicated points.

The prediction error for the space filling design was found to be about 0.0279 while for the optimal design the error was 0.0260, so in this example the optimal design reduced prediction error by about 7%. This improvement seems reasonable given the results found in our simulation study. With the design found, an applied researcher could use these points as locations to place buoys along the Florida coastal waterway for data collection and modeling of the waterway, as is done by (Merz, 2001; Weisberg et al., 2002).

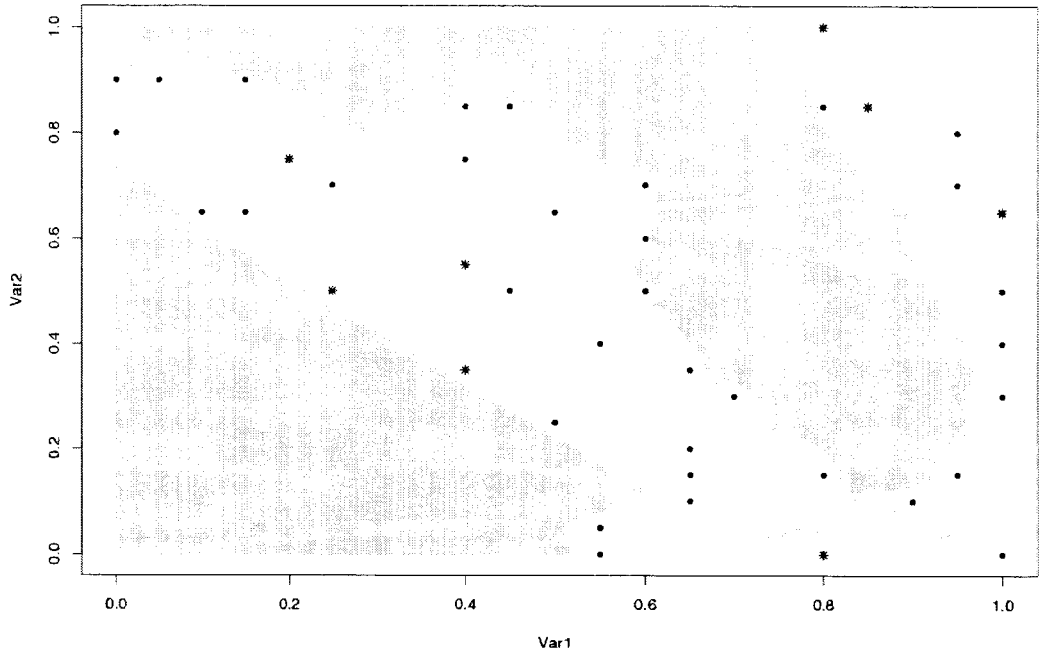


Figure 3.4: A 50 point I-Optimal design for the Florida coastal waterway

Chapter 4

Discussion

The results obtained in the simulation study provide some interesting insights into how the I-Optimal designs we have constructed behave with changes in the response surface. In particular, we saw that

- designs constructed with smaller ρ_{design} tended to be more robust
- designs constructed under Model 1 (distance-preserving approach) behaved well in almost all circumstances, even though there is only the single ρ_0 parameter
- designs constructed in the wrong embedding space often gave the worse performance.

The notion that designs constructed for smaller ρ_{design} have greater robustness might be expected. For instance, the behaviour of a flat plane response can be easily discerned from a small sample of points, but a highly variable response would require many strategically placed points to learn.

Although we could only investigate three design and modeling spaces in our study, there was some indication that designs constructed under Model 1 (distance preservation) would always perform reasonably well while designs constructed using Model 2 in the incorrect embedding space might not perform well. This was particularly true in the linear model example. This example also gave a strong indication that space-filling designs might not always give acceptable performance.

