PREDICTION AND DIMENSION REDUCTION METHODS IN COMPUTER EXPERIMENTS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Minyong R. Lee
June 2017
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Art Owen, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Paul Switzer

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jonathan Taylor

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

In many fields of engineering and science, computer experiments have become essential tools in studying physical processes. This dissertation reviews standard prediction methods and dimension reduction methods in the analysis of computer experiments and proposes new approaches.

Response surface modeling is the starting point of the analysis of computer experiments. Kriging or Gaussian process regression is widely used in constructing response surfaces. We propose Single Nugget Kriging, which is a method with better predictions at extreme values than the standard method of Kriging. Our prediction exhibits robustness to the model mismatch in the covariance parameters, a desirable feature for computer simulations with a restricted number of data points.

For high dimensional computer experiments, dimension reduction methods in regression are essential for solving optimization problems and inverse problems. We compare model-free sufficient dimension reduction methods and the active subspace for computer experiments. We propose a modification of the active subspace. We further discuss the analysis of dimension reduction methods in computer experiments, using projected Gaussian processes.
Acknowledgements

This dissertation could not have been completed without the support of many people over the years.

First and foremost, I would like to show my deepest gratitude to my advisor, Art B. Owen, for his guidance and support. His diverse academic interests and vast knowledge inspired me and guided me in developing my research. I am truly grateful for his mentorship and invaluable support over the years.

I am grateful to my dissertation committee members, Jonathan Taylor, Paul Switzer, Jef Caers, and Gianluca Iaccarino for their guidance and constructive criticism.

During my years at Stanford, I have been very fortunate to meet great people. I would like to thank all the members and visitors of Owen’s group, for the discussions and interactions we had. Also, I am thankful to my cohort of graduate students in the Statistics department, for being great friends and sharing unforgettable memories.

I would also like to thank my parents, brother, and in-laws for their emotional support.

Finally, I am indebted to my wife, Seung Ah Lee. Her support, encouragement, patience, and love are undeniably irreplaceable.
# Contents

Abstract iv

Acknowledgements v

1 Introduction 1

1.1 Introduction to computer experiments 1

1.2 Examples of computer experiments 3

1.2.1 Borehole function 3

1.2.2 ONERA-M6 6

2 Prediction in computer experiments 8

2.1 Kriging 9

2.2 Single Nugget Kriging 14

2.2.1 Conditional likelihood at the target and conditional bias 15

2.2.2 Single Nugget Kriging 18

2.2.3 Numerical experiments 27

2.2.4 Discussion 31

2.2.5 Proofs 32

2.2.6 Test functions 37

2.3 Universal Single Nugget Kriging 37

2.3.1 Advantages and disadvantages of universal Kriging 38

2.3.2 Universal Single Nugget Kriging 39

2.3.3 An illustrative example 40

2.3.4 Discussion 42
3 Dimension reduction in computer experiments 43

3.1 Sufficient dimension reduction methods 46
  3.1.1 Sliced Inverse Regression 47
  3.1.2 Sliced Average Variance Estimation 49
  3.1.3 Principal Hessian Directions 50
  3.1.4 Contour Regression 51
  3.1.5 Directional Regression 52
  3.1.6 More sufficient dimension reduction methods 52
  3.1.7 Discussion 55

3.2 Active subspaces 56

3.3 Comparison of dimension reduction methods for computer experiments 58
  3.3.1 Sufficient dimension reduction methods for deterministic functions 58
  3.3.2 Ridge functions 59
  3.3.3 Numerical examples 60
  3.3.4 Discussion 63

3.4 Modified active subspaces 65
  3.4.1 Introduction 66
  3.4.2 Modified active subspaces 67
  3.4.3 Criteria for dimension reduction 68
  3.4.4 Quadratic functions 70
  3.4.5 Numerical examples 73
  3.4.6 Discussion 79
  3.4.7 Test functions 79

3.5 Analysis of dimension reduction in Gaussian process regression 80
  3.5.1 Assumption about the Gaussian process 82
  3.5.2 Gaussian process projected on the dimensionally reduced space 83
  3.5.3 Estimation of ξ 88
  3.5.4 Extensions to general Matérn kernels 89
  3.5.5 Example: borehole function 91
  3.5.6 Discussion 93
List of Tables

1.1 Variables in borehole function. ................................................. 4

2.1 Performance comparison of ordinary Kriging and SiNK for a realization of Gaussian process and piston function. The EISE ratios are the EISE of SiNK divided by the EISE of ordinary Kriging. .......................... 28

2.2 Input variables $x$ for the piston function. ................................. 29

2.3 Performance comparison among ordinary Kriging, limit Kriging and SiNK. The average and standard deviation of metrics from 100 independent set of inputs are reported. The EISE ratios are the EISE of SiNK or limit Kriging divided by the EISE of ordinary Kriging. The Friedman function did not have extreme values for most of the simulations. ................................................................. 31

2.4 $R^2$ of the predictors in fitting the example function (2.23) with 4 function evaluations ................................................................. 42

3.1 Comparison of sufficient dimension reduction methods. The mean and standard deviation of the distance from the truth are computed, from 400 samples of inputs. ................................................................. 62

3.2 Example of a quadratic function in (3.18) ($p = 2, q = 1$). The quantities $VE_{Q_1}$ and $EV_{Q_2}$ are normalized by the variance of the function $V$. The goal is to maximize $VE_{Q_1}$ or $EV_{Q_2}$. ................................. 75
3.3 Example of an eight-dimensional function in (3.19). The mean and the standard deviation of normalized $\hat{EV}_{Q_1}$ and $\hat{EV}_{Q_2}$ by the estimated variance $V$ over 100 sets of inputs are reported, where $n = 200$. The goals are to minimize $EV_{Q_1}$ or maximize $EV_{Q_2}$.

3.4 The mean and the standard deviation of normalized $\hat{EV}_{Q_1}$ and $\hat{EV}_{Q_2}$ by the estimated variance $V$ are reported, over 100 sets of inputs, where $n = 200$. The goal is to minimize $EV_{Q_1}$ or maximize $EV_{Q_2}$.

3.5 Summary of the decomposition of $f(x) | Q_1 \top x$.

3.6 Estimated length-scale hyperparameters $\ell$ in fitting Gaussian process regression to the borehole function, using 1000 function evaluations.

3.7 Proportional variance explained in the Gaussian process by variable selection in borehole function. $\hat{\xi}$ is the plug-in estimate from the original data, and $\tilde{\xi}$ is the heuristic estimate from the dimensionally reduced data.
# List of Figures

1.1 Borehole scheme, Figure 2-2 in Harper and Gupta (1983) ............................... 4
1.2 ONERA-M6, figure in Schmitt and Charpin (1979) ............................... 6

2.1 Two-dimensional Zakharov function. The gray surface is the true function, four points on the surface are the observations. We compare the SiNK predictions and ordinary Kriging predictions on the dotted line $x_1 = x_2$ in Figure 2.2. ................................................................. 23
2.2 Illustration of the difference between ordinary Kriging and SiNK. SiNK performs better at extreme values than ordinary Kriging, more significantly when the correlations between function values are smaller. .......... 24
2.3 Relation between $\rho(x_0)$ and z-score. ......................................................... 26
2.4 Ordinary Kriging and SiNK for a realization of 7-dimensional Gaussian process. Rank is the order of the true function values of the test points. 28
2.5 Ordinary Kriging and SiNK for the piston function. Rank is the order of the true function values of the test points. ................................. 30
2.6 Example function (2.23) ................................................................. 41
2.7 Example function (2.23). The circle dots are the observations and the gray lines are the estimated trends in both models. ....................... 41

3.1 $\log(\text{dist}(Q_1, \hat{Q}_1))$ for directional regression on example function $\phi_2$, over the log of the number of samples. The distances are averaged over 25 sets of samples. ................................. 64
3.2 $\text{dist}(Q_1, \hat{Q}_1)$ for directional regression on example function $\phi_2$, over the number of samples. The distances are averaged over 25 sets of samples. 65
3.3 Contour plot of the quadratic function in (3.18). The black solid arrow is \( \arg \max_{Q_1} V E_{Q_1} \). The red solid arrow, which matches the black solid arrow, is the modified active subspace. The green dashed arrow is the orthogonal complement of \( \arg \max_{Q_2} E V_{Q_2} \). The blue dotted arrow is the active subspace.

3.4 Illustration of the variance of the Gaussian process at an input. Gray lines are samples of Gaussian processes. Gaussian process regression treats the function of interest as a realization of a Gaussian process, which is the black line in the plot. The red arrows illustrate the variance of the Gaussian process.

3.5 Estimates of \( \xi \) in Table 3.7. The solid lines are \( \hat{\xi} \), and the dashed lines are \( \tilde{\xi} \).
Chapter 1

Introduction

1.1 Introduction to computer experiments

Uncertainty Quantification (UQ) is the process of inferring the unknown quantities from data. The term ‘uncertainty quantification’ usually refers to applying statistical methods to data from natural sciences and engineering.

Computer experiments have become an essential part in uncertainty quantification for studying physical processes such as the subsurface of the earth, aerodynamic forces on bridge decks, and channel network flow. These experiments can be thought of as functions: given a set of input variables in a fixed domain, the computer experiment returns the output, which can be a single value, a vector, or even another function.

\[
\text{Input} \xrightarrow{f(\cdot)} \text{Output}
\]

The objectives of computer experiments vary across the applications. Typical goals include:

1. Prediction: make predictions and construct prediction intervals at unobserved inputs.

2. Inverse problem: find set of the inputs for a target response.
3. Optimization: find the optimizing input and the global optimum.

4. Sensitivity analysis: quantify the impact in the response from a certain change in an input variable.

In general, the process of computer simulation consists of many steps including deterministic functions and stochastic processes. For example, reservoir modelings in geosciences are heavily based on stochastic simulations. A reservoir model is a 3D model of the porous media and the flow of fluids and gasses in the media. Reservoir models are constructed based on various data sources, such as seismic data, logging data from wells, etc. The models are produced in numerous ways, for example by object models following physical laws. Practitioners quantify the uncertainty from multiple output images of the 3D models. Still, many deterministic functions are also used in the simulation process of the fluids, and it is important to understand the functions.

The field of computer experiments usually focuses on the deterministic part of the computer simulation. If we run the experiment with the same set of input variables, the output is identical. When there is noise, such as measurement error, usually they are relatively very small compared to the deterministic part of the output. In many cases, gathering the data for computer experiments are expensive concerning cost and time, so the number of function evaluations can be very limited.

Because of the limited resources to get data, the design of experiments is important. The traditional ideas in the design of experiments, such as randomization and blocking, may not be appropriate because the functions are deterministic, and can have complex interactions between the variables. Thus, designing in computer experiments needs meticulous approaches. Space-filling designs, such as Latin Hypercube Designs and low discrepancy sequences, are widely used for general purposes. The quality of design depends greatly on the objective of the analysis. For example, IMSE (Integrated Mean Squared Error)-optimal and MMSE (Maximum Mean Squared Error)-optimal are used for minimizing the mean squared errors. Extensive numbers of designs have been developed and studied; see Pronzato and Müller (2012).

The analysis of the data from computer experiments has also been developed
with the demand of uncertainty quantification. Response surface modeling, which is modeling an approximation function of the black-box function, plays a crucial role in data analysis. Response surfaces are also called surrogates, meta-models, and emulators. Interpolation methods, such as polynomial methods, may provide good predictions but do not provide the uncertainty of the unknown part of the function. Classical regression methods yield prediction intervals, but usually lack the accuracy in their predictions. By considering the black-box function as a realization of a stochastic process, statistical inference is applicable. Gaussian process regression is a standard method to analyze computer experiments, which we discuss in detail in Chapter 2.

A natural question arises when we model the functions by stochastic processes. What is stochastic in the function? The function itself is deterministic with no or very small noise, so it is not random. However, we conveniently model the function as a sample of a stochastic process to quantify the uncertainty. Uncertainty comes from the insufficiency of our information, and stochastic processes are tools that help us understand how much we know and we do not know about the function.

For more discussions of problems and examples in computer experiments, see Sacks et al. (1989) and Koehler and Owen (1996).

1.2 Examples of computer experiments

In this section, we introduce two examples of computer experiments.

1.2.1 Borehole function

The borehole model by Harper and Gupta (1983) is commonly used for testing a variety of analysis methods in computer experiments. The model describes the flow through a borehole, drilled through an aquifer above a nuclear waste repository, through the repository, to an aquifer below. Figure 1.1 provides a schematic view of the scenario. It is assumed that the potentiometric head is higher in the upper aquifer than in the lower, which results in a downward flow. Also, it is assumed that
there is no regional ground-water gradient.

The quantity of interest is the flow rate, $Q (m^3/yr)$. The following table shows the physical variables in the model and their range. In Harper and Gupta (1983), the ranges for $r_w$ and $r$ are also given as $[0.05, 0.15]$ and $[100, 50000]$ respectively.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>type</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_u$ (m²/yr)</td>
<td>transmissivity of upper aquifer</td>
<td>input</td>
<td>Unif [63070, 111600]</td>
</tr>
<tr>
<td>$T_l$ (m²/yr)</td>
<td>transmissivity of lower aquifer</td>
<td>input</td>
<td>Unif [63.1, 116]</td>
</tr>
<tr>
<td>$H_u$ (m)</td>
<td>potentiometric head of upper aquifer</td>
<td>input</td>
<td>Unif [990, 1110]</td>
</tr>
<tr>
<td>$H_l$ (m)</td>
<td>potentiometric head of lower aquifer</td>
<td>input</td>
<td>Unif [700, 820]</td>
</tr>
<tr>
<td>$r$ (m)</td>
<td>radius of influence</td>
<td>input</td>
<td>Lognormal (7.71, 1.0056)</td>
</tr>
<tr>
<td>$r_w$ (m)</td>
<td>radius of borehole</td>
<td>input</td>
<td>Normal (0.10, 0.0161812)</td>
</tr>
<tr>
<td>$K_w$ (m/yr)</td>
<td>hydraulic conductivity of borehole</td>
<td>input</td>
<td>Unif [9855, 12045]</td>
</tr>
<tr>
<td>$L$ (m)</td>
<td>length of borehole</td>
<td>input</td>
<td>Unif [1120, 1680]</td>
</tr>
<tr>
<td>$H_{uw}$ (m)</td>
<td>steady-state potentiometric head in borehole at upper aquifer</td>
<td>dependent</td>
<td></td>
</tr>
<tr>
<td>$H_{wl}$ (m)</td>
<td>steady-state potentiometric head in borehole at lower aquifer</td>
<td>dependent</td>
<td></td>
</tr>
</tbody>
</table>

From the Thiem solution, the steady-state flow through the upper aquifer to the
borehole in an aquifer system can be written as

\[ Q = 2\pi T_u \frac{H_u - H_{wu}}{\ln(r/r_w)}. \]  (1.1)

Similarly, the steady-state flow from the borehole to the lower aquifer can be written as

\[ Q = -2\pi T_l \frac{H_l - H_{wl}}{\ln(r/r_w)}. \]  (1.2)

Using Darcy’s equation, the steady-state, laminar, and isothermal flow of a homogeneous fluid through the borehole can be written as

\[ Q = \pi r_w^2 K_w \frac{H_{wu} - H_{wl}}{L}. \]  (1.3)

From (1.1), (1.2), and (1.3), the variables \( Q, H_{wu}, \) and \( H_{wl} \) depend on the input variables in Table 1.1. Solving (1.1) and (1.2) in terms of \( H_{wu} \) and \( H_{wl} \) leads to

\[ H_{wu} = H_u - \frac{Q \ln(r/r_w)}{2\pi T_u} \]  \text{and}  \quad (1.4)

\[ H_{wl} = H_l + \frac{Q \ln(r/r_w)}{2\pi T_l}. \]  (1.5)

Plugging these in (1.3) leads to

\[ Q = \frac{2\pi T_u (H_u - H_l)}{\log(r/r_w) \left(1.5 + \frac{2LT_u}{\log(r/r_w) r_w^2 K_w} + \frac{T_u}{T_l}\right)}. \]  (1.6)

Thus, the flow is an eight-dimensional function with inputs described in Table 1.1.

This function is not an example of computationally expensive high-dimensional computer experiments. However, the simplicity of the borehole function allows us to evaluate various analysis methods in computer experiments.
1.2.2 ONERA-M6

The ONERA-M6 wing is a classic computational fluid dynamics (CFD) validation case for external flows because of its simple geometry, complicated flow physics, and availability of experimental data. The wind tunnel tests are documented by Schmitt and Charpin in the AGARD Report AR-138 published in Schmitt and Charpin (1979). Figure 1.2 shows the design of the wing.

CFD codes such as SU2, an open-source simulation suite developed in the Aerospace Design Laboratory at Stanford University, provides solvers to simulate the flow past the wing at transonic speed in air.

![Figure 1.2: ONERA-M6, figure in Schmitt and Charpin (1979)](image)

A major goal in aerospace engineering is constrained shape optimization. A famous case is lift-constrained drag minimization of the ONERA-M6 wing. The shape optimization exercise has three main tasks; free-form deformation (FFD) of the surface, solving a CFD problem, and optimization.

The FFD parameterization leads to a set of design points, which can be controlled independently in the domain. The number of design variables depends on the deformation, and reasonable ranges are from ten to several hundred. For example, see Chauhan et al. (2010) and Lukaczyk et al. (2014).

For more complex problems in engineering, the number of dimensions is larger and the computer simulation is computationally more challenging. Thus, the aerospace engineering community has been interested in dimension reduction methods in the
CHAPTER 1. INTRODUCTION

analysis of computer experiments.
Chapter 2

Prediction in computer experiments

Fitting a regression model or a response surface to data from computer experiments is usually the first step of the analysis. Regression provides prediction and prediction intervals under certain assumptions. The regression fit can also be used for estimating a functional of the function, such as the maximum of the function and the integral of the function over a certain input distribution.

Standard parametric models in regression, such as generalized linear models, are often not appropriate in constructing response surfaces due to the high nonlinearity of the functions in computer experiments. Decision-tree based methods do not adequately represent the smoothness of the functions. Kriging, or Gaussian process regression, is a popular way to build metamodels in computer experiments. The method was initially proposed in Krige (1951), and improved by Matheron (1963). Kriging exactly interpolates the experimental data and produces predictions at unobserved inputs. The method also generates credible intervals which represent the uncertainty of the prediction. Stein (1999) and Switzer (2006) give summaries and in-depth discussions of Kriging.

This chapter is structured as follows: in Section 2.1, we review Kriging and Gaussian process regression and discuss their advantages and limitations. In Section 2.2, we propose a new prediction method which we call Single Nugget Kriging (SiNK).
We show that SiNK has several desirable properties with the same model complexity and computational cost as standard Kriging. In Section 2.3, we propose Universal Single Nugget Kriging that shares the benefits of SiNK and universal Kriging.

2.1 Kriging

Kriging treats the deterministic function \( f(x) \) as a realization of a real-valued random field

\[
Y(x) = m(x) + Z(x)
\]

where \( x \in \mathbb{R}^p \), \( m(x) \) is a deterministic mean function, and \( Z(x) \) is centered (mean zero) square-integrable process with covariance function \( K(\cdot, \cdot) \).

There are three widely used Kriging models based on the mean function. When the mean function is a known function, it is called \textit{simple Kriging}, and when the function is an unknown constant \( \beta \), it is called \textit{ordinary Kriging}. When the mean function is a linear combination of known functions \( f_0, \ldots, f_r \) but coefficients \( \beta_0, \ldots, \beta_r \) are unknown, namely \( m(x) = \sum_{s=0}^r \beta_s f_s(x) \), it is called \textit{universal Kriging}.

For the covariance function, stationary covariance functions that are tensor products of one-dimensional kernels are popular. Let \( C_\theta : \mathbb{R} \rightarrow [-1, 1] \) be a covariance kernel with length-scale parameter \( \theta \). Let

\[
K(x, z) = \sigma^2 C(h) = \sigma^2 \prod_{j=1}^p C_{\theta_j}(|h_j|) = \sigma^2 \prod_{j=1}^p C_1 \left( \frac{|h_j|}{\theta_j} \right)
\]

where \( h = x - z \) and \( \sigma^2 \) and \( (\theta_1, \ldots, \theta_p) \) are estimated from data. Matérn covariance kernels (Matérn, 1986) have the form

\[
C_{\nu,\theta}(d) = \frac{(\sqrt{2\nu d/\theta})^\nu}{\Gamma(\nu)(2\nu-1)} K_\nu \left( \sqrt{2\nu d/\theta} \right)
\]

where \( K_\nu(\cdot) \) is the modified Bessel function of the second kind. Matérn covariance kernels are commonly used in practice because of the smoothness of the associated Gaussian process, defined in terms of its mean square differentiability parametrized...
through $\nu$. Also, almost sure regularity of paths is determined by $\nu$; see Scheuerer (2009) for more details.

For high dimensional functions, isotropic covariances

$$K(x, z) = \sigma^2 C_\theta(\|h\|) = \sigma^2 C_1(\|h\|)$$

are often used, where $\| \cdot \|$ is the Euclidean norm. If there is a measurement error or noise in the function, then adding a nugget effect handles the discontinuity in the function, namely

$$K(x, z) = \sigma^2 C(h) = \sigma^2 \prod_{j=1}^p C_{\theta_j}(|h_j|) + \tau^2 \mathbb{I}_0(h)$$

where $\tau^2 > 0$ is a parameter and $\mathbb{I}_0$ is the indicator function of the set $\{0\} \subset \mathbb{R}^p$.

Now suppose we observe $y = (Y(x_1), \ldots, Y(x_n))$, and let $K = (K_{ij})$ be the $n \times n$ covariance matrix of $y$, $k(x_0, x_0)$ be the variance of $Y(x_0)$, and $k(x_0)$ be the covariance vector between $y$ and $Y(x_0)$. In a matrix form,

$$\text{Var} \begin{bmatrix} Y(x_0) \\ y \end{bmatrix} = \begin{bmatrix} k(x_0, x_0) & k(x_0)^T \\ k(x_0) & K \end{bmatrix}.$$

Let $1$ be the $n$-length vector of all ones. The simple Kriging predictor is the Best Linear Predictor (BLP) that minimizes the mean squared prediction error (MSPE). Specifically, the linear unbiased predictor $\hat{Y}(x_0) = \beta + \lambda^T (y - \beta 1)$ that minimizes

$$\mathbb{E}[(Y(x_0) - \hat{Y}(x_0))^2]$$

with respect to $\lambda$ is the simple Kriging predictor. Expanding the MSPE leads to

$$\mathbb{E}[(Y(x_0) - \hat{Y}(x_0))^2] = \mathbb{E}[(Y(x_0) - \beta + \lambda^T (y - \beta 1))^2]$$

$$= k(x_0, x_0) - 2\lambda^T k(x_0) + \lambda^T K \lambda.$$
Thus, the minimizing \( \lambda \) is \( K^{-1}k(\mathbf{x}_0) \), and the simple Kriging predictor is

\[
\hat{Y}_K(\mathbf{x}_0) = \beta + k(\mathbf{x}_0)^\top K^{-1}(\mathbf{y} - \beta 1).
\]

For ordinary Kriging and universal Kriging, we need to estimate the coefficients in the mean function. In the universal Kriging model

\[
Y(\mathbf{x}) = m(\mathbf{x}) + Z(\mathbf{x}) = \sum_{s=0}^{r} \beta_s f_s(\mathbf{x}) + Z(\mathbf{x}),
\]

let \( F \) be the \( n \times (r+1) \) matrix of basis function evaluations at \( \mathbf{x}_1, \ldots, \mathbf{x}_n \),

\[
F = \begin{pmatrix}
  f_0(\mathbf{x}_1) & \cdots & f_r(\mathbf{x}_1) \\
  \vdots & \ddots & \vdots \\
  f_0(\mathbf{x}_n) & \cdots & f_r(\mathbf{x}_n)
\end{pmatrix}.
\]

Let \( \mathbf{f}(\mathbf{x}_0) = (f_0(\mathbf{x}_0), \ldots, f_r(\mathbf{x}_0))^\top \). The universal Kriging predictor is the Best Linear Unbiased Predictor (BLUP) \( \lambda^\top \mathbf{y} \) that minimizes the MSPE with respect to \( \mathbb{E}[\hat{Y}(\mathbf{x}_0)] = \sum_{s=0}^{r} \beta_s f_s(\mathbf{x}_0) \). The unbiasedness constraint is equivalent to \( F^\top \lambda = \mathbf{f} \). The form of the MSPE is the same as the simple Kriging case:

\[
\mathbb{E}[(Y(\mathbf{x}_0) - \hat{Y}(\mathbf{x}_0))^2] = k(\mathbf{x}_0, \mathbf{x}_0) - 2\lambda^\top k(\mathbf{x}_0) + \lambda^\top K \lambda.
\]

We introduce the Lagrange multiplier

\[
g(\lambda, \gamma) = k(\mathbf{x}_0, \mathbf{x}_0) - 2\lambda^\top k(\mathbf{x}_0) + \lambda^\top K \lambda + \gamma^\top \{F^\top \lambda - \mathbf{f}(\mathbf{x}_0)\}
\]

where \( \gamma \) is a length \( r+1 \) vector. The system of equations from the Lagrange multiplier can be represented as

\[
\begin{pmatrix} 0 & F^\top \\ F & K \end{pmatrix} \begin{pmatrix} \gamma \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_0) \\ k(\mathbf{x}_0) \end{pmatrix}.
\]
Solving the first order condition is equivalent to solving

$$
\hat{\lambda}(x_0)^\top = f(x_0)^\top (F^\top K^{-1}F)^{-1}F^\top K^{-1} + k(x_0)^\top K^{-1} - k(x_0)^\top K^{-1}F(F^\top K^{-1}F)^{-1}F^\top K^{-1}.
$$

If we write

$$
\hat{\beta} = (F^\top K^{-1}F)^{-1}F^\top K^{-1}y, \quad (2.2)
$$

then the universal Kriging estimator can be written as

$$
\hat{Y}(x_0) = \hat{\lambda}^\top y = f^\top \hat{\beta} + k(x_0)^\top K^{-1}(y - F\hat{\beta}). \quad (2.3)
$$

The Kriging predictor can be also derived from the Gaussian process assumption on \(Y\); this approach is called Gaussian process regression. Throughout, we assume that \(Z\) in (2.1) is a centered Gaussian process with covariance function \(K(\cdot, \cdot)\). For simplicity, let us discuss the simple Kriging with known constant mean \(\beta\). Then,

$$
Y(x_0) \mid (Y(x_1), \ldots, Y(x_n)) = y \sim N(m, s^2),
$$

where

$$
m = \beta + k(x_0)^\top K^{-1}(y - \beta 1), \text{ and } \quad s^2 = k(x_0, x_0) - k(x_0)^\top K^{-1}k(x_0).
$$

That is, the conditional distribution of \(Y(x_0)\) given \(y\) is \(N(m, s^2)\). The simple Kriging predictor is the conditional mean

$$
\hat{Y}_K(x_0) = \mathbb{E}[Y(x_0) \mid y] = \beta + k(x_0)^\top K^{-1}(y - \beta 1).
$$

When the mean function is unknown in the Gaussian process regression, which is the universal Kriging case, the coefficients are estimated by maximizing the likelihood of \((Y(x_1), \ldots, Y(x_n))\). We can easily show that the maximum likelihood estimator
(MLE) of $\beta$ is $\hat{\beta} = (F^T K^{-1} F)^{-1} F^T K^{-1} y$ which matches (2.2) in universal Kriging. With

$$\rho(x_0) = \sqrt{\frac{k(x_0)^T K^{-1} k(x_0)}{k(x_0, x_0)}},$$

$\rho(x_0)^2$ is the variance explained by conditioning divided by the marginal variance of $y_0$. The quantity $\rho(x_0)$ always lies in $[0, 1]$, and can be understood as the correlation between the target function value and the data.

There are several limitations of Kriging. Kriging prediction depends on the covariance hyperparameters that are usually unknown, and need to be estimated. The variability of the predicted process depends on the hyperparameters, and the likelihood of the hyperparameters, usually computationally expensive to compute, can be quite flat, so that the variance of the maximum likelihood estimator can be huge. Stein (1999) gives asymptotic results showing that there is asymptotically no loss using the misspecified covariance function if the true and assumed covariance functions yield equivalent Gaussian measures. In practice, we might face functions that behave quite differently from the assumed covariance structure, and with only a small number of observations. Bachoc (2013) states that the fixed-domain asymptotics does not solve completely the issue of the estimation of the covariance function. There have been several approaches to stabilize the estimation of the hyperparameters, such as penalized Kriging by Li and Sudjianto (2005). We would like to find a predictor that is less affected by the hyperparameters.

Kriging prediction depends on the mean function whose specification needs form before looking at the data. In Kriging, there is a "regression effect", in which the predictions are pulled towards the mean function. This can give bad predictions at extreme function values. Using a better design of input points, such as Maximum Projection Designs (Joseph et al., 2015), can be effective. But, with a small number of observations, finding a good model can be challenging.

There are several approaches to mitigate the regression effect. Composite Gaussian process models (Ba and Joseph, 2012), Fractional Brownian Fields (Zhang and
Apley, 2014), and Gaussian process model with Brownian integrated covariance functions (Zhang and Apley, 2015) use flexible covariance functions that capture the non-stationarity of functions. Blind Kriging (Joseph et al., 2008) incorporates variable selection procedure in the mean function. Limit Kriging (Joseph, 2006) and Kernel Interpolation (Kang and Joseph, 2016) modify the standard Gaussian process regression model or the predictor for more accurate prediction in certain situations. Conditional Bias-Penalized Kriging (CBPK) (Seo, 2013) suggests minimizing the mean squared error plus the squared conditional bias to improve the performance at the extreme values. All these methods, except Limit Kriging and CBPK, introduce more complexity in the model than the stationary Gaussian process model.

2.2 Single Nugget Kriging

In this section, we propose a method with better predictions at extreme values than the standard method of Kriging. We construct our predictor in two ways: by penalizing the mean squared error through conditional bias and by penalizing the conditional likelihood at the target function value. Our prediction exhibits robustness to the model mismatch in the covariance parameters, a desirable feature for computer simulations with a restricted number of data points. Applications on several functions show that our predictor is robust to the non-Gaussianity of the function.

Throughout, we only consider deterministic functions and we use the model with a known (or estimated) constant mean \( \beta \) for simplicity. The simplification of the mean function to a constant does not affect predictive performance in general; see Sacks et al. (1989). We assume that the hyperparameters of the covariance function are known (or estimated from the data), and we focus on the prediction at a new point \( x_0 \).

In Section 2.2.1, we discuss conditioning the likelihood at the target, a fundamental idea of the SiNK. In Section 2.2.2, we define SiNK, and show that it gives smaller mean squared prediction error than usual Kriging when the function value is far from the mean function. In other words, SiNK is robust to misspecifying the mean function. In Section 2.2.3, comparison between the performance of SiNK and the performance of
usual Kriging and limit Kriging are given in several numerical experiments. Technical proofs are in Section 2.2.5

2.2.1 Conditional likelihood at the target and conditional bias

In this section, we investigate the idea of maximizing the conditional likelihood given the target function value, which is the supporting idea of the SiNK. We also define a class of predictors by generalizing the Conditional Bias-Penalized Kriging.

Conditional likelihood at the target

We formulate the prediction problem as an estimation problem. From the Gaussian process assumption, the density (which can be also viewed as the augmented likelihood in Jones (2001)) of \( (Y(x_0), Y(x_1), \ldots, Y(x_n)) \) is known. Instead of conditioning on the observed function values as in the Gaussian process regression, we condition on the unknown function value at the target point and compute the likelihood. We easily find that

\[
(Y(x_1), \ldots, Y(x_n)) \mid Y(x_0) = y_0 \sim N(\tilde{m}, \tilde{K}), \quad \text{where} \quad \tilde{m} = \beta \mathbf{1} + k(x_0, x_0)^{-1}(y_0 - \beta k(x_0)) \quad \text{and} \quad \tilde{K} = K - k(x_0, x_0)^{-1}k(x_0)k(x_0)^\top.
\]

Thus, we reverse the perspective by seeing \( y_0 \) as a parameter. The conditional log likelihood is

\[
l(y_0 \mid y) = l(y_0) = -\frac{1}{2}(y - \tilde{m})^\top \tilde{K}^{-1}(y - \tilde{m}) + \text{constant}.
\]

The maximizer of the conditional likelihood with respect to \( y_0 \) with penalty \(-\frac{(y_0 - \beta)^2}{2k(x_0, x_0)}\), which is the maximum a posteriori estimate of \( y_0 \) with the prior distribution \( y_0 \sim N(\beta, k(x_0, x_0)) \), is the simple Kriging predictor. However, when \( k(x_0) \neq 0 \), the maximizer of the conditional likelihood without penalty (CMLE)
exists and it is
\[ \hat{Y}_{CMLE}(x_0) = \beta + \frac{k(x_0, x_0) k(x_0)^\top K^{-1} (y - \beta 1)}{k(x_0)^\top K^{-1} k(x_0)} . \] (2.6)

The derivation is in Subsection 2.2.5. The CMLE is obtained by inflating the residual term of the simple Kriging predictor by \( 1/\rho(x_0)^2 \).

**Conditional bias**

In the geostatistical literature, there are two types of conditional bias (Katz and Murphy, 1997; Seo, 2013). *Type 1 conditional bias* is defined as \( \mathbb{E}[Y(x_0) | \hat{Y}(x_0) = \hat{y}] - \hat{y} \), which measures the degree of correspondence between the mean of the unknown function value given a particular prediction. This quantity has been used to measure the reliability of the forecast in geostatistics. For example, the simple Kriging predictor is type 1 conditionally unbiased. Type 1 conditionally unbiased predictors have been discussed with an interest in the issue of predicting tails better (Isaaks, 2005; David et al., 1984).

*Type 2 conditional bias* is defined as \( \mathbb{E}[\hat{Y}(x_0) | Y(x_0) = y_0] - y_0 \), which is computed by conditioning the true function value. If this bias is large for some \( y_0 \), then it means that the prediction could be bad for these \( y_0 \). There has not been much discussion on type 2 conditional bias, until Seo (2013) explicitly defined the Conditional Bias-Penalized Kriging, which will be discussed in the following subsection. We focus on type 2 conditional bias. Our intuition behind the new suggested predictor is that by reducing type 2 bias by the appropriate amount, it has better performance at the extreme values and it has analogous behavior as the nearest neighborhood regression.

For simple Kriging, we have \( \mathbb{E}[\hat{Y}_K(x_0) | Y(x_0) = y_0] = \beta + \rho(x_0)^2 (y_0 - \beta) \), which is not \( y_0 \) in general. Thus, \( \hat{Y}_K(x_0) \) is type 2 conditionally biased in general. We can expect that for a given \( y_0 \) that is far from the prior mean, the performance of standard Kriging can be poor.
Conditional Bias-Penalized Kriging

Conditional Bias-Penalized Kriging (CBPK) is defined as the linear predictor $\hat{Y}(x_0) = \beta + \lambda^\top (y - \beta 1)$ that minimizes the MSPE plus a multiple of squared type 2 conditional bias ($CB^2$)

$$E[(y_0 - \hat{Y}(x_0))^2] + \delta E[(y_0 - E[\hat{Y}(x_0)|y_0])^2] \quad \text{for some } \delta \geq 0 \quad (2.7)$$

with respect to $\lambda$. Seo (2013) suggests that we use $\delta = 1$. We show in proposition 1 that using $\delta = 1$ in (2.7) leads to the predictor

$$\hat{Y}_{CBPK}(x_0) = \beta + \frac{2}{1 + \rho(x_0)^2} k(x_0)^\top K^{-1} (y - \beta 1). \quad (2.8)$$

We observe that it is again a predictor with an inflated residual term. Different choices of $\delta$ in (2.7) lead to different predictors. If $\delta = 0$, (2.7) is the objective for simple Kriging, and thus the minimizer $\hat{Y}_{CBPK}(x_0)$ is the simple Kriging predictor. If $\delta \to \infty$, the minimizing predictor approaches the CMLE. This accords with the fact that the CMLE is type 2 conditionally unbiased, $E[\hat{Y}_{CMLE}(x_0)|Y(x_0) = y_0] = y_0$.

The main question when using a CBPK is over the ratio to use between MSPE and $CB^2$. We seek an automatic way to choose $\delta$ instead of simply using $\delta = 1$ or applying a cross-validation-style approach. We suggest varying the ratio spatially, using an appropriate function of $x_0$ as $\delta$. To distinguish from Seo’s CBPK ($\delta = 1$), we call the linear predictor $\hat{Y}$ that minimizes (2.7) (for general $\delta$) the generalized CBPK predictor. The proof of the following is in Subsection 2.2.5.

**Proposition 1.** For any nonnegative $\delta$ in (2.7), the generalized CBPK predictor for a constant mean model is of the form

$$\hat{Y}(x_0) = \beta + w(x_0) k(x_0)^\top K^{-1} (y - \beta 1)$$

where $w(x_0) \in [1, 1/\rho(x_0)^2)$ (if $\rho(x_0) = 0$, then the range is $[1, \infty)$). For every nonnegative $\delta$, there is a corresponding $w(x_0) \in [1, 1/\rho(x_0)^2)$.

In the next section, we focus on a generalized CBPK with specific $\delta$ that has
desired properties.

### 2.2.2 Single Nugget Kriging

In this section, we define the Single Nugget Kriging and discuss its properties. Verification of Definition 1 is in Subsection 2.2.5.

**Definition of SiNK**

**Definition 1.** The Single Nugget Kriging (SiNK) predictor is defined as

\[
\hat{Y}_{\text{SiNK}}(x_0) = \begin{cases} 
\beta + \frac{1}{\rho(x_0)} k(x_0)^\top K^{-1}(y - \beta 1) & \text{if } \rho(x_0) \neq 0 \\
\beta & \text{otherwise},
\end{cases}
\]

which is the maximizer of the conditional likelihood given \( Y(x_0) = y_0 \) with penalty

\[
\text{pen}(y_0) = -\frac{(y_0 - \beta)^2}{2k(x_0, x_0)} \frac{\rho(x_0)}{1 + \rho(x_0)}.
\]

Thus, the implicit prior distribution on \( y_0 \) is \( y_0 \sim N(\beta, k(x_0, x_0)(1 + 1/\rho(x_0))) \).

SiNK is defined as the *maximum a posteriori* estimator with a prior distribution on \( Y(x_0) \). We inflate the prior variance only at \( x_0 \) by the amount of uncertainty measured by \( \rho(x_0) \), to reduce the dependency on the prior. This is equivalent to assuming that there exist a nugget effect only on \( Y(x_0) \), so we call the method Single Nugget Kriging. We choose the specific penalty, or prior variance, because it yields two desirable properties which will be discussed in Section 2.2.2.

In the geostatistical literature, the nugget effect is designed to model functions that are discontinuous (Stein, 1999). In our definition, we inflate the prior variance, like a nugget of size \( k(x_0, x_0)/\rho(x_0) \) would, but only on \( Y(x_0) \) to introduce the additional uncertainty. This nugget is not from any additional noise assumption, and thus SiNK is also an interpolator like simple Kriging. If one wants to introduce a nugget effect or Gaussian noise to the Gaussian process, the SiNK predictor can be adjusted accordingly.
**Remark 1.** The SiNK predictor is the CBPK predictor with \( \delta = 1/\rho(x_0) \) in (2.7); it is the linear predictor
\[
\hat{Y}(x_0) = \beta + \lambda^\top(y - \beta \mathbf{1})
\]
where \( \lambda \) is the solution of the optimization problem
\[
\min_{\lambda} \mathbb{E}[(y_0 - \hat{Y}(x_0))^2] + \frac{1}{\rho(x_0)} \mathbb{E}[(y_0 - \mathbb{E}[\hat{Y}(x_0) | y_0])^2].
\]
This can be verified by plugging in \( \delta = 1/\rho(x_0) \) to (2.18).

**Remark 2.** The SiNK prediction can be discontinuous at points where \( \rho(x_0) = 0 \), which happens if and only if \( k(x_0) = 0 \). However, \( k(x_0) = 0 \) means that we do not have any information at the point \( x_0 \), so predicting with the prior mean is the best we can do. Note that \( k(x_0) = 0 \) could only happen if we use a kernel that has 0 in its range. If we use a strictly positive kernel, such as the Matérn kernel, then \( \rho(x_0) > 0 \) for every \( x_0 \). In practice, even though the prediction is theoretically well defined, dividing by \( \rho(x_0) \) can be numerically unstable when \( \rho \) is close to zero. A practical fix is to use
\[
\hat{Y}_{\text{SiNK}, \epsilon}(x_0) = \beta + \frac{1}{\max(\rho(x_0), \epsilon)} k(x_0)^\top K^{-1}(y - \beta \mathbf{1})
\]
for a small \( \epsilon \). We use \( \epsilon = 10^{-3} \) in our numerical work. A larger \( \epsilon \) would protect from bad estimators of length-scale parameters, something that we did not encounter in our numerical experiments.

**Remark 3.** One can construct the credible interval around the SiNK predictor based on the posterior from the implicit new prior \( y_0 \sim N(\beta, k(x_0, x_0)(1 + 1/\rho(x_0))) \) at \( x_0 \), which is wider than the credible interval from standard Kriging. Further theoretical and empirical study on constructing prediction intervals of SiNK is of interest.
As mentioned in Section 2.2.1, the ratio $\delta$ is now a function of $x_0$. The conditional bias penalty is larger when we have less information on the target function value. Penalizing by the conditional bias by an appropriate multiple of the conditional bias squared improves performance at extreme values. The rationale of using $\delta = 1/\rho(x_0)$ will be discussed in Section 2.2.2.

**Properties**

The main feature of SiNK is its stability which are represented as *boundedness* and *localness* in this section. We find that the SiNK predictor is the unique predictor with these properties in the class of generalized CBPK predictors with MSPE-CB ratio $\delta$ as a function of $\rho(x_0)$.

If the covariance function is stationary, then the SiNK predictor is bounded. On the contrary, CMLE is unbounded when $\rho(x_0)$ approaches 0. The proof is given in Subsection 2.2.5.

**Proposition 2** (Boundedness).

$$|\hat{Y}_{\text{SiNK}}(x_0) - \beta| \leq \sqrt{k(x_0, x_0)}\sqrt{(y - \beta 1)^\top K^{-1}(y - \beta 1)}$$ (2.10)

*Thus, if the covariance function is stationary, with probability 1,*

$$\sup_{x_0 \in \mathbb{R}^p} |\hat{Y}_{\text{SiNK}}(x_0)| < \infty.$$ (2.11)

For a predictor with inflated residual of simple Kriging predictor to be bounded, the maximum amount of inflation is of order $1/\rho(x_0)$. Roughly speaking, SiNK is the predictor with maximum inflation of the residual term that satisfies boundedness.