A notable limitation to the robustness study approach is the computational burden involved in constructing such a study. A practitioner might not look forward to weeks of computations to investigate the design properties for his/her non-convex region of interest under a few assumed values of the correlation parameter. It might be much nicer if, for instance, the practitioner could simply specifiy a prior distribution on the parameter ρ and achieve robustness in that way. This is an approach we would like to investigate in future work, however given the form of (2.13) and (2.14), we note that integration over $\mathscr{G}$ and σ_τ^2 were readily achievable since they are separable, but with integration over ρ this is not the case. So, the computational burden will be much higher than the criterions discussed in this thesis.

Finally, although we have made use of ISOMAP in deriving our approach to handling non-convex design regions, not all non-convex regions can be handled with ISOMAP. For instance, shapes such as rings or donuts result in eigenvalues that oscillate from being positive to negative when performing the eigendecomposition on the squared distance matrix. Since the eigenvalues no longer have intrinsic order, the criterions for selecting the appropriate embedding (2.10) and (2.11) no longer have meaning. This raises the question of what mapping is used in constructing the embedding of the design region. In fact, this is an active area of research, and other mappings (Roweis and Saul, 2000; Donoho and Grimes, 2003) would serve as good candidates for future investigations.

4.1 Conclusions

In this thesis we considered the problem of optimal design for Gaussian Processes over non-convex design regions. We considered Integrated Mean Squared Error Optimal (I-Optimal) designs when the question of interest is response surface prediction. We first extended existing results for the convex case by introducing a new formulation that allows one to consider the effect of measurement error in the construction of the designs. We then introduced a distance measure, the geodesic distance, which is gaining popularity in existing literature for modeling over non-convex regions. However, it is known that direct use of the geodesic distances in the correlation function

is not always valid. We then propose the novel approach of using a dimensionality reduction technique known as ISOMAP as a means to project our non-convex design region into a possibly higher dimensional Euclidean space with the requirement that the Euclidean space distances closely approximate the true geodesic distances. Since the correlation function is well defined for any Euclidean space, this allows us to both model and construct designs over the non-convex region.

As a result of this embedding space approach, we can take the view of designing over a "correct" embedding space, or simply using the embedding space for the purpose of geodesic distance approximation. In the former approach, the correlation of the response is assumed to behave along the axis of the embedding space. So, this allows more flexibility in the correlation function due to the increased number of parameters (one for each dimension), although it may not be valid if the assumption is incorrect. In the latter approach, we assume the correlation structure of the response is the same in all directions as our correlation function has only one parameter.

A simulation study was performed to investigate these issues by evaluating the robustness of designs at three levels of the correlation parameter and three different types of response functions. In addition, random and space-filling designs were constructed to serve as a comparison point. The simulation study revealed that lower values of the correlation parameter usually resulted in more robust designs in the range we considered. Designs constructed under distance preservation tended to perform reasonably well in all circumstances while designs constructed under an assumed embedding space sometimes performed noticeably worse when the assumption was incorrect. Space-filling designs often performed very well, and are certainly more computationally feasible. However, their performance tended to degrade as the response function complexity increased, and in the case of the linear model example, they did not perform well at all.

We also investigated a real world example involving surface temperature in the waters surrounding Florida. In this example we investigated the space filling design and the distance-approximating I-Optimal design. The I-Optimal design modestly reduced prediction errors as compared to the space-filling design.

There are notable areas that would make for interesting extensions and future

work. First, the process of performing a robustness study in the manner done for this thesis is quite computationally intensive, and yet does not give a fine enough resolution to answer all the questions we have on the effect of the correlation parameter in constructing our designs. A nice alternative would be to investigate designs constructed with prior distributions specified for the correlation parameter, although it is noted in our discussion that this will be computationally difficult to achieve. A final area for future study is the investigation of alternative embedding algorithms that may handle a larger class of non-convex regions than the ISOMAP approach considered in this thesis.