To discuss the localness property, we consider a more specific covariance class that contains many widely used covariance functions. A measurable function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is called *rapidly varying* of index $-\infty$, in the sense of de Haan (1970), if for any $t > 1$,

$$\lim_{x \to \infty} \frac{f(tx)}{f(x)} = 0.$$
holds. The class of rapidly varying functions is important in asymptotic analysis (see Bingham et al. (1989)). The Matérn kernel with $\nu > 0$ is rapidly varying of index $-\infty$, because the modified Bessel function of the second kind satisfies $K_\nu(z) \propto \exp(-z)/\sqrt{z} (1 + O(1/z))$ as $z \to \infty$. The squared exponential kernel is also rapidly varying of index $-\infty$. This technical assumption is necessary only for proving the localness property stated in Proposition 3. There are kernels that are not rapidly varying of index $-\infty$ such as rational quadratic covariance kernel (Rasmussen, 2006).

Now let $J_j$ be a set of points that have different distances from observations in the $j$'th coordinate,

$$J_j := \{x_0 \mid \|(x_0 - x_i)_j\| \neq \|(x_0 - x_l)_j\| \text{ for all } i \neq l, i, l \in \{1, 2, \ldots, n\}\}$$  \hspace{1cm} (2.12)

where $j \in \{1, 2, \ldots, p\}$. In Proposition 3 and Theorem 1, we assume that the new point $x_0$ and the observed points $x_1, \ldots, x_n$ are in $J_j$, $j = 1, \ldots, p$ to break the ties; we remove a measure zero set to simplify the argument. Also, define the neighborhood of an observation $x_i$ for $i \in \{1, 2, \ldots, n\}$ as

$$B(x_i) := \{x_0 \mid K(x_0, x_i) > K(x_0, x_l) \forall l \neq i, l \in \{1, 2, \ldots, n\}\}. \hspace{1cm} (2.13)$$

Thus, if $x_0 \in B(x_i)$, then $x_i$ is the closest observation to $x_0$ in terms of covariance.

**Proposition 3** (Localness). Suppose that the covariance function is a tensor product of stationary, rapidly varying of index $-\infty$, positive kernels with length scale parameter $\theta = (\theta_1, \ldots, \theta_p)$. Then

$$\lim_{\theta_j \to 0} \sup_{x_0 \in B(x_i) \cap J_j} |\hat{Y}(x_0) - Y(x_i)| = 0$$

with probability 1, where $J_j$ and $B(x_i)$ are sets of points defined in (2.12) and (2.13), respectively.

Thus, as $\theta_j \to 0$, if $x_i$ is the closest observation (in $j$'th coordinate) to $x_0$, then the SiNK predictor $\hat{Y}(x_0)$ converges to $Y(x_i)$. We can then show that the SiNK predictor is the only predictor that satisfies localness in the class of generalized CBPK
CHAPTER 2. PREDICTION IN COMPUTER EXPERIMENTS

predictors. As $\theta_j \to 0$, the simple Kriging predictor converges to the prior mean $\beta$.

**Theorem 1 (Uniqueness).** Consider a conditional biased penalized Kriging predictor

$$\hat{Y}(x_0) = \beta + w(x_0)k(x_0)^\top K^{-1}(y - \beta 1),$$

such that the covariance function is a tensor product of stationary, rapidly varying of
index $-\infty$, positive kernels with length scale parameter $\theta = (\theta_1, \ldots, \theta_p)$, and $w(x_0) \in [1, 1/\rho(x_0)^2]$ is a continuous function of $\rho(x_0)$. Suppose that $x_1, \ldots, x_n$ are in $J_j$, $j = 1, \ldots, p$. If there exists a $j \in \{1, 2, \ldots, p\}$ such that

$$\lim_{\theta_j \to 0} \sup_{x_0 \in B(x_i) \cap J_j} |\hat{Y}(x_0) - Y(x_i)| = 0$$

holds with probability 1, where $J_j$ and $B(x_i)$ are sets of points defined in (2.12) and
(2.13), respectively, then $w(\rho(x_0)) = 1/\rho(x_0)$, and $\hat{Y}(x_0)$ is the SiNK predictor.

The proofs of Proposition 3 and Theorem 1 are given in Subsection 2.2.5. Restricting $w(x_0)$ to be a function of $\rho(x_0)$ enables us to guarantee that $w(x_0) \in [1, 1/\rho(x_0)^2]$. For example, $w(x_0) = 1/\rho(x_0)$ is always in $[1, 1/\rho(x_0)^2]$. Another example for ne-
cessity of this condition is limit Kriging (Joseph, 2006) where the predictor has
$w(x_0) = 1/(k(x_0)^\top K^{-1}1)$. The limit Kriging predictor has the localness property,
but is not guaranteed to be a CBPK with nonnegative ratio $\delta$, which means we
cannot guarantee better performance at extreme values.

Figures 2.1 and 2.2 illustrate the property of SiNK and the difference from ordinary
Kriging. The function used in Figure 2.1 is the two-dimensional Zakharov function
in $[0, 1]^2$ is

$$f(x) = \sum_{j=1}^p x_j^2 + \left( \sum_{j=1}^p 0.5jx_j \right)^2 + \left( \sum_{j=1}^p 0.5jx_j \right)^4,$$

where $p = 2$, and the input points are 4 midpoints of the edges of a unit square.

We fitted ordinary Kriging and SiNK with an estimated constant mean and ten-
sor product Matérn 5/2 covariance. To visualize the difference, we evaluated the
predictors on \( x_1 = x_2 \), the dashed diagonal line on the square. Figure 2.2 shows the predictions with two different sets of parameters. For \( \theta_1 = \theta_2 = 1 \), the predictions are quite similar because \( \rho(\mathbf{x}_0) \approx 1 \) for all \( \mathbf{x}_0 \in [0, 1]^2 \). However, when \( \theta_1 = \theta_2 \) are close to zero (0.05), we observe significant differences between the two predictions. We also observe the localness property of SiNK. The \( \rho(\mathbf{x}_0) \) are close to zero for most of the plotted points, and thus the ordinary Kriging predictor is close to the estimated constant mean for points far from the observations. The SiNK predictor uses the function value of the observation that is the closest to the target point.

The localness property of SiNK is also related to the fact that the SiNK prediction at \( \mathbf{x}_0 \) only depends on the ratios of the correlations with observed function values. For instance, suppose that we predict at another point \( \mathbf{x}'_0 \) with covariance vector

Figure 2.1: Two-dimensional Zakharov function. The gray surface is the true function, four points on the surface are the observations. We compare the SiNK predictions and ordinary Kriging predictions on the dotted line \( x_1 = x_2 \) in Figure 2.2.
Figure 2.2: Illustration of the difference between ordinary Kriging and SiNK. SiNK performs better at extreme values than ordinary Kriging, more significantly when the correlations between function values are smaller.

\[ k(x_0') = c k(x_0), \text{ where } c \text{ is in } (0, 1). \quad \text{Then} \]

\[ \hat{Y}_{\text{SiNK}}(x_0') = \beta + \frac{k(x_0')^\top K^{-1}(y - \beta 1)}{\sqrt{k(x_0')^\top K^{-1} k(x_0')}} = \hat{Y}_{\text{SiNK}}(x_0). \]

Thus, the SiNK prediction at \( x_0' \) is the same as the prediction at \( x_0 \). However, the simple Kriging prediction is shrunk to \( \beta \) by a factor of \( c \). Thus, even if \( x_0' \) is far away from inputs, only the ratios of the correlation determine the SiNK prediction. Accordingly, SiNK does not automatically converge to the prior mean \( \beta \) as \( k(x_0) \to 0 \), for instance if one of the \( \theta_j \to 0 \).

**Mean squared prediction error at extreme values**

Since the simple Kriging predictor is the BLUP, the SiNK predictor has larger MSPE than the simple Kriging predictor. However, we can show that SiNK is only slightly inferior; the ratio of MSPEs is bounded. The proof is given in Subsection 2.2.5.
Proposition 4 (MSPE comparison).

\[
\mathbb{E}[(\hat{Y}_{\text{SiNK}}(x_0) - Y(x_0))^2] = \frac{2}{1 + \rho(x_0)} \mathbb{E}[(\hat{Y}_K(x_0) - Y(x_0))^2],
\]

the RMSPE of SiNK is at most \( \sqrt{2} \) times larger than the RMSPE of Kriging.

Here we show that SiNK has improved performance at extreme values. This can be represented in two ways; conditioning on a single extreme value of \( Y(x_0) \), and conditioning on a region of extreme \( Y(x_0) \) values.

Proposition 5. For a input \( x_0 \) with \( \rho(x_0) > 0 \), if

\[
\left| \frac{y_0 - \beta}{\sqrt{k(x_0, x_0)}} \right| \geq \sqrt{\frac{(1 + \rho(x_0))^2}{(1 + \rho(x_0))^2 - 1}}
\]

holds, then

\[
\mathbb{E}[(\hat{Y}_{\text{SiNK}}(x_0) - Y(x_0))^2 \mid Y(x_0) = y_0] \leq \mathbb{E}[(\hat{Y}_K(x_0) - Y(x_0))^2 \mid Y(x_0) = y_0].
\]

Proof. Directly follows from (2.21) in the proof of Proposition 4.

Proposition 6. Let \( \phi(\cdot) \) and \( \Phi(\cdot) \) be the density function and distribution function of the standard normal, respectively. Let \( S(x_0) = \left| (Y(x_0) - \beta) / (\sqrt{k(x_0, x_0)}) \right| \). For \( M > 0 \), if

\[
\rho(x_0) \geq -1 + \sqrt{1 + (1 - \Phi(M))/(M\phi(M))}
\]

holds, then

\[
\mathbb{E}[(\hat{Y}_{\text{SiNK}}(x_0) - Y(x_0))^2 \mid S(x_0) \geq M] \leq \mathbb{E}[(\hat{Y}_K(x_0) - Y(x_0))^2 \mid S(x_0) \geq M].
\]

The proof of Proposition 6 is given in Subsection 2.2.5. \( S(x_0) \) represents the z-score of the function value.

Figure 2.3 shows the relation between \( \rho(x_0) \) and the critical z-score or the threshold \( M \) for the z-score. The ratio of the region-conditional mean squared prediction
(a) The level curve for the mean squared error. SiNK outperforms simple Kriging in the region above and to the right of the given curves.

(b) Contour plot of CMSPE_{SiNK}/CMSPE_{K} (2.16). The ratio decreases as the threshold $M$ increases.

Figure 2.3: Relation between $\rho(x_0)$ and $z$-score.

The level curve for the mean squared error.

$$\frac{\text{CMSPE}_{\text{SiNK}}}{\text{CMSPE}_{\text{K}}} = \frac{\mathbb{E}[(\hat{Y}_{\text{SiNK}}(x_0) - Y(x_0))^2 | S(x_0) \geq M]}{\mathbb{E}[(\hat{Y}_{\text{K}}(x_0) - Y(x_0))^2 | S(x_0) \geq M]}$$

(2.16)

decreases as the threshold $M$ increases.

**Remark 4.** At function values around the model mean value $\beta$, the SiNK predictor has larger conditional MSPE than the simple Kriging predictor. For example, if $y_0 = \beta$,

$$\mathbb{E}[(\hat{Y}_{\text{SiNK}}(x_0) - Y(x_0))^2 | Y(x_0) = \beta] = k(x_0, x_0) - k(x_0)^\top K^{-1}k(x_0)$$

$$\mathbb{E}[(\hat{Y}_{\text{K}}(x_0) - Y(x_0))^2 | Y(x_0) = \beta] = k(x_0)^\top K^{-1}k(x_0) - \frac{(k(x_0)^\top K^{-1}k(x_0))^2}{k(x_0, x_0)}$$

which follow from the proof of Proposition 4 in Subsection 2.2.5. In this case, the conditional RMSPE of SiNK is $1/\rho(x_0)$ times larger than the conditional RMSPE of simple Kriging.
2.2.3 Numerical experiments

For numerical simulations, we used the DiceKriging package in R by Roustant et al. (2012). We fit the constant mean model for ordinary Kriging and SiNK, with the maximum likelihood estimator of the constant mean \( \hat{\beta} = (1^\top K^{-1} 1)^{-1} 1^\top K^{-1} y \). For the covariance function, we used tensor products of Matérn \( \nu = \frac{5}{2} \) kernels with maximum likelihood estimators of the length-scale parameters \( \theta_1, \ldots, \theta_p \), unless specified otherwise. We used \( \epsilon = 10^{-3} \) in (2.9).

To measure the performance of a predictor, we computed the empirical integrated squared error (EISE)

\[
\frac{1}{n_T} \sum_{i=1}^{n_T} (\hat{Y}(x_{test,i}) - Y(x_{test,i}))^2
\]  

with an independent set of \( n_T \) test points. Note that EISE is different from the MSPE; MSPE is the expected squared prediction error at a fixed point \( x_0 \). The EISE ratio of SiNK is computed by dividing the EISE of SiNK by the EISE of ordinary Kriging. We also report the test \( R^2 \) of the predictors, \( 1 - \text{EISE}/(\text{sample variance of } Y(x_{test})) \), to understand the relative errors of predictors. To measure the performance of a predictor at extreme values, we also computed the extreme empirical integrated squared error (EEISE)

\[
\frac{\sum_{i=1}^{n_T} (\hat{Y}(x_{test,i}) - Y(x_{test,i}))^2 1\{S(x_{test,i}) \geq 2\}}{\sum_{i=1}^{n_T} 1\{S(x_{test,i}) \geq 2\}}.
\]

Gaussian process example

We generated a realization of a 7-dimensional Gaussian process with zero mean and Matérn covariance with length-scale hyperparameters \( \theta = (1, 1, 1, 1, 1, 1, 1) \) and stationary variance \( k(x, x) = \sigma^2 = 1 \). The chosen observations were 100 points i.i.d. uniform in \([0, 1]^7\) and the test points were 2000 points i.i.d. uniform in \([0, 1]^7\).

To emulate the situation where the hyperparameters are unknown, we estimated
the hyperparameters by maximizing the likelihood. The estimated mean was \( \hat{\beta} = 0.143 \), the estimated length-scale hyperparameters were \( \hat{\theta} = (1.29, 0.92, 1.18, 1.41, 0.95, 0.76, 1.32) \), and the estimated stationary variance was \( \hat{\sigma}^2 = 0.94 \). The performance comparison between SiNK and ordinary Kriging is in Table 2.1. We observe that SiNK had slightly inferior EISE, but showed better performance at extreme values.

Table 2.1: Performance comparison of ordinary Kriging and SiNK for a realization of Gaussian process and piston function. The EISE ratios are the EISE of SiNK divided by the EISE of ordinary Kriging.

<table>
<thead>
<tr>
<th>Function</th>
<th>Gaussian Process</th>
<th>Piston Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>100</td>
<td>14</td>
</tr>
<tr>
<td>EEISE Ratio (SiNK/Ordinary Kriging)</td>
<td>0.820</td>
<td>0.814</td>
</tr>
<tr>
<td>EISE Ratio (SiNK/Ordinary Kriging)</td>
<td>1.020</td>
<td>0.887</td>
</tr>
<tr>
<td>( R^2 ) Ordinary Kriging</td>
<td>0.818</td>
<td>0.674</td>
</tr>
</tbody>
</table>

Figure 2.4 shows the prediction at test points with extreme function values. We first sorted the test points by the true function values and see the 1% largest and smallest function values. We observe that SiNK reduces the conditional bias by inflating the residual term. Differences are small but consistently in the right direction.

(a) Prediction at test points with 1% largest function values.  
(b) Prediction at test points with 1% smallest function values.

Figure 2.4: Ordinary Kriging and SiNK for a realization of 7-dimensional Gaussian process. Rank is the order of the true function values of the test points.
Piston function

We examined the performance of SiNK in a computer experiment. The piston simulation function in Kenett and Zacks (1998) models the circular motion of a piston within a cylinder. The response $C$ is the time it takes to complete one cycle, in seconds. The formula for the function is

$$C(x) = 2\pi \sqrt{\frac{M}{k + S^2 P_0 V_0 T_0 \over T_0} \over 2},$$

where

$$V = \frac{S}{2k} \left( \sqrt{A^2 + 4kP_0V_0 T_a - A} \right)$$

and $A = P_0 S + 19.62 M - kV_0 S$.

The description of the input variables is in Table 2.2. For this example, we adopted the Randomized QMC design (Faure sequence base 7) for observations and test points, because the number of observation is small. In Table 2.1, we see that in this case SiNK performs better not only at extreme values but also overall. This result possibly comes from the model mismatch of Gaussian process for the piston function; more specifically, the reduction of conditional bias may have had a large effect on the test error here.

Table 2.2: Input variables $x$ for the piston function.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>piston weight (kg)</td>
</tr>
<tr>
<td>$S$</td>
<td>piston surface area ($m^2$)</td>
</tr>
<tr>
<td>$V_0$</td>
<td>initial gas volume ($m^3$)</td>
</tr>
<tr>
<td>$k$</td>
<td>spring coefficient ($N/m$)</td>
</tr>
<tr>
<td>$P_0$</td>
<td>atmospheric pressure ($N/m^2$)</td>
</tr>
<tr>
<td>$T_a$</td>
<td>ambient temperature (K)</td>
</tr>
<tr>
<td>$T_0$</td>
<td>filling gas temperature (K)</td>
</tr>
</tbody>
</table>

Again, in Figure 2.5 the SiNK predictions are better at the test points with extreme function values than the ordinary Kriging predictions, and the difference is significant at the test points with 1% smallest function values. The inflation of the residual is consistently in the right direction, and larger than that of the Gaussian
process example.

(a) Prediction at the test points with 1% largest function values. (b) Prediction at the test points with 1% smallest function values.

Figure 2.5: Ordinary Kriging and SiNK for the piston function. Rank is the order of the true function values of the test points.

**Other functions**

We fitted ordinary Kriging, limit Kriging, and SiNK for several deterministic functions and compared the performances. The test function codes are from Bingham’s website (Bingham, 2013). The borehole function is described in Section 1.2.1. See Section 2.2.6 for the details of the other functions. For each test function, we trained with 100 independent sets of input points and computed the averages and standard deviations of metrics. The training points and test points were independent and uniformly distributed in the domain of inputs. Table 2.3 shows the dimension of the function, the number of observed points and test points, covariance type, $R^2$, EISE ratio, and EISE ratio at extreme values for each function. The number of training points for fitting each function was chosen so that the $R^2$ of ordinary Kriging is roughly 0.95, for all but the Robot Arm function which is a comparably difficult function to fit with our prediction methods.

We observe that for the five functions that we considered, SiNK performed better than ordinary Kriging in terms of EISE, and the performance gets even better
Table 2.3: Performance comparison among ordinary Kriging, limit Kriging and SiNK. The average and standard deviation of metrics from 100 independent set of inputs are reported. The EISE ratios are the EISE of SiNK or limit Kriging divided by the EISE of ordinary Kriging. The Friedman function did not have extreme values for most of the simulations.

<table>
<thead>
<tr>
<th></th>
<th>Borehole</th>
<th>Welch</th>
<th>Piston</th>
<th>Friedman</th>
<th>Robot Arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>8</td>
<td>20</td>
<td>7</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Training, test</td>
<td>32,500</td>
<td>320,5000</td>
<td>49,5000</td>
<td>50,5000</td>
<td>512,5000</td>
</tr>
<tr>
<td>Covariance type</td>
<td>Matérn 5/2</td>
<td>Matérn 5/2</td>
<td>Matérn 5/2</td>
<td>Matérn 5/2</td>
<td>Matérn 5/2</td>
</tr>
<tr>
<td>Fraction of extremes</td>
<td>0.091 (± 0.050)</td>
<td>0.249 (± 0.026)</td>
<td>0.072 (± 0.037)</td>
<td>0.000 (± 0.001)</td>
<td>0.051 (± 0.015)</td>
</tr>
<tr>
<td>EEISE Ratio (SiNK)</td>
<td>0.718 (± 0.084)</td>
<td>0.554 (± 0.068)</td>
<td>0.819 (± 0.063)</td>
<td>NA</td>
<td>0.696 (± 0.040)</td>
</tr>
<tr>
<td>EISE Ratio (Limit)</td>
<td>0.694 (± 0.134)</td>
<td>0.675 (± 0.033)</td>
<td>0.802 (± 0.121)</td>
<td>NA</td>
<td>0.877 (± 0.035)</td>
</tr>
<tr>
<td>R² (OK)</td>
<td>0.970 (± 0.015)</td>
<td>0.948 (± 0.005)</td>
<td>0.967 (± 0.011)</td>
<td>0.973 (± 0.013)</td>
<td>0.838 (± 0.011)</td>
</tr>
</tbody>
</table>

for extreme values in terms of EEISE. For instance, for the Welch function, the SiNK predictions at points with extreme function values (function values such that \(|z\text{-score}| > 2\)) have roughly half of the EISE of ordinary Kriging predictions. In addition, we observe that the performance of limit Kriging and SiNK is very similar in terms of EISE. Limit Kriging also shows improved performance at extreme values compared to ordinary Kriging, but the improvement is smaller or no different than the improvement of SiNK. For the Friedman function, there was not a test point function value which had \(|z\text{-score}|\) larger than 2 for most of the simulations, due to the large estimate of the stationary variance \(\sigma^2 = k(x, x)\). A suspicious estimate of the stationary variance can be found occasionally in practice, but it is not a problem for prediction because all three predictors that we are comparing do not depend on the estimate of \(\sigma^2\).

2.2.4 Discussion

We have presented an alternative to Kriging with improved predictions at the extreme values. We first found a link between conditional likelihood at the target and CBPK, and used it to define SiNK. In addition, we showed that SiNK has a boundedness and a localness property. In numerical experiments, we observed that SiNK generally performs better not only at extreme values but also in terms of overall integrated squared error. This result is possibly because the functions used in the examples may
not behave like typical realizations of stationary Gaussian processes.