Appendix A

Gaussian Process Model

Let $Z(X) \sim N(\mu, \sigma_z^2 R), \varepsilon(X) \sim N(0, \sigma_\varepsilon^2 I), \Sigma = \sigma_z^2 R, \tilde{\Sigma} = \sigma_z^2 R + \sigma_\varepsilon^2 I, R = [r_{ij}]$. Then,

$$Y(X) = Z(X) + \varepsilon(X) \sim N(\mu, \tilde{\Sigma}),$$

where

$$\begin{aligned} r_{ij} &= e^{-\sum_{d=1}^{D} \theta_d \left(\mathbf{x}_{id} - \mathbf{x}_{jd}\right)^2} \\ &= \prod_{d=1}^{D} \rho_d^{4\left(\mathbf{x}_{id} - x_{jd}\right)^2}, \end{aligned}$$

with $\rho_d = \exp^{-\frac{1}{4}\theta_d}$. Since $\theta \in (0, \infty)$, then $\rho \in (0, 1)$.

The likelihood is $L = \frac{1}{(2\pi)^{\frac{n}{2}} |\tilde{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(y - \mathbf{1}\mu)^T \tilde{\Sigma}^{-1} (y - \mathbf{1}\mu)}$, or taking the log, we find the log-likelihood as:

$$\begin{aligned} logL &= -\frac{n}{2} log(2\pi) - \frac{1}{2} log|\tilde{\Sigma}| - \frac{1}{2} (y - \mathbf{1}\mu)^T \tilde{\Sigma}^{-1} (y - \mathbf{1}\mu) \\ &\approx -\frac{1}{2} log|\tilde{\Sigma}| - \frac{1}{2} (y - \mathbf{1}\mu)^T \tilde{\Sigma}^{-1} (y - \mathbf{1}\mu) \end{aligned}$$

The maximum likelihood estimates are:

$$
\begin{aligned}
&\hat{\mu} : \frac{\partial logL}{\partial \mu} = 0 \\
&\Rightarrow -0 - \frac{1}{2}\left(\left(\frac{\partial}{\partial \mu}(y - \mathbf{1}\mu)^T\right)\tilde{\Sigma}^{-1}(y - \mathbf{1}\mu) + 0 + (y - \mathbf{1}\mu)^T\tilde{\Sigma}^{-1}\left(\frac{\partial}{\partial \mu}(y - \mathbf{1}\mu)\right)\right) \\
&\quad = 0 \\
&\Rightarrow -\mathbf{1}^T\tilde{\Sigma}^{-1}(y - \mathbf{1}\mu) - (y - \mathbf{1}\mu)^T\tilde{\Sigma}^{-1}1 = 0 \\
&\Rightarrow \hat{\mu} = \left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\tilde{\Sigma}^{-1}y
\end{aligned}
$$

and the remaining parameters $\theta, \sigma_\varepsilon^2$ and σ_z^2 need be found numerically, except in the case when $\sigma_\varepsilon^2 = 0$, where it is possible to determine $\hat{\sigma_z^2}$ directly:

$$
\begin{aligned}
&\hat{\sigma_z^2} : \frac{\partial logL}{\partial \sigma_z^2} = 0 \\
&\quad \Rightarrow -\frac{\partial}{\partial \sigma_z^2} log|\sigma_z^2 R| - \frac{\partial}{\partial \sigma_z^2}(y - \mathbf{1}\mu)^T\left(\sigma_z^2 R\right)^{-1}(y - \mathbf{1}\mu) = 0 \\
&\quad \Rightarrow -\frac{\partial}{\partial \sigma_z^2} log\left((\sigma_z^2)^n|R|\right) - \frac{\partial}{\partial \sigma_z^2}(y - \mathbf{1}\mu)^T\frac{R^{-1}}{\sigma_z^2}(y - \mathbf{1}\mu) = 0 \\
&\quad \Rightarrow -\frac{\partial}{\partial \sigma_z^2} log\left((\sigma_z^2)^n\right) + \frac{\partial}{\partial \sigma_z^2} log|R| - \frac{\partial}{\partial \sigma_z^2}\frac{(y - \mathbf{1}\mu)^T R^{-1}(y - \mathbf{1}\mu)}{\sigma_z^2} = 0 \\
&\quad \Rightarrow -\frac{n}{\sigma_z^2} + 0 + \frac{(y - \mathbf{1}\mu)^T R^{-1}(y - \mathbf{1}\mu)}{(\sigma_z^2)^2} = 0 \\
&\quad \Rightarrow \hat{\sigma_z^2} = \frac{(y - \mathbf{1}\mu)^T R^{-1}(y - \mathbf{1}\mu)}{n}
\end{aligned}
$$