2.2.5 Proofs

Proof of Proposition 1 (Generalized CBPK)

Proof. Without loss of generality, let $\beta = 0$. Expanding (2.7), we get

$$
\mathbb{E}[(y_0 - \lambda^\top y)^2] + \delta \mathbb{E}[(y_0 - \mathbb{E}[\lambda^\top y|y_0])^2]
= k(x_0, x_0) - 2\lambda^\top k(x_0) + \lambda^\top K\lambda + \delta \mathbb{E}[(y_0 - \lambda^\top \tilde{m})^2]
= k(x_0, x_0) - 2\lambda^\top k(x_0) + \lambda^\top K\lambda + \delta \left(1 - \frac{\lambda^\top k(x_0)}{k(x_0, x_0)}\right)^2 k(x_0, x_0).
$$

This is a quadratic function of $\lambda$. Thus, the minimizing $\lambda$ is

$$
\hat{\lambda} = \left(K + \frac{\delta}{k(x_0, x_0)} k(x_0) k(x_0)^\top\right)^{-1} (k(x_0) + \delta k(x_0)) = \frac{\delta + 1}{\delta \rho(x_0)^2 + 1} K^{-1}k(x_0)
$$

by the Woodbury formula. Then $w(x_0) = (\delta + 1)/(\delta \rho(x_0)^2 + 1)$, and

$$
\hat{Y}(x_0) = \beta + \frac{\delta + 1}{\delta \rho(x_0)^2 + 1} k(x_0)^\top K^{-1}(y - \beta 1).
$$

(2.18)

If $\rho(x_0) \neq 0$, then for $\delta \geq 0$, $w(x_0) \in [1, 1/\rho(x_0)^2)$, and $\lim_{\delta \to \infty} w(x_0) = 1/\rho(x_0)^2$.

If $\rho(x_0) = 0$, then for $\delta \geq 0$, $w(x_0) \in [1, \infty)$, and $\lim_{\delta \to \infty} w(x_0) = \infty$. \hfill \Box

Derivation of the CMLE

By the Woodbury formula,

$$
\hat{K}^{-1} = (K - k(x_0)k(x_0)^{-1}k(x_0)^\top)^{-1}
= K^{-1} + \frac{K^{-1}k(x_0)k(x_0)^\top K^{-1}}{k(x_0, x_0) - k(x_0)^\top K^{-1}k(x_0)}.
$$
Therefore
\[
k(x_0)^\top \tilde{K}^{-1} = k(x_0)^\top K^{-1} + \frac{k(x_0)^\top K^{-1}k(x_0)k(x_0)^\top K^{-1}}{k(x_0, x_0) - k(x_0)^\top K^{-1}k(x_0)} = \frac{1}{1 - \rho(x_0)^2} k(x_0)^\top K^{-1}.
\]

(2.19)

Thus, differentiating the conditional log likelihood (2.5) with respect to \( y_0 \),
\[
\frac{\partial l(y_0)}{\partial y_0} = \frac{1}{k(x_0, x_0)} (y - \tilde{m})^\top \tilde{K}^{-1}k(x_0) = \frac{1}{(1 - \rho(x_0)^2)k(x_0, x_0)} (y - \tilde{m})^\top K^{-1}k(x_0)
\]
from (2.19). Thus, when \( k(x_0) \neq 0 \), solving \( \frac{\partial l(y_0)}{\partial y_0} = 0 \) leads to
\[
\hat{y}_0 = \beta + \frac{1}{\rho(x_0)^2} k(x_0)^\top K^{-1} (y - \beta 1).
\]

Definition of SiNK

The logarithm of the posterior probability (up to a constant) is
\[
\log p(y_0|y) = -\frac{1}{2} (y - \tilde{m})^\top \tilde{K}^{-1} (y - \tilde{m}) - \frac{\rho(x_0)}{2(1 + \rho(x_0))} \frac{(y_0 - \beta)^2}{k(x_0, x_0)}.
\]

Differentiating with respect to \( y_0 \), we get
\[
\frac{\partial \log p(y_0|y)}{\partial y_0} = \frac{1}{k(x_0, x_0)} (y - \tilde{m})^\top \tilde{K}^{-1}k(x_0) - \frac{\rho(x_0)}{1 + \rho(x_0)} \frac{(y_0 - \beta)}{k(x_0, x_0)}
\]
\[
= \frac{1}{(1 - \rho(x_0)^2)k(x_0, x_0)} (y - \tilde{m})^\top K^{-1}k(x_0) - \frac{\rho(x_0)}{1 + \rho(x_0)} \frac{(y_0 - \beta)}{k(x_0, x_0)}
\]
from (2.19). If \( \rho(x_0) \neq 0 \), then solving \( \frac{\partial \log p(y_0|y)}{\partial y_0} = 0 \) leads to
\[
\hat{y}_0 = \beta + \frac{1}{\rho(x_0)} k(x_0)^\top K^{-1} (y - \beta 1).
\]
Proof of Proposition 2 (Boundedness)

Proof. By the Cauchy-Schwarz inequality,

\[
|\hat{Y}_{\text{SINK}}(x_0) - \beta| = \frac{1}{\rho(x_0)}|k(x_0)^\top K^{-1}(y - \beta 1)| \leq \frac{1}{\rho(x_0)} \sqrt{k(x_0)^\top K^{-1}k(x_0)} \sqrt{(y - \beta 1)^\top K^{-1}(y - \beta 1)}
\]

and equality holds when \( K^{-1/2}k(x_0) \) and \( K^{-1/2}(y - \beta 1) \) are parallel. If the covariance function is stationary, then the right hand side of (2.10) does not depend on \( x_0 \), thus (2.11) holds.

Proof of Theorem 1 and Proposition 3 (localness and uniqueness)

Proof. Let the stationary variance \( K(x, x) = \sigma^2 \). Now for a target point \( x_0 \in B(x_i) \cap J_j \), for \( l \neq i \), from the rapidly varying of index \(-\infty\) condition,

\[
\lim_{\theta_j \to 0} \frac{K(x_0, x_l)}{K(x_0, x_i)} = \lim_{\theta_j \to 0} \prod_{j' = 1}^{p} \frac{C_{\theta_j'}(|(x_l - x_0)_{j'}|)}{C_{\theta_j}(|(x_i - x_0)_{j'}|)} = \lim_{\theta_j \to 0} \prod_{j' = 1}^{p} C_1 \left( \frac{|(x_l - x_0)_{j'}|}{\theta_{j'}} \right) = 0.
\]

Thus we obtain

\[
\lim_{\theta_j \to 0} \frac{1}{K(x_0, x_i)} k(x_0) = e_i
\]

where \( e_i \) is the \( i \)-th unit vector. Noting that \( x_i \in B(x_i) \cap J_j \), we have

\[
\lim_{\theta_j \to 0} \frac{1}{\sigma^2} K = I_n,
\]
where \( I_n \) is the \( n \times n \) identity matrix. Thus,

\[
\lim_{\theta_j \to 0} \frac{\rho^2}{K(x_0, x_i)^2} = \lim_{\theta_j \to 0} \frac{k(x_0)^\top K^{-1}k(x_0)}{\sigma^2 K(x_0, x_i)^2} = \frac{1}{\sigma^4} \quad \text{and} \\
\lim_{\theta_j \to 0} \frac{\sigma^2 k(x_0)^\top K^{-1}(y - \beta 1)}{K(x_0, x_i)} = y_i - \beta.
\]

Thus,

\[
\lim_{\theta_j \to 0} \frac{K(x_0, x_i)}{\rho \sigma^2} = 1 \quad \text{and} \quad \lim_{\theta_j \to 0} \frac{\sigma^2 k(x_0)^\top K^{-1}(y - \beta 1)}{K(x_0, x_i)} = 1.
\]

Now we write the generalized CBPK as

\[
\hat{Y}(x_0) = \beta + w(\rho) k(x_0)^\top K^{-1}(y - \beta 1) \\
= \beta + w(\rho) \frac{K(x_0, x_i) \sigma^2 k(x_0)^\top K^{-1}(y - \beta 1)}{\rho \sigma^2}.
\]

Thus, to satisfy (2.14),

\[
\lim_{\theta_j \to 0} w(\rho) \rho = 1 \tag{2.20}
\]

is the condition that needs to hold. For the SiNK predictor, \( w(\rho) = 1/\rho \), so the condition holds, and therefore SiNK has the localness property and Proposition 3 holds.

The limit range of \( \rho \) as \( \theta_j \to 0 \) needs to be determined. For fixed \( x_0 \in B(x_i) \cap J_j \), \( \rho \to 0 \) as \( \theta_j \to 0 \). Now for any \( \delta \in (0, 1] \), let \( \epsilon = C^{-1}_\delta(\delta) \) and \( x_0 = x_i + \epsilon \theta_j e_j \). For all sufficiently small and positive \( \theta_j \), we have \( x_0 \in B(x_i) \cap J_j \). Then

\[
\lim_{\theta_j \to 0} \frac{K(x_0, x_i)}{\sigma^2} = \lim_{\theta_j \to 0} \prod_{j' = 1}^p C_{\theta_j}((x_l - x_0)_{j'}) = \lim_{\theta_j \to 0} C_{\theta_j}(\epsilon \theta_j) = C_1(\epsilon) = \delta
\]

Thus, \( \lim_{\theta_j \to 0} \rho = \delta \) for our selection of \( x_0 \). For (2.20) to hold, since \( w \) is a continuous function of \( \rho \), \( w(\delta) \delta = 1 \) must hold for all \( \delta \in (0, 1] \). To put it differently, if (2.14) holds, then it is the SiNK predictor. \( \square \)
CHAPTER 2. PREDICTION IN COMPUTER EXPERIMENTS

Proof of Proposition 4

Proof. From the conditional distribution of $y$ given $Y(x_0) = y_0$ (2.4),

$$
E[(\hat{Y}_K(x_0) - Y(x_0))^2 | Y(x_0) = y_0] = k(x_0)^\top K^{-1} k(x_0) - \frac{(k(x_0)^\top K^{-1} k(x_0))^2}{k(x_0, x_0)} + (y_0 - \beta)^2 \left(1 - \frac{k(x_0)^\top K^{-1} k(x_0)}{k(x_0, x_0)}\right)^2
$$

and

$$
E[(\hat{Y}_{\text{Sink}}(x_0) - Y(x_0))^2 | Y(x_0) = y_0] = k(x_0, x_0) - k(x_0)^\top K^{-1} k(x_0) + (y_0 - \beta)^2 \left(1 - \sqrt{\frac{k(x_0)^\top K^{-1} k(x_0)}{k(x_0, x_0)}}\right)^2 \quad (2.21)
$$

Now since $y_0 \sim N(\beta, k(x_0, x_0))$,

$$
E[(\hat{Y}_K(x_0) - Y(x_0))^2] = k(x_0, x_0) - k(x_0)^\top K^{-1} k(x_0) \quad \text{and}
$$

$$
E[(\hat{Y}_{\text{Sink}}(x_0) - Y(x_0))^2] = 2k(x_0, x_0) - 2\sqrt{k(x_0, x_0)k(x_0)^\top K^{-1} k(x_0)}
$$

$$
= \frac{2}{1 + \rho(x_0)} E[(\hat{Y}_K(x_0) - Y(x_0))^2].
$$

by the definition of $\rho(x_0)$.

Proof of Proposition 6

Proof. Let $\rho = \rho(x_0)$. From (2.21),

$$
E[(\hat{Y}_K(x_0) - Y(x_0))^2 | S(x_0)] = k(x_0, x_0)(\rho^2 - \rho^4 + S(x_0)^2(1 - \rho^2)^2) \quad \text{and}
$$

$$
E[(\hat{Y}_{\text{Sink}}(x_0) - Y(x_0))^2 | S(x_0)] = k(x_0, x_0)(1 - \rho^2 + S(x_0)^2(1 - \rho^2)).
$$
Using

\[ \mathbb{E}[S(x_0)^2 | S(x_0) > M] = \frac{1}{1 - \Phi(M)} \int_M^\infty s^2 \phi(s) ds = \frac{M \phi(M) + 1 - \Phi(M)}{1 - \Phi(M)} \]

we get the inequality for \( \rho \geq -1 + \sqrt{1 + (1 - \Phi(M))/(M \phi(M))}. \)

\[ \square \]

### 2.2.6 Test functions

**Welch function**

(Welch et al., 1992)

\[ f(x) = \frac{5x_1^{12}}{1 + x_1} + 5(x_4 - x_20)^2 + x_5 + 40x_1^{19} - 5x_19 \]

\[ + 0.05x_2 + 0.08x_3 - 0.03x_6 + 0.03x_7 - 0.09x_9 - 0.01x_{10} - 0.07x_{11} + 0.25x_1^{2} \]

\[ - 0.04x_{14} + 0.06x_{15} - 0.01x_{17} - 0.03x_{18}, \quad x \in [-0.5, 0.5]^{20}. \]

**Friedman function**

(Friedman et al., 1983)

\[ f(x) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5, \quad x \in [0, 1]^{5}. \]

**Robot arm function**

(An and Owen, 2001)

\[ f(x) = (u^2 + v^2)^{0.5} \]

where \( u = \sum_{i=1}^{4} L_i \cos \left( \sum_{j=1}^{i} \theta_j \right), \quad v = \sum_{i=1}^{4} L_i \sin \left( \sum_{j=1}^{i} \theta_j \right), \) and

\[ x = (\theta_1, \ldots, \theta_4, L_1, \ldots, L_4) \in [0, 2\pi]^4 \times [0, 1]^4. \]
CHAPTER 2. PREDICTION IN COMPUTER EXPERIMENTS

2.3 Universal Single Nugget Kriging

In this section, we discuss the limitations of universal Kriging and propose universal Single Nugget Kriging.

2.3.1 Advantages and disadvantages of universal Kriging

When the spatial data have a strong trend, and if the trend can be modeled by simple functions, then universal Kriging becomes handy. In the universal Kriging model

\[ Y(x) = \sum_{i=0}^{k} \beta_i f_i(x) + Z(x), \]  

(2.22)

the basis functions \( f_0, \ldots, f_k \) are known but the coefficients \( \beta_0, \ldots, \beta_k \) are unknown. If the covariance function is known, the model is essentially a linear regression model with spatially correlated residuals. Ordinary Kriging is the case when \( k = 0 \), and universal Kriging is the case when \( k \geq 1 \).

There are three main advantages of universal Kriging over ordinary Kriging. First, the universal Kriging predictor can be better than the ordinary Kriging predictor if there is a strong trend in the function. The mean function can capture the global trend of the function where the stochastic process \( Z \) can fit the local behavior of the function when the covariance function is chosen properly. Second, modeling the trend matters, especially in extrapolation situations. As discussed in Section 2.1, a limitation of the Kriging predictor is the regression to the mean function, and thus a good estimate of the mean function is critical at points far from the observed inputs. Last, the coefficients \( \beta_k \) are useful to interpret the relationship between the basis functions and the response.

However, these advantages are effective under certain conditions. Finding a good model that captures the trend can be very difficult, especially in high dimensional input space. Also, many studies claim that the simplification to the ordinary Kriging does not affect predictive performance.

A major drawback of universal Kriging is that the mean function and the stochastic process are confounding when we estimate the parameters in the model, which is
called *spatial confounding* by Clayton et al. (1993) and Reich et al. (2006). One way to remove the confounding is to orthogonalize the mean function and the stochastic process, using a specific covariance function. Orthogonal Gaussian process model by Plumlee and Joseph (2016) solves this identifiability issue. The paper also shows an example where universal Kriging cannot find an obvious trend in the function.

In sum, careful consideration should be made before fitting universal Kriging. In this section, we define universal Single Nugget Kriging and discuss its beneficial properties.

From the derivation of universal Kriging in Section 2.1, we observe that the universal Kriging predictor can be understood as a plug-in predictor, plugging in \( \hat{\beta} \) to the simple Kriging predictor with the known mean function \( f(x_0)^\top \beta \). \( \hat{\beta} \) in (2.2) is the generalized least square estimate of \( \beta \). The reason that the universal Kriging predictor matches with the plug-in predictor is because the loss function minimized to derive \( \hat{\beta} \) and \( \hat{Y}(x_0) \) are both quadratic loss functions. In general, the plug-in predictor is not a BLUP.

### 2.3.2 Universal Single Nugget Kriging

**Definition 2.** The universal Single Nugget Kriging (USiNK) predictor is defined as

\[
\hat{Y}_{\text{USiNK}}(x_0) = \begin{cases} 
    f(x_0)^\top \hat{\beta} + \frac{1}{\rho(x_0)} k(x_0)^\top K^{-1} (y - F \hat{\beta}) & \text{if } \rho(x_0) \neq 0 \\
    f(x_0)^\top \hat{\beta} & \text{otherwise,}
\end{cases}
\]

where \( \hat{\beta} = (F^\top K^{-1} F)^{-1} F^\top K^{-1} y \).

USiNK predictor is defined to be the plug-in predictor, using the generalized least square estimate of \( \beta \). Note that USiNK is not the minimizer of the MSPE penalized by the conditional bias (2.7) with an unbiasedness constraint. The plug-in predictors can be improved with a better estimate of the deterministic mean function.

USiNK predictor is designed to share the properties of SiNK. The residuals \( y - F \hat{\beta} \) are inflated by \( 1/\rho(x_0) \) to reduce the dependency on the mean function, which results in reducing the type 2 conditional bias. The predictor is also bounded given \( f(x_0)^\top \hat{\beta} \).
for any stationary covariance functions, and thus the prediction is guaranteed to not diverge for any misspecified stationary covariance function.

USiNK is favorable in many circumstances. First, when we know the right regressors \( f_0, \ldots, f_p \) and when there exists a strong global trend in the function, then universal Kriging and USiNK are expected to be good choices. Second, in the extrapolation setting, where the new input point \( x_0 \) is far from the inputs of the data, the universal Kriging predictor is pulled to the estimate of the global trend \( f(x_0)\hat{\beta} \). However, USiNK predictor depends more on the neighbor data points of the unobserved input point, which leads to more robust prediction. When the number of observations \( n \) is comparably small to the dimension \( p \), the benefit of using the neighbors can be amplified.

### 2.3.3 An illustrative example

We consider an illustrative example where USiNK works well. Let us consider an example function

\[
f(x) = \exp(-1.4x) + \cos(3.5\pi x),
\]

which is a one-dimensional function, evaluated in \([0, 1]\). Figure 2.6 is the plot of the function. Within the domain, there is a negative trend in the function with a large oscillation.

Suppose that we have four evaluations at \( x = 0.15, 0.4, 0.5, 0.75 \). The points are deliberately chosen to show the difference between the predictors. The ordinary Kriging model and the universal Kriging model are

\[
Y(x) = \beta_0 + Z(x) \quad \text{and} \quad Y(x) = \beta_0 + \beta_1 x + Z(x)
\]

respectively, where the covariance function of \( Z \) is a Matérn covariance function with \( \nu = 5/2 \) and estimated length-scale hyperparameter.

Figure 2.7 shows the predictions of ordinary Kriging (OK), ordinary Single nugget
Kriging (OSiNK), universal Kriging (UK), and universal single nugget Kriging (USiNK). We observe that the observations are not enough to fully understand the behavior of
Table 2.4: $R^2$ of the predictors in fitting the example function (2.23) with 4 function evaluations

<table>
<thead>
<tr>
<th></th>
<th>OK</th>
<th>OSiNK</th>
<th>UK</th>
<th>USiNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.417</td>
<td>0.617</td>
<td>0.586</td>
<td>0.690</td>
</tr>
</tbody>
</table>

the function in $[0, 1]$, which often happens in computer experiments. Especially, the regions $[0, 0.15)$ and (0.75, 1] need extrapolation of the data. We observe that OSiNK and USiNK are less sensitive to the estimated trend of the function, and make better predictions at points far from the observed inputs.

Table 2.4 shows the $R^2$ of the predictors, $1 - \text{EISE}/(\text{sample variance of } Y(x_{\text{test}}))$, in the damped cosine function example. We observe that OSiNK and USiNK are better than OK and UK respectively, mainly because there is a vast region of extrapolation in the input domain. We also observe that USiNK benefits from a better estimate of the global trend so that the predictor is better than OSiNK in this example.

2.3.4 Discussion

Universal Single Nugget Kriging shares the characteristics of universal Kriging and Single Nugget Kriging. When there is only a small number of function evaluations, and the function is expected to have a strong trend, then USiNK is a good choice for better predictions at unknown inputs. However, before applying USiNK, one should be aware of the limitations in universal Kriging discussed in Section 2.3.1, because USiNK also can have the same limitations.
Chapter 3

Dimension reduction in computer experiments

Many parametric studies in computer experiments, such as optimization and inverse problems, can be limited due to high-dimensionality. Reducing the input dimension helps us apply more complicated methods of analyzing the experiment. We often find that the relation between the input and the output can be mostly explained with only a few dimensions. Undoubtedly, not every function from computer experiments admits a low-dimensional structure. Still, dimension reduction methods can tell us the key factors — the original variables or the transformed variables — in an experiment.

The general goal of dimension reduction in regression can be formulated as searching a dimension reduction mapping \( R : \mathbb{R}^p \rightarrow \mathbb{R}^q \) such that

\[
y \approx g(R(x))
\]

where \( y \) is the response and \( g : \mathbb{R}^q \rightarrow \mathbb{R} \) is a low-dimensional function \((q < p)\). Dimension reduction methods in statistics are developed for a noisy response \( y \).