Appendix B

Best Linear Unbiased Predictor (BLUP)

The GP model is $y(\mathbf{x}_i) = \sum_h \beta_h f_h(\mathbf{x}_i) + Z(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i)$, a realization of the process $Y(\mathbf{x}) = f^T(\mathbf{x})\beta + Z(\mathbf{x}) + \varepsilon(\mathbf{x})$. In our case, $h = 1$ and we simply have $y(\mathbf{x}_i) = \mu + Z(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i)$. Let $F = \left[[f_1(\mathbf{x}_1), \ldots, f_1(\mathbf{x}_n)]^T, \ldots, [f_h(\mathbf{x}_1, \ldots, f_h(\mathbf{x}_n)]^T\right] = \mathbf{1}$ (in our case), and $\mathbf{y} = [y_1, \ldots, y_n]^T$, $Z = [Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)]^T$, $\varepsilon = [\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n)]^T$. Consider predicting $Y(\mathbf{x})$ at some new point $\mathbf{x}$ by the linear predictor $\hat{Y}(\mathbf{x}) = a^T(\mathbf{x})\mathbf{y} = a^T(\mathbf{x})(F\beta + Z + \varepsilon) = a^T(\mathbf{x})(\mu\mathbf{1} + Z + \varepsilon)$. Then, the Best Linear Unbiased Predictor (BLUP) $\hat{Y}(\mathbf{x})$ will satisfy:

$$E\left(\hat{Y}(\mathbf{x})\right) = E\left(Y(\mathbf{x})\right)$$

$$\min_{a(\mathbf{x})} E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right]$$

Then,

$$E\left(\hat{Y}(\mathbf{x})\right) = E\left(a^T(\mathbf{x})y\right) = a^T(\mathbf{x})E(y) = a^T(\mathbf{x})F\beta = a^T(\mathbf{x})\mu\mathbf{1}$$

and

$$\begin{aligned} E\left(Y(\mathbf{x})\right) &= E(f^T(\mathbf{x})\beta + Z(\mathbf{x}) + \varepsilon(\mathbf{x})) = E(f^T(\mathbf{x})\beta) + 0 + 0 = \mu \\ &\Rightarrow a^T(\mathbf{x})F\beta = \beta \\ &\Rightarrow a^T(\mathbf{x})1 = 1 \\ &\Rightarrow a^T(\mathbf{x})1 - 1 = 0 \end{aligned}$$