Dimension reduction methods in regression can be categorized into model-based...
dimension reduction methods and model-free dimension reduction methods. Model-based dimension methods fit specific regression models on $y|x$, for example,

$$y = g_\theta(R(x)) + \epsilon,$$  \hspace{1cm} (3.2)

where $g_\theta$ is a parametric model. In a model-based approach, the mapping $R$ can usually be determined from the model itself. Projection pursuit regression (Friedman and Stuetzle, 1981) is a common model-based dimension reduction method. Model-based methods are useful when one has a good understanding of the data-generating mechanism, so that the information can be exploited when modeling the association. Even when the model does not perfectly represent the mechanism, simple and robust models can help us understand the association between response and input variables.

However, in the exploratory stage of data analysis, we may not have enough knowledge to construct a good dimension-reduced regression model. Also, when we are interested in sensitivity analysis and inference on certain variables, then the model selection process can introduce a selection bias. There are many studies that address post-selection inference in the linear regression setting (Fithian et al., 2014; Lee et al., 2016), but post-selection inference is still challenging for complex regression models.

In this chapter, we focus on model-free dimension reduction methods, where we do not assume a specific regression model on $y$ given $x$. Model-free methods are especially useful for regression graphics and exploratory analysis. The methods are less affected by model selection, although determining the number of reduced dimensions could affect inference on the transformed input variables.

Model-free dimension reduction methods are especially useful in the analysis of computer experiments, where we usually do not have prior knowledge of the structure of black-box functions. We can fit regression models after reducing the dimensionality of the problem, so that we can avoid the curse of dimensionality and fit more complex models. However, in spite of not having a model of the association between $x$ and $y$, the performance of a model-free method depends strongly on the characteristics of the association. We discuss the performance of various model-free methods in this chapter.
The characteristics of $y|\mathbf{x}$ also depend heavily on input domain and distribution. For instance, the relation can be linear in a particular region and highly nonlinear in another region. The relation can be mostly explained by a single variable in a particular region, but not in other regions. It is possible that the input variables are correlated. Usually, the ranges of inputs are known in computer experiments; the ranges can stem from a physical law or can represent the domain of interest. Sometimes, the distributions of inputs are known as well. In this dissertation, we assume that the distributions of input variables are given, independent and continuous, which is the simplest case. We also assume that the input distributions can be transformed to simple distributions such as the Gaussian distribution.

Dimension reduction mappings are typically constrained to be linear in model-free approaches. We focus on linear dimension reduction mappings $R(\mathbf{x}) = Q_1^\top \mathbf{x}$.

Choosing an appropriate low-dimension $q$ is important and challenging in dimension reduction. An appropriate low-dimension $q$ depends on the dimension reduction mapping. For instance, consider the function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ such that

$$f(\mathbf{x}) = x_1 + x_2.$$ 

If the dimension reduction mapping is $R(\mathbf{x}) = (x_1 + x_2, x_1 - x_2)$, then the appropriate low dimension is one. However, if $R(\mathbf{x}) = (x_1, x_2)$, then the appropriate low dimension is two. Some dimension reduction methods propose asymptotic hypothesis testing for determining $q$. Other methods provide heuristic guides to choosing a reasonable $q$.

The rest of this chapter is organized as follows: in Section 3.1, we review classical sufficient dimension reduction methods. In Section 3.2, we review the theory of active subspaces. In Section 3.3, we study the reviewed dimension reduction methods in computer experiments and empirically analyze the performance with numerical examples. In Section 3.4, we propose the modified active subspaces which can improve the dimension reduction from the original active subspace. In Section 3.5, we analyze the quality of dimension reduction results using projected Gaussian processes. We quantify the proportional variance of the Gaussian process explained by the dimensionally reduced space. This helps us choose the dimension of the dimensionally
CHAPTER 3. DIMENSION REDUCTION IN COMPUTER EXPERIMENTS

reduced space.

3.1 Sufficient dimension reduction methods

Dimension reduction in regression has been a major point of interest in the field of statistics. The sufficient dimension reduction paradigm was developed by R.D. Cook to provide a framework to combine statistical sufficiency and dimension reduction.

In regression, we are interested in extracting information from data about the statistical dependence of a response variable \( y \) on a \( p \times 1 \) vector \( x \). The structural dimension of regression is the smallest number of distinct linear combinations of the predictors required to characterize the conditional distribution \( y \mid x \).

A dimension reduction subspace (DRS) for \( y \mid x \) is defined as the column space \( S(\eta) \) of a \( p \times q \) (\( q \leq p \)) matrix \( \eta \) such that

\[
y \perp \perp x \mid \eta^\top x. \tag{3.3}
\]

In other words, there is no loss of information if \( x \) is replaced by \( \eta^\top x \) in understanding the conditional distribution \( y \mid x \).

A subspace \( S \) is a central subspace \( S_{y\mid x} \) for \( y \mid x \) if \( S \) is a DRS and \( S \subset S_{drs} \) for all DRSs \( S_{drs} \). The definition of central subspace is based on the statistical sufficiency. Cook (1998) shows that a central subspace exists when \( x \) has a density with a convex support. The concept of central subspace is a convenient way to address the existence and uniqueness of a DRS.

Cook and Li (2002) defined the mean dimension reduction subspace (mean DRS) as the column space \( S(\eta) \) such that

\[
y \perp \perp \mathbb{E}[y \mid x] \mid \eta^\top x. \tag{3.4}
\]

It follows that a DRS is a mean DRS. The central mean subspace is defined analogously to the central subspace: A subspace \( S \) is a central mean subspace \( S_{\mathbb{E}[y \mid x]} \) for \( y \mid x \) if \( S \) is a mean DRS and \( S \subset S_{mdrs} \) for all mean DRSs \( S_{mdrs} \). In many standard regression models, \( S_{y \mid x} = S_{\mathbb{E}[y \mid x]} \).
Sufficient dimension reduction methods aim to estimate the central subspace, or a portion of the central subspace. Many methods find a $p \times p$ kernel matrix $K_{y|x}$ that satisfies $\mathcal{S}(K_{y|x}) \subset \mathcal{S}_{y|x}$. Then a DRS can be estimated using the eigendecomposition of the estimated kernel matrix $\hat{K}_{y|x}$.

Methods of estimating the central subspace or portions thereof include Average Derivative Estimation (ADE) (Härdle and Stoker, 1989), Sliced Inverse Regression (SIR) (Li, 1991a), Sliced Average Variance Estimation (SAVE) (Cook, 2000), Principal Hessian Directions (pHd) (Li, 1992), Simple contour regression (SCR), General contour regression (GCR) (Li et al., 2005), and Directional Regression (DR) (Li and Wang, 2007). In this dissertation, we review a few of the fundamental methods. Variations, extensions, and combinations have been developed in the literature, which are briefly discussed in Section 3.1.6.

Many dimension reduction methods only guarantee an estimate of a subspace of the central subspace. If a method can estimate the central subspace, then we say the method is exhaustive.

Most methods require certain assumptions on the input distribution. Earlier methods required Gaussian distribution on the inputs, which are relaxed to elliptical distributions. Many recent methods further relax this assumption for broader applications. For simplicity, we assume that the inputs are standardized. If $\text{Var}[x] = \Sigma$ and we define

$$z = \Sigma^{-1/2}(x - \mathbb{E}[x]),$$

we can easily verify that $\mathcal{S}_{y|x} = \Sigma^{-1/2}\mathcal{S}_{y|z}$. Thus, for simplicity, throughout this chapter, we assume $\mathbb{E}[x] = 0$ and $\text{Var}[x] = I$.

### 3.1.1 Sliced Inverse Regression

The Sliced Inverse Regression (SIR) proposed in the seminal paper (Li, 1991a) is based on the inverse regression $x|y$. We call

$$\mathcal{S}_{\mathbb{E}[x|y]} = \text{span}\{\mathbb{E}[x|y]\}$$
Suppose that the central mean subspace is spanned by $\eta$, $S_{y|x} = s(\eta)$. Then

$$E[x|y] = E[E[x|y, \eta^\top x]|y] = E[E[x|\eta^\top x]|y] = P_{\eta}E[x|y].$$

Thus, $E[x|y]$ belongs to $S_{y|x}$. The idea of SIR is to use $\text{Var}[E[x|y]]$, a positive semidefinite matrix, as a kernel matrix, and to use the first $d$ eigenvectors to estimate the central subspace.

To estimate $\text{Var}[E[x|y]]$, the response $y$ is sliced into bins $b_1, \ldots, b_B$ to approximately compute $E[x|y]$. The sample kernel matrix is defined as

$$\frac{1}{n} \sum_{i=1}^{B} n_i \bar{x}_{b_i} \bar{x}_{b_i}^\top$$

where $\bar{x}_{b_i}$ is the sample mean of $x$ in bin $b_i$. Then, the column space of the first $q$ eigenvectors is the SIR estimate of the central subspace. Many studies have shown that the SIR estimator is not very sensitive to the number of slices. One major drawback of SIR is that it is not exhaustive. SIR is essentially a first-moment based method, and fails when the response is symmetric about the origin. For example, when $E[x|y] = 0$, the SIR estimator will be irrelevant to the central subspace. Suppose that $y = x^2 + \epsilon$, where $\epsilon \sim N(0,1)$. Then $E[x|y] = 0$, and the SIR estimator of the central subspace will be only based on the noise.

After Cook and Weisberg (1991) pointed out this drawback in SIR, Li (1991b) proposed SIRII in the rejoinder of the SIR paper, which takes into account the curvature of the conditional variance. It can be shown that the column space of $\text{Var}[x|y] - E[\text{Var}[x|y]]$ is a subspace of $S_{y|x}$. The kernel matrix of SIRII is defined as

$$E[(\text{Var}[x|y] - E[\text{Var}[x|y]])^2].$$

The sample kernel matrix is defined analogously to the SIR sample kernel matrix. Li also suggested combining SIR and SIRII to increase the chance of discovering the
central space more, by defining

\[
\text{SIRII}_\alpha = (1 - \alpha)(\text{Var}[\mathbb{E}[\mathbf{x}|y]])^2 + \alpha \mathbb{E}[(\text{Var}[\mathbf{x}|y] - \mathbb{E}[\text{Var}[\mathbf{x}|y]])^2].
\]  

(3.6)

When \(\alpha = 0.5\), then SIRII_{0.5} is equivalent to SAVE, which we describe in the following section.

### 3.1.2 Sliced Average Variance Estimation

Sliced Average Variance Estimation (SAVE) is proposed in Cook and Weisberg (1991), a discussion paper on SIR. SAVE is based on the first and second moments of the conditional distribution \(\mathbf{x}|y\). It can be shown that the columns of

\[
I - \text{Var}[\mathbf{x}|y]
\]

are in \(S_{y|x}\). The kernel matrix of SAVE is defined as

\[
\mathbb{E}[(I - \text{Var}[\mathbf{x}|y])^2].
\]

The estimation procedure is the same as SIR; the response is sliced into bins to estimate \(\text{Var}[\mathbf{x}|y]\).

Under regularity conditions, SAVE is more comprehensive than SIR. However, many studies have shown that SAVE is more sensitive to the number of slices than SIR. It is also known that SAVE is not very efficient in estimating monotone trends for small to moderate sample sizes.

SAVE is exhaustive when the conditional distribution of \(\mathbf{x}\) given \(y\) is multivariate normal, which is quite restrictive (see Cook and Lee (1999)).
3.1.3 Principal Hessian Directions

The Principal Hessian Direction (pHd) utilizes the Hessian of the regression function. Li (1992) considered the case when the regression function takes the form

$$\mathbb{E}[y|x] = f(x) = h(\eta^T x)$$

for a twice-differentiable function $h : \mathbb{R}^q \rightarrow \mathbb{R}$. From the chain rule,

$$\mathbb{E}[\nabla^2 f(x)] = \eta \mathbb{E}[\nabla^2 h(t)] \eta^T.$$ 

Thus, the $q$ eigenvectors of $\mathbb{E}[\nabla^2 f(x)]$ span the column space of $\eta$.

The mean of the Hessian can be estimated using Stein’s lemma. If $z \sim N(0, 1)$, then for any twice-differentiable function $h : \mathbb{R} \rightarrow \mathbb{R}$ such that $\mathbb{E}[h''(z)] < \infty$,

$$\mathbb{E}[zh(z)] = \mathbb{E}[h'(z)]$$

and

$$\mathbb{E}[z^2h(z)] = \mathbb{E}[h(z)] + \mathbb{E}[h''(z)].$$

The lemma can be extended to the multivariate case. Suppose that $x \sim N(0, I)$ and let $h : \mathbb{R}^q \rightarrow \mathbb{R}$ be a twice-differentiable function such that $\nabla^2 h$ is integrable with respect to the distribution $N(0, I)$. Then,

$$\mathbb{E}[\nabla^2 h(x)] = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])xx^T].$$ (3.7)

Thus, an unbiased estimate of $\mathbb{E}[\nabla^2 f(x)]$ is

$$\hat{\Sigma}_{yxx} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})(x_i - \bar{x})^T.$$ 

$\hat{\Sigma}_{yxx}$ is the sample kernel matrix for pHd. The method estimates $\eta$ by the first $q$ eigenvectors that have the (absolute) largest $q$ eigenvalues of $\hat{\Sigma}_{yxx}$. Cook and Li (2002) showed that pHd targets the central mean subspace, not the central subspace.

pHd does not require slicing the response, because it is not based on inverse regression. There is no additional tuning parameter such as the number of slices. Empirically, the pHd estimator has large variance.
An obvious shortcoming of pHd is that it cannot find the linear trend because it is based on the Hessian. Li (1992) proposed fitting a linear regression first and replace $y$ by the residual. Also, pHd requires Gaussian distribution of $\mathbf{x}$. Li (1992) studied the behavior of the pHd estimator under weaker conditions on the distribution of $\mathbf{x}$.

### 3.1.4 Contour Regression

Li et al. (2005) defined the *empirical directions* as the vectors $\mathbf{x}_i - \mathbf{x}_j$, $1 \leq i \neq j \leq n$. Contour regression essentially extracts a subset of the empirical directions that have small variation in $y$.

Let $(\bar{\mathbf{x}}, \bar{y})$ be an independent copy of $(\mathbf{x}, y)$. Then Li et al. (2005) shows that the eigenvectors corresponding to the smallest $q$ eigenvalues of $K(c) = \mathbb{E}[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^\top | |y - \bar{y}| \leq c]$ span the central subspace, under certain assumptions on the central subspace $S_{y|x}$ associated with a positive constant $c$. A sample estimate of $K(c)$ using the empirical directions is

$$\hat{K}(c) = \frac{1}{n(n-1)/2} \sum_{i,j} (\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^\top I(|y_j - y_i| \leq c).$$

We use the smallest $q$ eigenvalues to estimate the central subspace. Li et al. (2005) call this method simple contour regression and extend it to generalized contour regression (GCR). It replaces the contour identifier $I(|y_j - y_i| \leq c)$ by a more sensitive one to improve the efficiency. It has been shown that contour regression is exhaustive and robust to the misspecification of input distribution.

One downside of contour regression is that the cutoff $c$ needs to be chosen, analogously to the number of slices in SIR and SAVE. Also, the computational cost of computing the empirical directions, which is $O(n^2)$, can be a burden for a large $n$. 
3.1.5 Directional Regression

Directional regression (Li and Wang, 2007) utilizes empirical directions differently than contour regression. Let \((\tilde{x}, \tilde{y})\) be an independent copy of \((x, y)\). Define

\[
A(y, \tilde{y}) = \mathbb{E}[(x - \tilde{x})(x - \tilde{x})^\top | y, \tilde{y}].
\]

Then it can be shown that the column space of \(2I - A(y, \tilde{y})\) is in \(S_{y|x}\), under certain conditions on the distribution \(x\). The conditions are claimed to be weaker than the conditions used in preceding methods. Analogously to other classic sufficient dimension reduction methods, the goal is to estimate \(G = \mathbb{E}[(2I - A(y, \tilde{y}))^2]\). Li shows that \(G\) can be expressed as

\[
G = 2\mathbb{E}[(\mathbb{E}[x x^\top | y])^2] + 2(\mathbb{E}[^\mathbb{E}[x | y]\mathbb{E}[x^\top | y])^2
+ \mathbb{E}[^\mathbb{E}[x^\top | y]\mathbb{E}[x | y]\mathbb{E}[x^\top | y]\mathbb{E}[x | y]\mathbb{E}[x^\top | y] - 2I_p,
\]

which can be estimated by slicing the response, as with SIR and SAVE.

We observe that \(G\) is a combination of the first and second moments. The computational cost is \(O(n)\). From numerical experiments, it is shown that DR is empirically more efficient in finding the central subspace than SIR, SAVE, and SCR.

3.1.6 More sufficient dimension reduction methods

Due to various types of data and limitations of the classical methods described in previous sections, dimension reduction in regression remains an active field of research. We introduce a few directions in the literature that improve the traditional methods in different circumstances. This categorization is neither exclusive nor exhaustive, and can be different from other categorizations in the literature.

Methods based on kernel smoothing

A big branch of sufficient dimension reduction methods is methods based on kernel smoothing. These methods aim to approximate the local behavior of \(y|x\) or \(\mathbb{E}[y|x]\)
nonparametrically. These methods are also called local methods, as compared to the global methods or moment-based methods that comprise SIR, SAVE, and pHd. Local methods require less restriction on the design distribution and practically work well for small sample sizes. It is challenging to apply these methods to high-dimensional problems because the kernels suffer from the curse of dimensionality. These methods are usually computationally more complex and have slow convergence rates than the moment-based methods.

A typical example of a kernel-smoothing-based method is Minimum Average Deviance Estimation (MAVE) (Xia et al., 2002), which approximates the derivative of $E[y|x]$ linearly. MAVE puts less restriction on the design distribution, but is sensitive to extreme values and can infer only about the central mean subspace. Another example is Sliced Regression (Wang and Xia, 2008), which applies local kernel regression to each slice. Kernel Dimension Reduction by Fukumizu et al. (2009) utilizes the reproducing kernel Hilbert space to characterize the conditional independence in the formulation of the central subspace, to directly estimate the central subspace.

**Ensemble methods and higher-moment-based methods**

SIR, SAVE, and pHd are based on the first two moments of $x$ and $y$. Using higher moments may provide more information on the conditional distribution. Estimating central subspaces via inverse third moments (Yin and Cook, 2003) is an example of using higher moments.

To mitigate certain limitations of traditional dimension reduction methods, many hybrid methods have also been studied. As discussed in the description of SIR, Li (1991b) proposed a convex combination of SIR$^2$ and SIRII, and Gannoun and Saracco (2003a,b) studied the asymptotic properties of this convex combination. Ye and Weiss (2003) proposed a bootstrap method to find an optimal combination of SIR, SAVE and pHd. Zhu et al. (2007) further investigated the hybrids of classical methods. Directional regression (Li and Wang, 2007) is a method that naturally combines the first and second moments of the inverse regression.

Applying transformations, including polynomials, to the response can also be regarded as ensemble methods. Yin and Cook (2003) used conditional $k$th moments,

Methods for a comparably large number of predictors

Model-free dimension reduction methods were first developed for graphical representation of data where the number of observations \( n \) is large enough compared to the number of the predictors. In many research fields, scientists work with data with many variables \( p \) and comparably few observations \( n \). Numerous studies have attempted to refine model-free dimension reduction methods for high-dimensional data.

The first technical difficulty that arises when \( n < p \) is the estimation of the covariance of the predictors \( \Sigma \), and the estimation of the inverse of the covariance upon which many classical dimension reduction methods rely. Cook et al. (2007) proposed a dimension reduction method that does not require matrix inversion, using the powers of the sample covariance matrix \( \hat{\Sigma} \).

Li and Yin (2008) proposed Sliced Inverse Regression with L1 or L2 regularization to work with \( n < p \) and highly correlated predictors. Cook et al. (2012) formulated the inverse regression with the assumption that \( \text{Var}[x|y] \) is nonstochastic, and proposed the Abundant Sufficient Dimension Reduction. Wu and Li (2011) proposed dimension reduction through least squares regression of a set of transformations of the response, and proposed regularization using the Smoothly Clipped Absolute Deviation (SCAD) penalty for variable selection.

Nonlinear sufficient dimension reduction methods

A more general sufficient dimension reduction as formulated in Cook (2007) is to find a dimension reduction mapping \( R \) such that

\[
y|x = y|R(x)
\]

which we call nonlinear sufficient dimension reduction. Wu et al. (2008) and Zhu and Li (2011) studied Kernel-based Sliced Inverse Regression (kSIR) to allow for nonlinear mapping. Li et al. (2011) proposed the Principal Support Vector Machines (PSVM)
to find the optimal separating hyperplane that separates the slices of the response variable.

Naturally, kernel mappings are essential in incorporating nonlinearity in the dimension reduction space, and many studies have formulated this problem to propose extensions of existing methods. Fukumizu et al. (2009) was one of the early studies in this direction, and more recently Lee et al. (2013) proposed the Generalized Sliced Inverse Regression (GSIR) and the Generalized Sliced Average Variance estimator (GSAVE) using their characterization of dimension reduction at the level of sigma-fields and function classes.

**Likelihood-based methods and semiparametric methods**

One class of dimension reduction methods is based on the likelihood of observations. Cook and Forzani proposed Likelihood Acquired Directions (LAD) (Cook and Forzani, 2009), which they claimed is more efficient than SIR, SAVE, and DR under certain conditions. Although the method was developed under conditional normality of the predictors given the response, Cook and Forzani (2009) verified that it is robust to non-normality. Ma and Zhu (2012) proposed semiparametric counterparts of SIR, SAVE, DR, pHd that require less assumption about the conditional distribution. A variety of new methods frame the problem around criteria optimization that differs from the existing methods.