which gives our constraint, and,

$$\begin{aligned} E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right] &= E\left[\left(f^T(\mathbf{x})\beta + Z(\mathbf{x}) + \varepsilon(\mathbf{x}) - a^T(\mathbf{x})(F\beta + Z + \varepsilon)\right)^2\right] \\ &= E\left\{\left[(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F)\beta\right]^2\right\} \\ &+ 2E\left\{(f^T(\mathbf{x}) - a^T(\mathbf{x})F)\beta(Z(\mathbf{x}) + \varepsilon(\mathbf{x})\right\} \\ &- E\left\{a^T(\mathbf{x})(Z + \varepsilon)) + (Z(\mathbf{x}) + \varepsilon(\mathbf{x}) - a^T(\mathbf{x})(Z + \varepsilon))^2\right\} \\ &= \left[\left(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F\right)\beta\right]^2 + E\left[(Z(\mathbf{x}) + \varepsilon(\mathbf{x}))^2\right] \\ &- 2E\left[(Z(\mathbf{x}) + \varepsilon(\mathbf{x}))(a^T(\mathbf{x})Z + a^T(\mathbf{x})\varepsilon)\right] \\ &+ E\left[a^T(\mathbf{x})(Z + \varepsilon)(Z + \varepsilon)^T a(\mathbf{x})\right] \\ &= \left[\left(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F\right)\beta\right]^2 + E(Z(\mathbf{x})Z^T(\mathbf{x})) + 2E(Z(\mathbf{x})\varepsilon(\mathbf{x})) \\ &+ E(\varepsilon(\mathbf{x})\varepsilon^T(\mathbf{x})) - 2a^T(\mathbf{x})E(Z(\mathbf{x})Z) - 2a^T(\mathbf{x})E(\varepsilon Z(\mathbf{x})) \\ &- 2a^T(\mathbf{x})E(\varepsilon(\mathbf{x})Z) - 2a^T(\mathbf{x})E(\varepsilon(\mathbf{x})\varepsilon) + a^T(\mathbf{x})E(ZZ^T)a(\mathbf{x}) \\ &+ 2a^T(\mathbf{x})E(Z\varepsilon^T)a(\mathbf{x}) + a^T(\mathbf{x})E(\varepsilon\varepsilon^T)a(\mathbf{x}) \\ &= \left[\left(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F\right)\beta\right]^2 + Var(Z(\mathbf{x})) + 2Cov(Z(\mathbf{x}), \varepsilon(\mathbf{x})) \\ &+ Var(\varepsilon(\mathbf{x})) - 2a^T(\mathbf{x})Cov(Z(\mathbf{x}), Z) - 2a^T(\mathbf{x})Cov(\varepsilon(\mathbf{x}), Z) \\ &- 2a^T(\mathbf{x})Cov(\varepsilon(\mathbf{x}), \varepsilon) + a^T(\mathbf{x})Cov(Z)a(\mathbf{x}) + 2a^T(\mathbf{x})Cov(Z, \varepsilon) \\ &+ a^T(\mathbf{x})Cov(\varepsilon)a(\mathbf{x}) \\ &= \left[\left(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F\right)\beta\right]^2 + Var(Z(\mathbf{x})) + Var(\varepsilon(\mathbf{x})) \\ &- 2a^T(\mathbf{x})Cov(Z(\mathbf{x}), Z) + a^T(\mathbf{x})Cov(Z)a(\mathbf{x}) \\ &+ a^T(\mathbf{x})Cov(\varepsilon)a(\mathbf{x}), \end{aligned}$$

where $Var(Z(\mathbf{x})) = \sigma_z^2$, $Cov(Z) = \sigma_z^2 R = \Sigma$, $Cov(Z(\mathbf{x}), Z) = \sigma_z^2 r(\mathbf{x})$, $Cov(\varepsilon) = \sigma_\varepsilon^2 I$, $Var(\varepsilon(\mathbf{x})) = \sigma_\varepsilon^2$, and $Cov(Z(\mathbf{x}), \varepsilon(\mathbf{x})) = Cov(\varepsilon, Z(\mathbf{x})) = Cov(\varepsilon(\mathbf{x}), Z) = Cov(\varepsilon(\mathbf{x}), \varepsilon) = Cov(Z, \varepsilon) = 0$. Since, under our constraint,

$$\left[\left(f^T(\mathbf{x})\beta - a^T(\mathbf{x})F\right)\beta\right]^2 = 0,$$

we then have:

$$E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right] = \sigma_z^2\left(1 - 2a^T(\mathbf{x})r(\mathbf{x}) + a^T(\mathbf{x})Ra(\mathbf{x})\right) + \sigma_\varepsilon^2\left(1 + a^T(\mathbf{x})Ia(\mathbf{x})\right)$$

We can then minimize our expected mean square error subject to the constraint of unbiasedness using the method of LaGrange multipliers, ie,

$$\min_{a(\mathbf{x})} E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right]$$

subject to

$$E\left(\hat{Y}(\mathbf{x})\right) = E\left(Y(\mathbf{x})\right)$$

which is equivalent to

$$\min_{a(\mathbf{x})} \sigma_z^2\left(1 - 2a^T(\mathbf{x})r(\mathbf{x}) + a^T(\mathbf{x})Ra(\mathbf{x})\right) + \sigma_\varepsilon^2\left(1 + a^T(\mathbf{x})Ia(\mathbf{x})\right)$$

subject to

$$a^T(\mathbf{x})1 - 1 = 0.$$

Solving using LaGrange multipliers,

$$
\begin{aligned}
&\frac{\partial}{\partial a(\mathbf{x})}\left(\sigma_z^2\left(1-2a^T(\mathbf{x})r(\mathbf{x})+a^T(\mathbf{x})Ra(\mathbf{x})\right)+\sigma_\varepsilon^2\left(1+a^T(\mathbf{x})Ia(\mathbf{x})\right)\right)\\
&=\lambda\frac{\partial}{\partial a(\mathbf{x})}\left(a^T(\mathbf{x})\mathbf{1}-1\right)\\
&\Rightarrow\frac{\partial}{\partial a(\mathbf{x})}\sigma_z^2\left(1-2a^T(\mathbf{x})r(\mathbf{x})+a^T(\mathbf{x})Ra(\mathbf{x})\right)+\sigma_\varepsilon^2\left(1+a^T(\mathbf{x})Ia(\mathbf{x})\right)\\
&-\lambda\left(a^T(\mathbf{x})\mathbf{1}-1\right)=0\\
&\Rightarrow -2\sigma_z^2r(\mathbf{x})+2\sigma_z^2Ra(\mathbf{x})+2\sigma_\varepsilon^2Ia(\mathbf{x})-\lambda\mathbf{1}^T=0\\
&\Rightarrow\lambda\mathbf{1}^T=-2\sigma_z^2r(\mathbf{x})+2\sigma_z^2Ra(\mathbf{x})+2\sigma_\varepsilon^2Ia(\mathbf{x})\\
&\Rightarrow\lambda'\mathbf{1}^T=\sigma_z^2r(\mathbf{x})-\sigma_z^2Ra(\mathbf{x})+\sigma_\varepsilon^2Ia(\mathbf{x})\quad(*)\\
&\Rightarrow\lambda'\mathbf{1}^T=\sigma_z^2r(\mathbf{x})-a^T(\mathbf{x})\tilde{\Sigma}\\
&\Rightarrow\lambda'\mathbf{1}^T\tilde{\Sigma}\mathbf{1}=\sigma_z^2r(\mathbf{x})\tilde{\Sigma}\mathbf{1}-a^T(\mathbf{x})\mathbf{1}\\
&\Rightarrow\lambda'\mathbf{1}^T\tilde{\Sigma}\mathbf{1}=\sigma_z^2r(\mathbf{x})\tilde{\Sigma}\mathbf{1}-1\\
&\Rightarrow\lambda'=\left(\sigma_z^2r(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}-1\right)\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}\\
&\Rightarrow\lambda'=-\left(1-\sigma_z^2r(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\right)\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}
\end{aligned}
$$

and plugging into $(*)$, we solve for $a(\mathbf{x})$:

$$
\begin{aligned}
\lambda'\mathbf{1}^T&=\sigma_z^2r(\mathbf{x})-\sigma_z^2Ra(\mathbf{x})-\sigma_\varepsilon^2Ia(\mathbf{x})\\
-\left(1-\sigma_z^2r(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\right)\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}&=\sigma_z^2r(\mathbf{x})-\tilde{\Sigma}a(\mathbf{x})\\
a(\mathbf{x})&=\sigma_z^2r(\mathbf{x})\tilde{\Sigma}^{-1}\\
&\quad+\left(1-\sigma_z^2r(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\right)\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}\tilde{\Sigma}^{-1}
\end{aligned}
$$