**3.1.7 Discussion**

This chapter reviews methods of sufficient dimension reduction in regression that focus on estimating the central subspace or the central mean subspace for continuous predictors and univariate response. Estimation has been the primary focus of the literature on sufficient dimension reduction since Li (1991a) introduced the SIR. More recently, studies have addressed further problems of applying sufficient dimension reduction. To name a few, these include inference, evaluation, determination of the dimension of the dimensionally reduced space, regression on a multivariate or functional response, and regression on categorical variables.
The traditional model-free dimension reduction methods described in the previous sections have a few common characteristics. All of them find a symmetric kernel matrix and use eigendecomposition to estimate the central subspace or the central mean subspace. Also, the methods require certain conditions on the distribution of the inputs such as ellipticity.

The methods can be combined with each other or can be combined with nonparametric techniques such as kernel methods. There are many variations of the classic methods to adapt with certain background knowledge on data.

### 3.2 Active subspaces

Recently, the *active subspace* (Constantine, 2015) has become popular in the analysis of high-dimensional computer experiments. The approach finds the linear subspace of the input space that best describes variability in the function. Constantine et al. (2014) give the underlying theory of active subspaces with an application to an elliptic PDE model. Constantine et al. (2014) also fit a Kriging model to construct a response surface on the active subspace. Extensive work has been carried out using gradient information in estimating active subspaces; Constantine and Gleich (2014) estimate the number of gradient samples needed using random matrix theory, and Constantine et al. (2015b) discuss how to estimate gradients efficiently when they are not available directly from the data. Numerous applications on real simulators using the active subspaces have been studied, for example, optimizing the ONERA-M6 wing design (Lukaczyk et al., 2014) and analyzing the reactive flow in a hypersonic scramjet engine (Constantine et al., 2015a).

Consider a continuously-differentiable scalar-valued function $f : \mathbb{R}^p \rightarrow \mathbb{R}$. Let the gradient $\nabla f = \nabla_x f(x)$ be oriented as a column vector. We assume that there is a probability density function $\rho : \mathbb{R}^p \rightarrow \mathbb{R}_+ \cup \{0\}$ on the input space. Let

$$C = \mathbb{E}_x[(\nabla f)(\nabla f)^\top]$$  \hspace{1cm} (3.8)

where $\mathbb{E}_x$ stands for the integration over $\mathbb{R}^p$ with respect to $\rho$. The matrix $C$ is
symmetric and positive semi-definite, so we denote the real eigenvalue decomposition of $C$ as

$$C = QAQ^\top$$

where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, and $Q$ is the orthogonal matrix of eigenvectors. We are interested in the first $q$ eigenvalues and eigenvectors, where $1 \leq q < p$, so we write

$$Q = (Q_1 \; Q_2) \quad \text{and} \quad \Lambda = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix}$$

where $Q_1$ consists of the first $q$ columns of $Q$ and $\Lambda_1$ consists of the largest $q$ eigenvalues. We call the column space of $Q_1$ the *active subspace*. We call $Q_1^\top x$ the active variables and call $Q_2^\top x$ the inactive variables. The main idea of the active subspaces is to approximate $f(x)$ by

$$f_{AS}(Q_1^\top x) := \mathbb{E}_x[f(x)|Q_1^\top x].$$

For certain density functions $\rho$ on the input $x$, there exist constants $c_1$ such that

$$\mathbb{E}_x[(f(x) - f_{AS}(Q_1^\top x))^2] \leq c_1 \sum_{r=q+1}^p \lambda_r \quad (3.9)$$

which is a Poincâré-type inequality. For example, if the input $x$ follows $N(0, I_p)$, then the inequality holds with the optimal constant $c_1 = 1$ (Beckner, 1989).

Now suppose we have $n$ function values $f(x_1), \ldots, f(x_n)$, and $n'$ gradient samples $\nabla_x f(x_1), \ldots, \nabla_x f(x_{n'})$. The inputs for the function values and the inputs of the gradients are not necessarily the same, but we use the same $x_i$ for notational convenience. To predict a function value at a new input $x_0$, we can do regression analysis on the dataset $(Q_1^\top x_1, f_{AS}(Q_1^\top x_1)), \ldots, (Q_1^\top x_n, f_{AS}(Q_1^\top x_n))$ to construct a surrogate surface, for instance, by fitting a Kriging model.

In practice, we may not be able to compute $C$ analytically, or we may not be able to compute the conditional expectation to get $f_{AS}$. As substitutes for $C$ and $f_{AS}$, we
can use the Monte-Carlo estimates
\[ \hat{C} = \frac{1}{n'} \sum_{i=1}^{n'} (\nabla_x f(x_i))(\nabla_x f(x_i))^\top = \hat{Q} \hat{\Lambda} \hat{Q}^\top, \quad \hat{f}_{AS}(Q_1^\top x_j) = \frac{1}{N} \sum_{k=1}^{N} f(x_{j,k}) \]

where \( N \) is the number of Monte Carlo samples from the conditional distribution. In other words, each \( x_{j,k} \) satisfies \( \hat{Q}_1^\top x_{j,k} = Q_1^\top x_j \). We can use \( x_j \) as \( x_{j,1} \). Thus, for the Monte-Carlo estimate \( \hat{f}_{AS}(Q_1^\top x_j) \), we need at least \( N - 1 \) more function values for every \( j = 1, \ldots, n \) after estimating the active subspace. Samples from the conditional distribution may not be obtainable, so the special case \( N = 1 \) which leads to \( \hat{f}_{AS}(Q_1^\top x_j) = f(x_j), j = 1, \ldots, n \) can be used in practice.

### 3.3 Comparison of dimension reduction methods for computer experiments

In the standard regression setting with noise, sufficient dimension reduction methods have been developed to reduce the input dimension. We investigate the conditions where sufficient dimension reduction methods are successful or limited in the analysis of computer experiments. We also compare these methods with the active subspace that utilizes gradient evaluations of the function. The empirical performance of each method is examined via numerical examples.

#### 3.3.1 Sufficient dimension reduction methods for deterministic functions

Sufficient dimension reduction is developed based on the statistical sufficiency of the conditional distribution \( y|x \). When \( y \) is a deterministic function of \( x \), sufficient dimension reduction methods can still be applied. Especially, if \( y = f(x) = g(\eta^\top x) \) for a matrix \( \eta \) and a function \( g \), it is trivial that \( y \perp \perp x|\eta^\top x \). The reverse also holds.

**Proposition 7.** Suppose that \( x \) is random and \( f : \Omega \rightarrow \mathbb{R}^p \) is a deterministic function of \( x \). If \( f(x) \) is independent of \( x \) conditionally on \( A^\top x \) where \( A \) is a \( p \times q(p > q) \),
matrix, then \( f(x) = g(A^\top x) \) for some function \( g \) with probability 1.

**Proof.** Without loss of generality, let \( p = 1 \) and \( f \) be a scalar function. Let \( B \subset \mathbb{R} \) a measurable set. Then, due to conditional independence, it follows that

\[
P[f(x) \in B \mid A^\top x] = P[f(x) \in B \& x \in f^{-1}(B) \mid A^\top x]
\]

\[
= P[f(x) \in B \mid A^\top x] P[x \in f^{-1}(B) \mid A^\top x]
\]

\[
= P[f(x) \in B \mid A^\top x]^2.
\]

Thus,

\[
P[f(x) \in B \mid A^\top x] = 0 \text{ or } 1
\]

for any measurable set \( B \). Considering sets \( B = (-\infty, b] \), for every \( z \) in the domain of \( A^\top x \), there exists a unique \( b \) such that

\[
P[f(x) = b \mid A^\top x = z] = 1.
\]

Thus, there exists a function \( g \) such that \( g(A'x) = f(x) \) with probability 1.

Thus, the statistical sufficiency in deterministic functions is equivalent to assuming the representation \( y = g(A^\top x) \), which we discuss further in detail in the following subsection.

### 3.3.2 Ridge functions

Logan et al. (1975) started using the term *ridge function*, also referred to as plane waves, in computerized tomography. The original definition of a ridge function \( f : \mathbb{R}^p \to \mathbb{R} \) is

\[
f(x) = g(a^\top x)
\]

(3.10)

where \( a \) is a length \( p \) vector. Thus, ridge functions are essentially one-dimensional
functions. In statistics, approximation by linear combinations of ridge functions is studied under the name Projection Pursuit Regression (Friedman and Stuetzle, 1981). Ridge functions are also the basis of neural networks. Pinkus (2015) defined generalized ridge functions as the functions of the form

\[ f(x) = g(A^T x) \]  

where \( A \) is a \( p \times q \) matrix, with \( q < p \). Pinkus discussed the properties of general ridge functions in Pinkus (2015); Lin and Pinkus (1993), Pinkus (1997), and Pinkus (2013). We simply call functions of the form (3.11) as ridge functions following Constantine et al. (2016b).

Constantine et al. (2016a) state that many physical laws are ridge functions. According to Buckingham Pi theorem, a key theorem in dimensional analysis, all physical systems can be represented as functions of dimensionless parameters. Suppose that a physical system has a quantity of interest \( y \), with input parameters \( x = (x_1, \ldots, x_p) \) with \( k \) different units (m, kg, etc.). According to the derivation in Constantine et al. (2016a), physical laws can be represented as

\[ f(x) = h(A^T \log(x)) \]  

where \( h : \mathbb{R}^{p-k+1} \to \mathbb{R} \), and \( \log(x) \) is the vector of the logs of the input parameters.

For example, consider the eight-dimensional Borehole function (1.6) with inputs in Table 1.1. There are two different units (m, yr), and thus it can be represented as a seven-dimensional function. For physical systems with a comparably large number of units, the function can be represented in a comparably smaller number of dimensions.

Thus, Buckingham Pi theorem tells us that it is reasonable to attempt to find low-dimensional structures in high-dimensional functions in computer experiments.

### 3.3.3 Numerical examples

There has not been much study of the performance of sufficient dimension reduction methods for deterministic functions. We compare the performance of sufficient
dimension reduction methods for computer experiments with discussions.

We consider ridge functions in which we know there is a low-dimensional structure. Suppose that the true dimension reduction mapping is \( R(x) = Q_1^T x \), and \( Q_1 \) is estimated by \( \hat{Q}_1 \), where \( Q_1^T Q_1 = \hat{Q}_1^T \hat{Q}_1 = I_q \). We use the projection 2-norm,

\[
\text{dist}(Q_1, \hat{Q}_1) = ||Q_1 Q_1^T - \hat{Q}_1 \hat{Q}_1^T||_2,
\]

to evaluate the dimension reduction methods. The projection 2-norm is the metric of the \( q \)-dimensional Grassmannian, and also related to the largest canonical angle \( \theta_q \) between subspaces; \( \text{dist}(Q_1, Q_1^T) = \sin(\theta_q) \) (Golub and Van Loan, 2012). Thus, the distance is bounded by one.

Consider the following four functions:

\[
\begin{align*}
\phi_1(\mathbb{R}^{10} \rightarrow \mathbb{R}) : \phi_1(x) &= \cos(1.5x_1) + 0.5x_3^2, \\
\phi_2(\mathbb{R}^4 \rightarrow \mathbb{R}) : \phi_2(x) &= \frac{x_1}{0.5 + (x_2 + 1.5)^2} + (1 + x_2)^2, \\
\phi_3(\mathbb{R}^{10} \rightarrow \mathbb{R}) : \phi_3(x) &= \cos(2x_1) + \cos(x_2), \text{ and} \\
\phi_4(\mathbb{R}^5 \rightarrow \mathbb{R}) : \phi_4(x) &= \sin(1.5x_1) + 0.5\cos(1.5x_2).
\end{align*}
\]

The functions \( \phi_1 \) and \( \phi_2 \) are used in Li et al. (2005) with noise to demonstrate the performance of various dimension reduction methods. All four functions depend only on the first two inputs. Thus, the central subspace is \( S(Q_1) \), where \( Q_1 \) is a \( p \times 2 \) matrix whose columns are the first two columns of the identity matrix. Suppose that the input follows the standard Gaussian process \( N(0, I_p) \) where \( p \) represents the dimension of the function. We evaluate the example functions, without noise, at \( n \) inputs sampled from the input distribution and estimate the two-dimensional \( \hat{Q}_1 \).

We compare five standard sufficient dimension reduction methods: SIR, pHd, SAVE, SCR, and DR. For pHd, Li (1992) recommends that we estimate the directions of the residuals of the linear regression so we included this step in the analysis.

For SIR, SAVE and DR, we need to slice the response into bins so that we can estimate the moments of the inverse regression \( x|y \). For SCR, we must choose the cutoff for the contouring directions. We use \( \max(8, p + 3) \) for the number of slices,
which is recommended in Weisberg (2002), and set the cutoff for the SCR at 0.1 times the standard deviation of the response. Tuning these parameters can improve the performance of each method.

Table 3.1 shows dist($Q_1, \hat{Q}_1$) of the six methods. We chose the number of function evaluations $n$ differently because it was more difficult to estimate the central subspace in higher dimensions.

Table 3.1: Comparison of sufficient dimension reduction methods. The mean and standard deviation of the distance from the truth are computed, from 400 samples of inputs.

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$n$</th>
<th>SIR</th>
<th>pHd(y)</th>
<th>pHd(residual)</th>
<th>SAVE</th>
<th>SCR</th>
<th>DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>0.95(0.06)</td>
<td>0.77(0.14)</td>
<td>0.72(0.13)</td>
<td>0.32(0.10)</td>
<td>0.48(0.17)</td>
<td>0.25(0.07)</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0.94(0.08)</td>
<td>0.76(0.15)</td>
<td>0.71(0.14)</td>
<td>0.31(0.11)</td>
<td>0.47(0.17)</td>
<td>0.24(0.06)</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>0.41(0.21)</td>
<td>0.76(0.22)</td>
<td>0.71(0.24)</td>
<td>0.58(0.26)</td>
<td>0.55(0.22)</td>
<td>0.37(0.19)</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>0.88(0.17)</td>
<td>0.92(0.13)</td>
<td>0.72(0.20)</td>
<td>0.64(0.23)</td>
<td>0.76(0.18)</td>
<td>0.57(0.23)</td>
</tr>
</tbody>
</table>

We observe interesting results in Table 3.1. First of all, the performance of each method depends substantially on the function. For example, SAVE works better than SIR for $\phi_1$, but not for $\phi_3$. Sometimes the method completely fails. pHd(y) for $\phi_4$ is close to estimating $Q_1$ by a random orthogonal matrix.

We see that DR is the best among the classical methods, as argued by Li and Wang (2007). Although we can improve the methods by tuning and applying more complicated extensions, DR should be the first method to try in practice.

Apparently, none of the classical methods are very accurate for moderately small $p$ and moderately large $n$. For example, dist($Q_1, \hat{Q}_1$) = 0.24 for $\phi_2$ using DR translates to 13.9°. The distance becomes larger for larger $p$. For computer experiments where the prediction accuracy of the response surfaces is expected to be around 99%, the performances of the model-free dimension reduction methods are not in the same range.

Gradient evaluations are powerful for ridge functions. Let us consider the case where we also have the gradient evaluations at the points with the same cost. The
gradients of the example functions are

\[ \nabla \phi_1(x) = (-1.5 \sin(1.5x_1), 1.5x_2^2, 0, \ldots, 0) \]  \hspace{1cm} (3.14a)

\[ \nabla \phi_2(x) = \left( \frac{1}{0.5 + (x_2 + 1.5)^2}, 2(1 + x_2) - \frac{2x_1(x_2 + 1.5)}{(0.5 + (x_2 + 1.5)^2)^2}, 0, 0 \right) \]  \hspace{1cm} (3.14b)

\[ \nabla \phi_3(x) = (-2 \sin(2x_1), -\sin(x_2), 0, \ldots, 0) \]  \hspace{1cm} (3.14c)

\[ \nabla \phi_4(x) = (1.5 \cos(1.5x_1), -0.75 \sin(1.5x_2), 0, 0, 0) \]  \hspace{1cm} (3.14d)

respectively. Note that all the elements of the gradients are zero except the first two. Thus, the active subspace will be almost the same as the true \( Q_1 \) for each of the above examples. We discuss the gradient evaluations in more detail in Section 3.3.4.

We also empirically analyze the effect of the dimension \( p \) in sufficient dimension reduction. For function \( \phi_2 \), DR was the best method regarding the sample mean and the variance of the distance between \( \hat{Q}_1 \) and \( Q_1 \). For DR, we vary \( n \) from 25 to 2500 to see how the distance decreases over \( n \). We also consider the same function in 20-dimension:

\[ \phi_2'(\mathbb{R}^{20} \to \mathbb{R}) : \phi_2'(x) = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + (1 + x_2)^2. \]

Figure 3.1 shows the log-log plot of the sample size and the distance. For both \( \phi_2 \) and \( \phi_2' \), we observe a linear decay, which is expected from the \( \sqrt{n} \)-consistency of DR.

Figure 3.2 shows the plot of the distance over the sample size. Although the convergence is \( \sqrt{n} \), we observe that the performance is not great for moderately large \( n \).

### 3.3.4 Discussion

Sufficient dimension reduction methods are guaranteed to find the low-dimensional structure of noise-free ridge functions. From numerical studies, we found that sufficient dimension reduction methods do not have good accuracy for finite samples. We also observed that gradient evaluations can be very powerful.

Our study opens up many questions for future research. In some cases, gradient
evaluation comes with function evaluations; for instance, adjoint methods in engineering naturally estimate the gradients. Active subspace is a natural choice in these cases. In general, however, the gradients have to be estimated from function evaluations, for example, by finite-differences methods. The cost of one gradient evaluation becomes $O(p)$. But even in these cases, gradient evaluations can be $p$ times more important than the function evaluation themselves for estimating the low-dimensional structure of the function, and so it can be worth distributing the resources to estimate gradients. Incorporating information from gradient evaluations into the function evaluations can also improve accuracy. Gaussian process regression can be useful in incorporating both data.
When there is noise in the function, estimating the gradients is much more challenging. There are many methods of numerical differentiation for noisy functions, for example, Chartrand (2011), but they require careful tuning and designing of inputs.

### 3.4 Modified active subspaces

We propose a definition of modified active subspace using the average of gradients. For general quadratic functions, we show that the modified subspace minimizes an upper bound on the variance explained by its complement. In the examples that follow, we show that modification can substantially improve dimension reduction, as
compared to the original active subspace.

3.4.1 Introduction

Active subspaces (Constantine, 2015) have become popular in the analysis of high-dimensional computer experiments. Given a differentiable function $f: \mathbb{R}^p \rightarrow \mathbb{R}$ and an input distribution $x \sim \rho$, the active subspace of dimension $q$ is the space spanned by the first $q < p$ eigenvectors of $E_\rho[\nabla f(x)\nabla f(x)^\top]$.

The active subspace essentially minimizes an upper bound on the mean squared error when the function is projected to the subspace of active variables. Let $Q_1$ be the $p \times q$ matrix of first $q$ eigenvectors of $E[\nabla f(x)\nabla f(x)^\top]$ with associated eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q$. Let $Q_2$ be the $p \times (p-q)$ matrix of remaining $(p-q)$ eigenvectors with associated eigenvalues $\lambda_{q+1} \geq \ldots \geq \lambda_p$. The approximation of the function is naturally done by $f(x) \approx f_{AS}(Q_1^\top x) := E[f(x)|Q_1^\top x]$. The bound on the mean squared error is based on a Poincaré-type inequality:

$$
E[(f(x) - f_{AS}(Q_1^\top x))^2] = E[\text{Var}[f(x)|Q_1^\top x]] \\
\leq C_1 E[E[\nabla Q_2 f(x)^\top \nabla Q_2 f(x)|Q_1^\top x]] \\
= C_1 E[\nabla Q_2 f(x)^\top \nabla Q_2 f(x)] \\
= C_1 (\lambda_{q+1} + \ldots + \lambda_p).
$$

(3.15)

The inequality holds under certain conditions on the distribution $\rho$ (Constantine, 2015). Still, even under these conditions, equality in (3.15) holds if and only if $f(x)$ given $Q_1^\top x$ is a linear function of $Q_2^\top x$, which is quite restrictive. Based on a tighter inequality than (3.15), we show that using the first $q$ eigenvectors of

$$\frac{1}{2} E[\nabla f(x)\nabla f(x)^\top] + \frac{1}{2} E[\nabla f(x)]E[\nabla f(x)]^\top$$

can yield a better dimension reduction under certain conditions.
3.4.2 Modified active subspaces

**Definition 3 (Modified active subspace).** Given an input distribution $\mathbf{x} \sim \rho$, suppose that $f : \mathbb{R}^p \to \mathbb{R}$ is a differentiable function with a square-integrable gradient with respect to $\rho$. The modified active subspace is defined as the space spanned by the first $q < p$ eigenvectors of

$$
\frac{1}{2} \mathbb{E}[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top] + \frac{1}{2} \mathbb{E}[\nabla f(\mathbf{x})] \mathbb{E}[\nabla f(\mathbf{x})]^\top
$$

where the expectation is defined with respect to the input distribution $\rho$.

Suppose that we have $n$ gradient evaluations of the function, with inputs sampled independently from $\rho$. The estimated modified active subspace is defined as the space spanned by the first $q < p$ eigenvectors of

$$
\frac{1}{2n} \sum_{i=1}^{n} \nabla f(\mathbf{x}_i) \nabla f(\mathbf{x}_i)^\top + \frac{1}{2} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla f(\mathbf{x}_i) \right) \left( \frac{1}{n} \sum_{i=1}^{n} \nabla f(\mathbf{x}_i) \right)^\top.
$$

Note that if $\mathbb{E}[\nabla f(\mathbf{x})] = 0$, the modified active subspace is the same as the active subspace. For example, if $f(\mathbf{x}) = \mathbf{x}^\top C \mathbf{x}$ and $\mathbb{E}[\mathbf{x}] = 0$, then $\mathbb{E}[\nabla f(\mathbf{x})] = 0$.