Plugging this result back into our expected mean square error, we can (after some algebraic manipulation) arrive at:

$$
\begin{aligned}
E\left[\left(Y(\mathbf{x})-\hat{Y}(\mathbf{x})\right)^2\right]&=\sigma_z^2-\left(\sigma_z^2\right)^2r^T(\mathbf{x})\tilde{\Sigma}^{-1}r(\mathbf{x})\\
&\quad+\left(1-\sigma_z^2r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\right)^2\left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)^{-1}+\sigma_\varepsilon^2
\end{aligned}
$$

which, when $\sigma_\varepsilon^2 = 0$ simplifies to:

$$E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right] = \sigma_z^2 - \left(\sigma_z^2\right)^2 r^T(\mathbf{x})\Sigma^{-1}r(\mathbf{x}) + \left(1 - \sigma^2 r^T(\mathbf{x})\Sigma^{-1}\mathbf{1}\right)^2 \left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}$$

If we let $\sigma_\tau^2 = \frac{\sigma_\varepsilon^2}{\sigma^2}$, then factoring out σ_z^2, we will have $\tilde{\Sigma} = \sigma_z^2\check{\Sigma} = \sigma_z^2(R + \sigma_\tau^2 I)$, and so our MSE can now be written as:

$$E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right] = \sigma_z^2\left(1 - r^T(\mathbf{x})\check{\Sigma}^{-1}r(\mathbf{x}) + \left(1 - r^T(\mathbf{x})\check{\Sigma}^{-1}\mathbf{1}\right)^2 \left(\mathbf{1}^T\check{\Sigma}^{-1}\mathbf{1}\right)^{-1} + \sigma_\tau^2\right)$$

In a similar fashion, we can solve for the BLUP:

$$\begin{aligned}
\hat{Y}(\mathbf{x}) &= a^T(\mathbf{x})\mathbf{y} \\
&= \sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{y} + \left(\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\right)\mathbf{1}^T\tilde{\Sigma}^{-1}y - \sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\left(1^T\tilde{\Sigma}^{-1}1\right)^{-1}\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{y} \\
&= \sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{y} + \hat{\mu} - \sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\mathbf{1}\hat{\mu} \\
&= \sigma_z^2 r^T(\mathbf{x})\tilde{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{1}\hat{\mu}\right) + \hat{\mu}
\end{aligned}$$

and again by factoring out σ_z^2, we arrive at

$$\hat{Y}(\mathbf{x}) = r^T(\mathbf{x})\check{\Sigma}^{-1}\left(y - \mathbf{1}\hat{\mu}\right) + \hat{\mu}$$

Appendix C

IMSE Design Formulations

When searching for designs, we typically deal with the σ_z^2-normalized version of the expected mean square error, hence

$$\begin{aligned}\frac{IMSE}{\sigma_z^2} &= \int_\Omega \frac{E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right]}{\sigma_z^2} d\mathbf{x} \\ &= \int_\Omega 1 - r^T(\mathbf{x})\check{\Sigma}^{-1}r(\mathbf{x}) + \left(1 - r^T(\mathbf{x})\check{\Sigma}^{-1}\mathbf{1}\right)^2 \left(\mathbf{1}^T\check{\Sigma}^{-1}\mathbf{1}\right)^{-1} + \sigma_\tau^2 \, d\mathbf{x}\end{aligned}$$

or, in the deterministic case ($\sigma_\varepsilon^2 = 0$), this reduces simply to

$$\begin{aligned}\frac{IMSE}{\sigma_z^2} &= \int_\Omega \frac{E\left[\left(Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\right)^2\right]}{\sigma_z^2} \\ &= \int_\Omega 1 - r^T(\mathbf{x})R^{-1}r(\mathbf{x}) + \left(1 - r^T(\mathbf{x})R^{-1}\mathbf{1}\right)^2 \left(\mathbf{1}^T R^{-1}\mathbf{1}\right)^{-1} \, d\mathbf{x}\end{aligned}$$