Let us also show that if a function is a ridge function, the active subspace and the modified active subspace can both identify the lower subspace under certain conditions. Given an input distribution $\mathbf{x} \sim \rho$, let $f : \mathbb{R}^p \to \mathbb{R}$ be a function such that

$$
f(\mathbf{x}) = g(Q_1^\top \mathbf{x})
$$

for a $p \times q$ matrix $Q_1$ ($q < p$) and a differentiable function $g : \mathbb{R}^q \to \mathbb{R}$ with square-integrable gradient with respect to $\rho$. Then, if we write $\mathbf{z} = Q_1^\top \mathbf{x}$,

$$
\mathbb{E}[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top] = Q_1 \mathbb{E}[\nabla g(\mathbf{z}) \nabla g(\mathbf{z})^\top] Q_1^\top
$$
and
\[ \frac{1}{2} \mathbb{E}[\nabla f(x) \nabla f(x)^\top] + \frac{1}{2} \mathbb{E}[\nabla f(x)] \mathbb{E}[\nabla f(x)]^\top = Q_1 \left( \frac{1}{2} \mathbb{E}[\nabla g(z) \nabla g(z)^\top] + \frac{1}{2} \mathbb{E}[\nabla g(z)] \mathbb{E}[\nabla g(z)]^\top \right) Q_1^\top. \]

Thus, if \( \mathbb{E}[\nabla g(z) \nabla g(z)^\top] \) is a full \( q \)-rank matrix, then both approaches will identify the column space of \( Q_1 \) exactly. If \( \mathbb{E}[\nabla g(z) \nabla g(z)^\top] \) is not full rank, the modified active subspace will at least find the subspace that the active subspace finds, because \( \mathbb{E}[\nabla g(z)] \mathbb{E}[\nabla g(z)]^\top \) is a non-negative definite matrix. To be more specific, let \( Q_{\text{AS}} \) and \( Q_{\text{MAS}} \) be the matrix with first \( q \) eigenvalues of \( \mathbb{E}[\nabla f(z) \nabla f(z)^\top] \) and \((1/2)\mathbb{E}[\nabla f(z) \nabla f(z)^\top] + (1/2)\mathbb{E}[\nabla f(z)] \mathbb{E}[\nabla f(z)]^\top \) respectively. Then,
\[ C(Q_{\text{AS}}) \subseteq C(Q_{\text{MAS}}) \subseteq C(Q_1) \]
where \( C(Q) \) is the column space of matrix \( Q \). This holds for any input distribution \( \rho \). The ranks of the matrices \( \mathbb{E}[\nabla g(z) \nabla g(z)^\top] \) and \((1/2)\mathbb{E}[\nabla g(z) \nabla g(z)^\top] + (1/2)\mathbb{E}[\nabla g(z)] \mathbb{E}[\nabla g(z)]^\top \) are determined by not only the function \( g \) but also by the input distribution \( \rho \).

For example, if \( f \) is a linear function \( f(x) = b^\top x \) with \( b \neq 0 \), then by the above argument, both active subspace and modified active subspace can identify \( b \).

However, for more general functions such as quadratic function with a linear trend, the active subspace and the modified active subspace are different. To compare the two subspaces, we discuss criteria for dimension reduction.

### 3.4.3 Criteria for dimension reduction

A popular way to reduce the dimension of function is finding a low-dimensional subspace that preserves the variance of the function as much as possible.

Throughout this section, we use \( Q_1 \) and \( Q_2 \) as matrix parameters that span low-dimensional subspaces. \( Q_1 \) is a \( p \times q \) orthogonal matrix that spans the low-dimensional subspace of interest. \( Q_2 \) is a \( p \times (p - q) \) orthogonal matrix that makes \( Q = [Q_1 \ Q_2] \) orthonormal. In other words, \( Q_1 \) spans the important subspace and \( Q_2 \) spans the
unimportant subspace.

Let us define four quantities

\[ EV_{Q_1} = \mathbb{E}[\text{Var}[f(x)|Q_1^T x]], \]
\[ EV_{Q_2} = \mathbb{E}[\text{Var}[f(x)|Q_2^T x]], \]
\[ VE_{Q_1} = \text{Var}[\mathbb{E}[f(x)|Q_1^T x]], \] and
\[ VE_{Q_2} = \text{Var}[\mathbb{E}[f(x)|Q_2^T x]]. \]

\( EV \) stands for the expected conditional variance, and \( VE \) stands for the variance of conditional expectation. From the law of total variance,

\[ \text{Var}[f(x)] = EV_{Q_1} + VE_{Q_1} = EV_{Q_2} + VE_{Q_2}. \]

Thus,

\[ \arg \min_{Q_1} EV_{Q_1} = \arg \max_{Q_1} VE_{Q_1}, \] and
\[ \arg \max_{Q_2} EV_{Q_2} = \arg \min_{Q_2} VE_{Q_2}. \]

Several well-known dimension reduction approaches can be formulated as \( VE_{Q_1} \) maximization problems. For example, when screening variables in computer experiments using first order Sobol indices, the objective is to maximize \( VE_{Q_1} \). Active subspace minimizes an upper bound on \( EV_{Q_1} \) in (3.15), which is equivalent to maximizing a lower bound on \( VE_{Q_1} \). Minimum Average Variance Estimation (MAVE) proposed in Xia et al. (2002) minimizes \( EV_{Q_1} \) using a local linear approximation. Constantine et al. (2016b) define near-stationary subspace based on the \( EV_{Q_1} \) minimization problem.

On the other hand, we can define the important subspace by identifying the unimportant subspace, which can be done by minimizing \( EV_{Q_2} \) or maximizing \( VE_{Q_2} \) and finding the orthogonal complement \( Q_1 \) of \( Q_2 \). Note that this solution is not equal to \( \arg \max_{Q_1} EV_{Q_1} \) in general; we show an example in Section 3.4.5. The idea of identifying the orthogonal complement of the important subspace has also been discussed.
in the sufficient dimension reduction literature. For example, contour regression (Li et al., 2005) utilizes empirical directions to identify the orthogonal complement of the central subspace.

### 3.4.4 Quadratic functions

Quadratic approximation of functions from computer experiments is useful in understanding the local behavior of the function as well as the non-monotonic global trend. Quadratic approximation can be done by fitting a second-order Taylor expansion at a single point, by fitting a linear model on the gradient of the function, or by building a least squares approximation (with respect to $\rho$) with a quadratic basis. Let us consider quadratic functions with linear trend

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c$$

where $A$ is a symmetric matrix. Then, if $x \sim N(0, I)$,

$$\nabla f(x) = A x + b, \ E[\nabla f(x)] = b, \text{ and } E[\nabla f(x)\nabla f(x)^\top] = A^2 + bb^\top.$$ 

Thus, the solutions of active subspace and modified active subspace are

$$\arg \max_{Q_1} tr[Q_1^\top (A^2 + bb^\top) Q_1] \quad \text{and}$$

$$\arg \max_{Q_1} tr \left[ Q_1^\top \left( \frac{1}{2} A^2 + bb^\top \right) Q_1 \right]$$

(3.16)

respectively. Noting that

$$\text{Var}[f(x)] = tr\left[ \frac{1}{2} A^2 + bb^\top \right] \quad \text{and}$$

$$\text{Var}[E[f(x)|Q_1^\top x]] = \frac{1}{2} tr[Q_1^\top A Q_1 Q_1^\top A Q_1] + b^\top Q_1 Q_1^\top b,$$

we intuitively expect that the solution of modified active subspace would be closer to the optimal $Q_1$ for $VE_{Q_1}$ or $VE_{Q_2}$. In this section, we show that the modified active subspace minimizes an upper bound for $VE_{Q_2}$. We first present the trace optimization
property of eigenvectors, which many dimension reduction methods are based upon (Kokiopoulou et al., 2011).

**Proposition 8** (Trace optimization). Let $M$ be a $p \times p$ symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$ in decreasing order. Let $u_1, \ldots, u_p$ be the unit eigenvectors associated with the eigenvalues. Given $q \leq p$, let $U_{1:q} = [u_1, \ldots, u_q]$ and $U_{(p-q+1):p} = [u_{p-q+1}, \ldots, u_p]$. Then,

$$\max_{Q \in \mathbb{R}^{p \times q}, Q^\top Q = I} tr[Q^\top MQ] = tr[U_{1:q}^\top MU_{1:q}] = \lambda_1 + \ldots + \lambda_q.$$ 

and

$$\min_{Q \in \mathbb{R}^{p \times q}, Q^\top Q = I} tr[Q^\top MQ] = tr[U_{(p-q+1):p}^\top MU_{(p-q+1):p}] = \lambda_{p-q+1} + \ldots + \lambda_p.$$ 

The following proposition gives an upper bound on $VE_{Q_2}$ for quadratic functions.

**Proposition 9.** Let $x \sim N(0, I_p)$ and $f : \mathbb{R}^p \to \mathbb{R}$ such that

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c$$

where $A$ is a symmetric matrix. Let $Q_2$ be an $p \times (p-q)$ orthogonal matrix such that $Q_2^\top Q_2 = I_{p-q}$. Then,

$$\text{Var}[\mathbb{E}[f(x)|Q_2^\top x]] \leq tr\left[Q_2^\top \left(\frac{1}{2} A^2 + bb^\top\right)Q_2 \right].$$

**Proof.** Let $Q = [Q_1 \ Q_2]$ be the $p \times p$ orthonormal matrix such that $Q^\top Q = QQ^\top = I_p$. Then since $x \sim N(0, I_p)$, $Q_1^\top x$ and $Q_2^\top x$ are independent. Using $Q_1 Q_1^\top + Q_2 Q_2^\top = I_p$,

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c$$

$$= \frac{1}{2} x^\top Q_1 Q_1^\top A Q_1 Q_1^\top x + x^\top Q_1 Q_1^\top A Q_2 Q_2^\top x + \frac{1}{2} x^\top Q_2 Q_2^\top A Q_2 Q_2^\top x$$

$$+ b^\top Q_1 Q_1^\top x + b^\top Q_2 Q_2^\top x + c.$$
Thus,
\[
\mathbb{E}[f(\mathbf{x})|Q_2^\top \mathbf{x}] = \frac{1}{2} \mathbf{x}^\top Q_2 Q_2^\top A Q_2^\top x + b^\top Q_2^\top \mathbf{x} + c'
\]
where \( c' \) is a constant that does not depend on \( Q_2^\top \mathbf{x} \). For simplicity, writing \( Q_2^\top \mathbf{x} = \mathbf{z} \), \( \mathbf{z} \sim \mathcal{N}(0, I_{p-q}) \) and
\[
\mathbb{E}[f(\mathbf{x})|\mathbf{z}] = \frac{1}{2} \mathbf{z}^\top Q_2^\top A Q_2 \mathbf{z} + b^\top Q_2 \mathbf{z} + c'.
\]
Using the variance of the quadratic form of Gaussian variables, we get
\[
\text{Var}[\mathbb{E}[f(\mathbf{x})|Q_2^\top \mathbf{x}]] = \frac{1}{2} \text{tr}[Q_2^\top A Q_2 Q_2^\top A Q_2] + b^\top Q_2^\top b
\]
\[
\leq \frac{1}{2} \text{tr}[Q_2^\top A Q_2 Q_2^\top A Q_2] + \frac{1}{2} \text{tr}[Q_2^\top A Q_1 Q_1^\top A Q_2] + b^\top Q_2 Q_2^\top b
\]
\[
= \text{tr} \left[ Q_2^\top \left( \frac{1}{2} A^2 + bb^\top \right) Q_2 \right].
\]
Equality holds if and only if \( Q_2^\top A Q_1 = 0 \).

Now, if we apply Proposition 8 to the upper bound in Proposition 9, the upper-bound-minimizing \( Q_2 \) is the matrix with smallest \( p - q \) eigenvalues of \( (1/2)A^2 + bb^\top \), which is equal to \( (1/2)\mathbb{E}[\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top] + (1/2)\mathbb{E}[\nabla f(\mathbf{x})]\mathbb{E}[\nabla f(\mathbf{x})]^\top \) for quadratic functions.

The original active subspace essentially minimizes a looser upper bound, \( \text{tr}[Q_2^\top (A^2 + bb^\top) Q_2] \), of \( VE_{Q_2} \). Thus, we can expect the modified active subspace to be better, regarding \( VE_{Q_2} \), for functions that are well approximated by quadratic functions.

Remark 5. From the modification, we no longer have the same theoretical upper bound on the mean squared error \( EV_{Q_1} \) in (3.15) for general functions. However, we still have an upper bound. Let \( Q_1 \) be the orthogonal \( p \times q \) matrix that spans the modified active subspace, and let \( Q_2 \) be the orthogonal \( p \times (p - q) \) matrix that completes the orthonormal matrix with \( Q_1 \). Then, following the same logic as in
\[(3.15),
\]
\[
E[\text{Var}[f(x)|Q_1^\top x]] \leq C_1 E[\nabla Q_2 f(x)\nabla Q_2 f(x)] \\
\leq 2C_1 \left[ \frac{1}{2} E[\nabla Q_2 f(x)\nabla Q_2 f(x)] + \frac{1}{2} E[\nabla Q_2 f(x)] E[\nabla Q_2 f(x)] \right] \\
= 2C_1 (\lambda'_{q+1} + \ldots + \lambda'_{p})
\]
\[(3.17)\]

where $\lambda'_{q+1}, \ldots, \lambda'_{p}$ are the eigenvalues associated with $Q_2$.

Generalizing the modification to $\lambda E[\nabla f(x)\nabla f(x)^\top] + (1 - \lambda) E[\nabla f(x)] E[\nabla f(x)]^\top$ and tuning $\lambda$ would work in practice. However, $\lambda = 1/2$ gives us the tightest upper bound on $VE_{Q_2}$ for quadratic functions.

### 3.4.5 Numerical examples

We consider examples to compare modified active subspace with original active subspace. First, we consider a two-dimensional quadratic function where the two subspaces lead to very different dimension reductions. Second, we consider an eight-dimensional function, with additional discussion on how to estimate $EV_{Q_1}$ and $EV_{Q_2}$ when closed-form expressions are not available. Last, we consider real computer simulation functions.

**Two-dimensional quadratic function example**

We consider an extreme example where the original active subspace is not a good dimension reduction subspace. Consider a two-dimensional quadratic function

\[
f(x) = \frac{1}{2} x^\top Ax + b^\top x = 2^{-3/4} x_1^2 + x_2
\]
\[(3.18)\]

where

\[
A = \begin{pmatrix} 2^{1/4} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad b = (0, 1)^\top.
\]
Suppose that \( x \sim N(0, I_2) \). We are interested in the one-dimensional subspace that best explains the variance of the function. We compute the one-dimensional subspace in four ways: \( \arg\min_{Q_1} EV_{Q_1} \) (Oracle-\( EV_{Q_1} \)), orthogonal complement of \( \arg\max_{Q_2} EV_{Q_2} \) (Oracle-\( EV_{Q_2} \)), active subspace, and modified active subspace.

We can numerically compute
\[
\arg\min_{Q_1} EV_{Q_1} = \arg\max_{Q_1} VE_{Q_1} = (0, 1)^\top, \text{ and }
\arg\max_{Q_2} EV_{Q_2} = \arg\min_{Q_2} VE_{Q_2} = (-0.8408, 0.5414)^\top.
\]

However, the orthogonal complement of \( \arg\max_{Q_2} EV_{Q_2} \) is \( Q_1 = (0.5414, -0.8408)^\top \), which is different from \( \arg\min_{Q_1} EV_{Q_1} \).

We compute the solution of active subspace and the solution of modified active subspace, assuming that \( A, b \) are known. From (3.16), the solutions are
\[
\arg\max_{Q_1} tr\left[Q_1^\top (A^2 + bb^\top)Q_1\right] = (0, 1)^\top \text{ and }
\arg\max_{Q_1} tr\left[Q_1^\top \left(\frac{1}{2} A^2 + bb^\top\right)Q_1\right] = (1, 0)^\top
\]
respectively. Figure 3.3 shows the contour plot of the function with active subspace direction, modified active subspace direction, and \( VE_{Q_1} \)-maximizing direction. Table 3.2 shows the angles between the computed directions \( Q_1 \). The table also shows \( VE_{Q_1}, EV_{Q_2} \) normalized by the variance of the function. \( VE_{Q_1}/V \) and \( EV_{Q_2}/V \) can be interpreted as the proportion of variance explained by the direction \( Q_1 \). We observe that the modified active subspace exactly matches Oracle-\( EV_{Q_1} \), and the active subspace is orthogonal to the Oracle-\( EV_{Q_1} \).

This example is an extreme case that shows the active subspace and the modified active subspace can be very different subspaces, and the active subspace can perform poorly in reducing the mean squared error \( EV_{Q_1} \) for some functions. The same phenomenon can happen for high-dimensional functions as well. In practice, although the difference might not be as extreme as in the above example, the solutions can be fairly different when the first eigenvector of \( E[\nabla f(x)\nabla f(x)^\top] \) is far from the direction.
of average gradients $\mathbb{E} [\nabla f(x)]$.

Table 3.2: Example of a quadratic function in (3.18) ($p = 2$, $q = 1$). The quantities $VE_{Q_1}$ and $EV_{Q_2}$ are normalized by the variance of the function $V$. The goal is to maximize $VE_{Q_1}$ or $EV_{Q_2}$.

<table>
<thead>
<tr>
<th></th>
<th>Angle from $\overrightarrow{(0,1)}$</th>
<th>$VE_{Q_1}/V$</th>
<th>$EV_{Q_2}/V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle-$EV_{Q_1}$</td>
<td>0 °</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>Oracle-$EV_{Q_2}$</td>
<td>32.8 °</td>
<td>0.45</td>
<td>0.62</td>
</tr>
<tr>
<td>Active subspace</td>
<td>90 °</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>Modified active subspace</td>
<td>0 °</td>
<td>0.59</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Figure 3.3: Contour plot of the quadratic function in (3.18). The black solid arrow is $\arg \max_{Q_1} VE_{Q_1}$. The red solid arrow, which matches the black solid arrow, is the modified active subspace. The green dashed arrow is the orthogonal complement of $\arg \max_{Q_2} EV_{Q_2}$. The blue dotted arrow is the active subspace.
Eight-dimensional function example

Let us consider an eight-dimensional function \( f : [0, 1]^8 \rightarrow \mathbb{R} \) from Dette and Pepelyshev (2010),

\[
f(x) = 4(x_1 - 2 + 8x_2 - 8x_2)^2 + (3 - 4x_2)^2 + 16\sqrt{x_3 + 1}(2x_3 - 1)^2 \\
+ \sum_{k=4}^{8} k \log \left(1 + \sum_{i=3}^{k} x_i \right).
\] (3.19)

This function is highly curved in some variables and has less curvature in other variables.

Suppose that we are interested in a one or two-dimensional subspace that best explains the variance of the function, where the inputs are uniformly distributed in the input space. We transform the input distribution to follow standard Gaussian distribution by applying the inverse distribution function and standardization to each coordinate. We compare the estimated modified active subspace and the estimated active subspace when we have \( n = 200 \) gradient evaluations.

Unlike the quadratic function example, we do not have a closed form for the variance or for the conditional variances to compute the optimal subspace that minimizes \( EV_{Q_1} \) or maximizes \( EV_{Q_2} \). We can still compare the estimated quantities of \( EV_{Q_1} \) and \( EV_{Q_2} \) for active subspace and modified active subspace by the Monte Carlo method. We use a pick-freeze type method; note that for \( x, x' \) such that \( Q_1^\top x = Q_1^\top x' \) and \( x \perp \perp x'|Q_1^\top x \),

\[
E\left[\frac{1}{2}(f(x) - f(x'))^2\right] = E[\text{Var}[f(x)|Q_1^\top x]].
\]

Thus, we estimate \( EV_{Q_1} \) by

\[
EV_{Q_1} = \frac{1}{B_1} \sum_{i=1}^{B_1} \left[\frac{1}{2}(f(x_{i,1}) - f(x_{i,2}))^2\right].
\] (3.20)

The Monte Carlo points \( x_{i,j} \) are sampled by the following procedure:
1. Sample $\mathbf{x}_{i,1} \sim N(0, I)$ and compute $Q_1^T \mathbf{x}_{i,1}, i = 1, \ldots, B_1$.

2. Sample $\mathbf{x}_{i,2} \mid Q_1^T \mathbf{x}_{i,2} = Q_1^T \mathbf{x}_{i,1}, i = 1, \ldots, B_1$.

Essentially, we use a two-sample Monte Carlo estimate for the conditional variance; given a computational budget, fixing the number of inner Monte Carlo samples and increasing the number of outer Monte Carlo samples is usually more efficient in nested Monte Carlo estimation. Sun et al. (2011) show this in estimating the variance of conditional expectation.

In Table 3.3, we compare estimates of the active subspace and the modified active subspace. We use $B_1 = 10^6$ in (3.20). We ran the experiment 100 times with independent sets of inputs. We reported the mean and the standard deviation of the quantities and counted the number of times out of 100 when the modified active subspace was better than the active subspace for each quantity. We observe that the modified active subspace produced a better result for both $EV_{Q_1}$ and $EV_{Q_2}$. For example, when $q = 1$, the proportion of mean squared error ($EV_{Q_1}/V$) was reduced by 3.03% on average, and 86 replicates out of 100 showed a smaller mean squared error in the modified active subspace approach than in the original active subspace approach.