It can be shown (Sacks et al., 1989a; Sacks et al., 1989b) that this can be expressed in the much nicer form below:

$$\begin{aligned}\frac{IMSE}{\sigma_z^2} &= \int_\Omega \left[1 - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma} \end{pmatrix}^{-1} \begin{pmatrix} 1 & r^T(\mathbf{x}) \\ r(\mathbf{x}) & r(\mathbf{x})r^T(\mathbf{x}) \end{pmatrix}\right] d\mathbf{x}\right] \\ &= \int_\Omega 1 d\mathbf{x} - trace\left[\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \check{\Sigma} \end{pmatrix}^{-1} \begin{pmatrix} \int_\Omega 1 d\mathbf{x} & \int_\Omega r^T(\mathbf{x}) d\mathbf{x} \\ \int_\Omega r(\mathbf{x}) d\mathbf{x} & \int_\Omega r(\mathbf{x})r^T(\mathbf{x}) d\mathbf{x} \end{pmatrix}\right]\end{aligned}$$

which we then alter accordingly for the two models discussed in this dissertation.

Bibliography

Bengio, Y., Paiement, J., and Vincent, P.: 2003, *Out-of-sample extensions for LLE, ISOMAP, MDS, eigenmaps and spectral clustering*, Technical report, Unversity de Montreal

Box, G. and Draper, N.: 1987, *Empirical model-building and response surface*, Wiley, New York

Chen, T., Ciscon, K., and Ratkus, A.: 1984, in *Second symposium on taguchi methods*, pp 70–77, American Supplier Institute, Romulus, Michigan

Chipman, H.: 1997, *Fast model search for designed experiments with complex aliasing*, Technical report, University of Waterloo

Cormen, T., Leiserson, C., Rivest, R., and Stein, C.: 1990, *Introduction to algorithms, 2nd edition*, McGraw-Hill

Cox, T. and Cox, M.: 2001, *Multidimensional scaling, 2nd edition*, Chapman and Hall/CRC, Boca Raton

Donoho, D. and Grimes, C.: 2003, *Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data*, Technical report, Stanford University

Hamada, M. and Wu, C.: 1995, *Journal of Quality Technology* **27(1)**, 45

Hedayat, A., Sloane, N., and Stufken, J.: 1999, *Orthogonal arrays: theory and applications*, Springer-Verlag New York, Inc., 175 Fith Avenue, New York, NY 10010, USA

Johnson, M., Moore, L., and Ylvisaker, D.: 1990, *Journal of Statistical Planning and Inference* **26**, 131

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K.: 2005, *Variable selection for gaussian process models in computer experiments*, to appear

Loland, A. and Host, G.: 2003, *Environmetrics* **14**, 307

Merz, C.: 2001, in *MTS/IEEE Oceans*

Muller, W.: 2001, *Collecting spatial data: optimum design of experiments for random fields, 2nd edition*, Physica-Verlag

Press, W., Vetterling, W., Teukolsky, S., and Flannery, B.: 1992, *Numerical recipes in C: The art of scientific computing*, Cambridge University Press

Rathbun, S.: 1998, *Environmetrics* **9**, 109

Ravishanker, N. and Dey, D.: 2002, *A first course in linear model theory*, Chapman and Hall/CRC, Boca Raton

Roweis, S. and Saul, L.: 2000, *Science* **290**, 2323

Sacks, J., Schiller, S., and Welch, W.: 1989a, *Technometrics* **31(1)**, 41

Sacks, J., Welch, W., Mitchell, T., and Wynn, H.: 1989b, *Statistical Science* **4(4)**, 409

Silvey, S.: 1980, *Optimal design: an introduction to the theory for parameter estimation*, Chapman and Hall

Taguchi, G.: 1987, *System of experimental design*, Unipub/Kraus International Publications, White Plains, NY

Tenenbaum, J., de Silva, V., and Langford, J.: 2000, *Science* **290**, 2319

Weisberg, R., He, R., Luther, M., Walsh, J., Cole, R., Donovan, J., Merz, C., and Subramanian, V.: 2002, in *MTS/IEEE Oceans*

Wolfinger, R., Tobias, R., and Sall, J.: 1994, *Journal of Scientific Computing* **15(6)**, 1294