Table 3.3: Example of an eight-dimensional function in (3.19). The mean and the standard deviation of normalized $\hat{EV}_{Q_1}$ and $\hat{EV}_{Q_2}$ by the estimated variance $V$ over 100 sets of inputs are reported, where $n = 200$. The goals are to minimize $EV_{Q_1}$ or maximize $EV_{Q_2}$.

<table>
<thead>
<tr>
<th></th>
<th>$q = 1$</th>
<th>$q = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$EV_{Q_1}/V$</td>
<td>$EV_{Q_2}/V$</td>
</tr>
<tr>
<td>Active subspace</td>
<td>0.4533 (0.0026)</td>
<td>0.6194 (0.0087)</td>
</tr>
<tr>
<td>Modified active subspace</td>
<td>0.4230 (0.0572)</td>
<td>0.6952 (0.0555)</td>
</tr>
<tr>
<td>Number of times improved</td>
<td>86/100</td>
<td>100/100</td>
</tr>
</tbody>
</table>

Examples of computer experiments

We compare the modified active subspace with the active subspace in more complex functions that model computer simulations. We consider the borehole function, wing weight function, OTL circuit function, and power circuit function. The borehole function is described in Subsection 1.2.1 and the other functions are described in Subsection 3.4.7.
Again, suppose that we are interested in a one or two-dimensional subspace that best explains the variance of a function, where the inputs are uniformly distributed in the input space. We transform the input distribution to follow standard Gaussian distribution, and compare the estimated modified active subspace and the estimated active subspace when we have $n = 200$ gradient evaluations for each example functions.

In Table 3.4, we compare the estimates of the active subspace and the modified active subspace for the four example functions. We use $B_1 = 10^6$ in (3.20). We ran the experiment 100 times with independent sets of inputs. We reported the mean and the standard deviation of the quantities and counted the number of times out of 100 when the modified active subspace was better than the active subspace for each quantity. For each function, performance in terms of $EV_{Q_1}$ and $EV_{Q_2}$ was not very different, which means that the first eigenvector of the active subspace was quite close to the average gradient direction. Overall, the modified active subspace showed better performance, especially regarding $EV_{Q_1}$, when $q = 1$. For $q = 2$, we observe reduced differences. Especially for the OTL circuit function, the active subspace performed better than the modified active subspace regarding $EV_{Q_1}$, although the difference was small.

Table 3.4: The mean and the standard deviation of normalized $\hat{EV}_{Q_1}$ and $\hat{EV}_{Q_2}$ by the estimated variance $V$ are reported, over 100 sets of inputs, where $n = 200$. The goal is to minimize $EV_{Q_1}$ or maximize $EV_{Q_2}$.

<table>
<thead>
<tr>
<th>Function</th>
<th>$q = 1$</th>
<th>$q = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{EV}_{Q_1}/V$</td>
<td>$\hat{EV}_{Q_2}/V$</td>
</tr>
<tr>
<td>Borehole</td>
<td>Active subspace</td>
<td>Modified active subspace</td>
</tr>
<tr>
<td></td>
<td>Number of times improved</td>
<td></td>
</tr>
<tr>
<td>Wing weight</td>
<td>Active subspace</td>
<td>Modified active subspace</td>
</tr>
<tr>
<td></td>
<td>Number of times improved</td>
<td></td>
</tr>
<tr>
<td>OTL circuit</td>
<td>Active subspace</td>
<td>Modified active subspace</td>
</tr>
<tr>
<td></td>
<td>Number of times improved</td>
<td></td>
</tr>
<tr>
<td>Power circuit</td>
<td>Active subspace</td>
<td>Modified active subspace</td>
</tr>
<tr>
<td></td>
<td>Number of times improved</td>
<td></td>
</tr>
</tbody>
</table>
3.4.6 Discussion

In practice, it is hard to estimate $EV_{Q_1}$ or $EV_{Q_2}$ to compare the performance of dimension reduction approaches. However, by comparing the first eigenvector of $E[\nabla f(x)\nabla f(x)^\top]$ to the direction of $E[\nabla f(x)]$, we can see whether the modified active subspace makes a big difference in the resulting $Q_1$ from the original active subspace. If the eigenvector and the direction are close, then the modified active subspace will be close to the original active subspace.

3.4.7 Test functions

Wing weight function

(Forrester et al., 2008)

$$f(x) = 0.036S_w^{0.758}W_f^{0.0035} \left( \frac{A}{\cos^2(\Lambda)} \right)^{0.6} q^{0.006} \lambda^{0.04} \left( \frac{100t_c}{\cos(\Lambda)} \right)^{-0.3} (N_zW_{dg})^{0.49} + S_wW_p.$$  

The ranges of the 10 input variables are $S_w$: [150, 200], $W_f$: [220, 300], $A$: [6,10], $\Lambda$: [-10,10], $q$: [16,45], $\lambda$: [0.5,1], $t_c$: [0.08, 0.18], $N_z$: [2.5,6], $W_{dg}$: [1700, 2500], and $W_p$: [0.025, 0.08].

OTL circuit function

(Ben-Ari and Steinberg, 2007)

$$f(x) = \frac{(V_{b1} + 0.74)\beta(R_{c2} + 9)}{\beta(R_{c2} + 9) + R_f} + \frac{11.35R_f}{\beta(R_{c2} + 9) + R_f} + \frac{0.74R_f\beta(R_{c2} + 9)}{(\beta(R_{c2} + 9) + R_f)R_{cl}}$$  

where

$$V_{b1} = \frac{12R_{b2}}{R_{b1} + R_{b2}}.$$  

The ranges of the 6 input variables are $R_{b1}$: [50, 150], $R_{b2}$: [25, 70], $R_f$: [0.5, 3], $R_{c1}$: [1.2, 2.5], $R_{c2}$: [0.25, 1.2], and $\beta$: [50, 300].
CHAPTER 3. DIMENSION REDUCTION IN COMPUTER EXPERIMENTS 80

Power circuit function

(Kenett et al., 2013)

\[ f(x) = \frac{136.67(a + \frac{b}{x_{10}}) + d(c + e)\frac{g}{f} - h}{1 + d\frac{c}{f} + b[x_{10} + 0.006(1 + \frac{13.67}{x_{10}})] + 0.08202a} \]

where

\[ a = \frac{x_{2}}{x_{1} + x_{2}}, \]
\[ b = \frac{1}{x_{12} + x_{13}} \left( x_{3} + \frac{x_{1}x_{2}}{x_{1} + x_{2}} \right) + x_{9}, \]
\[ c = x_{5} + \frac{x_{7}}{2}, \]
\[ d = x_{11} \frac{x_{1}x_{2}}{x_{1} + x_{2}}, \]
\[ e = x_{6} + \frac{x_{7}}{2}, \]
\[ f = (c + e)(1 + x_{11})x_{8} + ce, \]
\[ g = 0.6 + x_{8}, \text{ and} \]
\[ h = 1.2. \]

The nominal values of the 13 input variables are \(x = (8200, 220000, 1000, 33000, 56000, 5600, 3300, 58.5, 1000, 120, 130, 100, 130)\). The values are assumed to vary \(\pm 1/60\) of their nominal values.

3.5 Analysis of dimension reduction in Gaussian process regression

Dimension reduction in computer experiments is often a stepping-stone in the analysis of data. Reducing the input dimension reduces the computational complexity in many problems, such as optimization problems and inverse problems. However, reducing the dimension of the input space induces uncertainty; even when the original function
is deterministic, the dimension-reduced inputs will not fully describe the function. It is challenging to find a principled way of modeling the variation from the removed dimensions in the functions. Assuming an independent noise in the response surface model can be an ad hoc way to model the variation.

Another challenge is to evaluate dimension reduction methods. In classical linear regression, $R^2$, the proportion of the variance explained by a linear model, can be an indicator of the goodness of fit. In the analysis of computer experiments, many response surface models can fit the data perfectly by interpolating. Thus, $R^2$ is not an appropriate measure of the goodness of fit for response surface models. Keeping hold-out data for evaluation is costly when the number of data points is small.

In this section, we propose a way to analyze dimension reduction methods using Gaussian process regression. Many studies have proposed dimension reduction methods based on Gaussian process models (Constantine et al., 2014; Tripathy et al., 2016). The length-scale hyperparameters in the covariance function of the Gaussian process represent the variability of the function. We quantify the importance of the variables in the dimensionally reduced space using the length-scale parameters in the Gaussian process. We also propose a consistent Gaussian process model on the original function, which supports fitting a standard Gaussian process regression on the dimensionally reduced subspace.

Suppose that we are given a dimension reduction mapping

$$\mathbf{x} \in \mathbb{R}^p \rightarrow Q_1^\top \mathbf{x} \in \mathbb{R}^q$$

on the input $\mathbf{x}$, where $Q_1$ is a $p \times q$ ($p > q$) orthogonal matrix. Many linear dimension reduction mappings such as Principal Component Analysis, Projection Pursuit Regression (Friedman and Stuetzle, 1981), and active subspaces (Constantine et al., 2014) can be represented in this form. When one is interested in variable selection, $Q_1$ consists of the first $q$ columns of a permutation matrix. Let $Q = [Q_1 \ Q_2]$ be a $p \times p$ orthogonal matrix. Many dimension reduction methods naturally produce $Q$; if not, we can complete $Q$ by padding an orthogonal complement $Q_2$.

The rest of this section is organized as follows: in Subsection 3.5.1, we propose a
CHAPTER 3. DIMENSION REDUCTION IN COMPUTER EXPERIMENTS

Gaussian process model that incorporates the dimension reduction mapping. In Subsection 3.5.2, the projected Gaussian process is defined and its mean and covariance function are discussed. In Subsection 3.5.3, we propose estimators of the proportion of the variance of the Gaussian process explained by dimension reduction. In Subsection 3.5.4, we extend the analysis to Matérn covariance functions. In Subsection 3.5.5, we apply the analysis to an example function. In Subsection 3.5.6, we conclude with a discussion.

3.5.1 Assumption about the Gaussian process

Suppose that \( f: \mathbb{R}^p \rightarrow \mathbb{R} \) is a realization of a Gaussian process with a constant mean \( \beta \) and a stationary covariance function. Suppose further that the covariance function is a form of

\[
\text{Cov}_f(f(x), f(x')) = \sigma_K^2 K(Q^T(x - x'))
\]

(3.21)

where \( \sigma_K^2 \) is the stationary variance of the process, and \( K: \mathbb{R}^p \rightarrow \mathbb{R} \) is a tensor product of one-dimensional covariance kernels. Vivarelli and Williams (1999) and Tripathy et al. (2016) used similar forms of covariance functions for dimension reduction. We first focus on the squared exponential covariance kernel, which is a standard covariance kernel for analyzing computer experiments,

\[
K(z) = \prod_{i=1}^{p} \exp \left( -\frac{z_i^2}{2\ell_i^2} \right),
\]

(3.22)

where \( \ell_1 \leq \ell_2 \leq \ldots \leq \ell_p \). A smaller length-scale \( \ell_i \) indicates that the variation in coordinate \( i \) is larger, which means that \( z_i \) is more important than other variables in explaining the variation of the function.

From (3.21), we implicitly assume that the rotation by \( Q^T \) arranges the variables in order of importance and also makes the covariance a tensor product. We discuss extensions to Matérn covariance kernels in Subsection 3.5.4.

We further assume that the input \( x \) follows standard Gaussian distribution \( N(0, I_p) \). This assumption simplifies the quantification.
There are two sources of randomness in our formulation; \( f \) is a Gaussian process, and the input follows the standard Gaussian distribution, and we assume these two are independent. Formally, \( f : \mathbb{R}^p \times \Omega \rightarrow \mathbb{R} \) is a stochastic process on a probability space \( \Omega \) with \( \sigma \)-algebra \( \mathcal{F} \). Based on the assumptions in the covariance functions (3.21) and (3.22), \( f(\cdot, \omega) \) is continuous for all \( \omega \in \Omega \). Functions that are continuous in one variable and measurable in another are called \textit{Carathéodory functions}, and these functions are jointly measurable (Aliprantis and Border, 1999). For a random input \( \mathbf{x} : \Omega_0 \rightarrow \mathbb{R}^p \), \( f(\mathbf{x}, \omega) \) is measurable.

Throughout, we write \( \mathbb{E}_x \) when the expectation is over the random input \( \mathbf{x} \), and we write \( \mathbb{E}_\mathcal{F} \) when the expectation is over \( \omega \). For example, \( \mathbb{E}_\mathcal{F}[f(\mathbf{x})|\mathbf{x}] \) denotes the mean of the Gaussian process given a fixed input, which is \( \beta \).

**Remark 6.** We can apply transformations to the input if its distribution does not follow the standard Gaussian distribution. A nonlinear transformation can change the dimension reduction space and the correlation structure in the Gaussian process.

### 3.5.2 Gaussian process projected on the dimensionally reduced space

We define

\[
\mathbf{f}_{\text{proj}}(Q_1^T \mathbf{x}) := \mathbb{E}_x[f(\mathbf{x})|Q_1^T \mathbf{x}, f]
\]  

which is the projected process on the dimensionally reduced subspace. Since integration is a linear operator, \( \mathbf{f}_{\text{proj}} \) is also a Gaussian process. We easily find that

\[
\mathbb{E}_\mathcal{F}[\mathbf{f}_{\text{proj}}(Q_1^T \mathbf{x})] = \beta.
\]

The covariance of \( \mathbf{f}_{\text{proj}} \) inherits the covariance structure of \( f \).
CHAPTER 3. DIMENSION REDUCTION IN COMPUTER EXPERIMENTS

Proposition 10 (Covariance of projected Gaussian process). Define

\[ \xi := \prod_{r=q+1}^{p} \left( 1 + \frac{2}{\ell_r^2} \right)^{-1/2}. \]  \hspace{1cm} (3.24)

Then, if \( Q_2^\top \mathbf{x} \) and \( Q_2^\top \mathbf{x}' \) are independent,

\[ \text{Cov}_r \left[ f_{\text{proj}}(Q_1^\top \mathbf{x}), f_{\text{proj}}(Q_1^\top \mathbf{x}') \mid Q_1^\top \mathbf{x}, Q_1^\top \mathbf{x}' \right] = \xi \sigma^2_K \exp \left( - \sum_{r=1}^{q} \frac{(Q_1^\top \mathbf{x} - Q_1^\top \mathbf{x}')^2}{2\ell_r^2} \right). \]

Proof. Joint measurability allows us to change the order of expectation. From the definition of \( f_{\text{proj}} \),

\[ \text{Cov}_r \left[ f_{\text{proj}}(Q_1^\top \mathbf{x}), f_{\text{proj}}(Q_1^\top \mathbf{x}') \mid Q_1^\top \mathbf{x}, Q_1^\top \mathbf{x}' \right] = \text{Cov}_r \left[ \mathbb{E}_x[f(x) \mid Q_1^\top \mathbf{x}, f], \mathbb{E}_{x'}[f(x') \mid Q_1^\top \mathbf{x}', f] \right] \]

\[ = \mathbb{E}_x \left[ \mathbb{E}_{x'}[K(Q^\top (\mathbf{x} - \mathbf{x}')) \mid Q_1^\top \mathbf{x} \mid Q_1^\top \mathbf{x}'] \right] \]

\[ = \sigma^2_K \exp \left( - \sum_{r=1}^{q} \frac{(Q_1^\top \mathbf{x} - Q_1^\top \mathbf{x}')^2}{2\ell_r^2} \right) \mathbb{E}_{x,x'} \left[ \exp \left( - \sum_{r=q+1}^{p} \frac{(Q_2^\top \mathbf{x} - Q_2^\top \mathbf{x}')^2}{2\ell_r^2} \right) \right]. \]

Since \( Q_2^\top \mathbf{x} \) and \( Q_2^\top \mathbf{x}' \) follow \( N(0, I_{p-q}) \), \( (Q_2^\top \mathbf{x} - Q_2^\top \mathbf{x}')^2 / 2\ell_r^2 \sim \chi^2_{(1)} \) for all \( r \). Using the moment generating function of \( \chi^2_{(1)} \) which is \( (1 - 2t)^{-1/2} \),

\[ \mathbb{E}_{x,x'} \left[ \exp \left( - \sum_{r=q+1}^{p} \frac{(Q_2^\top \mathbf{x} - Q_2^\top \mathbf{x}')^2}{2\ell_r^2} \right) \right] = \prod_{r=q+1}^{p} \left( 1 + \frac{2}{\ell_r^2} \right)^{-1/2}. \]

Thus,

\[ \text{Cov}_r \left[ f_{\text{proj}}(Q_1^\top \mathbf{x}), f_{\text{proj}}(Q_1^\top \mathbf{x}') \mid Q_1^\top \mathbf{x}, Q_1^\top \mathbf{x}' \right] = \xi \sigma^2_K \exp \left( - \sum_{r=1}^{q} \frac{(Q_1^\top \mathbf{x} - Q_1^\top \mathbf{x}')^2}{2\ell_r^2} \right). \]

Thus, the covariance function of \( f_{\text{proj}} \) is a multiplicative squared exponential covariance, aligned to the transformed input \( Q_1^\top \mathbf{x} \). The stationary covariance is reduced
by a factor of $\xi$. The variance lost by projection is quantified in Proposition 11.

**Proposition 11** (Variance lost by dimension reduction).

$$
\mathbb{E}_f\left[\text{Var}_x[f(x) \mid Q_1^\top x, f] \mid Q_1^\top x\right] = (1 - \xi)\sigma_K^2.
$$

**Proof.**

$$
\text{Cov}_f\left[f_{\text{proj}}(Q_1^\top x), f(x)\right] = \text{Cov}_f\left[\mathbb{E}_{x'}[f(x') \mid Q_1^\top x' = Q_1^\top x, f], f(x)\right]
\quad = \mathbb{E}_{x'}\left[\text{Cov}_f[f(x'), f(x)] \mid Q_1^\top x' = Q_1^\top x\right]
\quad = \sigma_K^2 \mathbb{E}_{x'}\left[\exp\left(-\sum_{r=q+1}^{p} \frac{(Q_2^\top x' - Q_2^\top x)_{r-q}}{2\ell_r^2}\right)\right].
$$

Since $Q_2^\top x'_{r-q} \sim N(0, 1)$, $(Q_2^\top x' - Q_2^\top x)_{r-q}^2 \mid x$ follows a noncentral chi-squared distribution. We use the moment generating function of the noncentral chi-squared distribution with noncentrality parameter $\lambda$ and degree of freedom one,

$$
\mathbb{E}[\exp(tZ)] = \exp\left(\frac{\lambda t}{1 - 2t}\right) \frac{1}{(1 - 2t)^{1/2}}.
$$

Let us define

$$
\eta = \prod_{r=q+1}^{p} \left(1 + \frac{1}{\ell_r^2}\right)^{-1/2}.
$$

Then,

$$
\text{Cov}_f\left[f_{\text{proj}}(Q_1^\top x), f(x)\right] = \eta\sigma_K^2 \exp\left(-\sum_{r=q+1}^{p} \frac{(Q_2^\top x)_{r-q}^2}{2(1 + \ell_r^2)}\right).
$$
Thus,
\[
\mathbb{E}_r[(f(x) - f_{\text{proj}}(Q_1^T x))^2] = \sigma_K^2 \left( 1 - \eta \exp \left( - \sum_{r=q+1}^{p} \frac{(Q_2^T x)^2_{r-q}}{2(1 + \ell_r^2)} \right) \right)
\]
and
\[
\mathbb{E}_x[\mathbb{E}_r[(f(x) - f_{\text{proj}}(Q_1^T x))^2]|Q_1^T x] = (1 - \xi)\sigma_K^2.
\]

Proposition 10 and 11 are summarized in Table 3.5. $K_{Q_1}$ represents the correlation function in Proposition 10.

Table 3.5: Summary of the decomposition of $f(x) | Q_1^T x$.

| Mean | $f_{\text{proj}}(Q_1^T x) | Q_1^T x \sim GP(\beta, \xi\sigma_K^2 K_{Q_1}(\cdot, \cdot))$ |
|------|------------------------------------------------------------------|
| Variance | $\mathbb{E}_r \left[ \text{Var}_x[f(x) | Q_1^T x, f] | Q_1^T x \right] = (1 - \xi)\sigma_K^2$ |

From Proposition 10, the proportion of variance of the Gaussian Process explained by the dimension reduction is

\[
\frac{\text{Var}_r[f_{\text{proj}}(Q_1^T x) | Q_1^T x]}{\text{Var}_r[f(x) | Q_1^T x]} = \xi = \prod_{r=q+1}^{p} \left( 1 + \frac{2}{\ell_r^2} \right)^{-1/2}
\]

which depends only on the variability in the reduced $p - q$ dimensions. The length-scale hyperparameters can be interpreted to the importance of the transformed variables in $Q_1^T x$. Dropping the $r$th variable in $Q_1^T x$ reduces the proportion by a factor of $(1 + 2/\ell_r^2)^{-0.5}$. If $\ell_r$ is large, the change in the proportion is negligible. On the other hand, $\xi$ can be close to zero if any of the dropped $\ell_r$'s are very small.

In (3.25), although both the numerator and the denominator are functions of $Q_1^T x$, they do not depend on $Q_1^T x$, and thus the ratio $\xi$ does not depend on $Q_1^T x$. This is because we assume a stationary covariance function in the Gaussian process model.
The proportion $\xi$ should be distinguished from $R^2$, the fraction of variance explained in regression. The usual variance in classical regression context is $\text{Var}_x[f(x)]$, the expectation of the squared deviation from its mean over the input distribution. The variance we focus on in the Gaussian process framework is the variance of the underlying Gaussian process. Figure 3.4 illustrates the variance of the Gaussian process.

We easily find that

$$
\xi = \prod_{r=q+1}^{p} \left(1 + \frac{2}{\ell_r^2}\right)^{-1/2} \geq \exp \left(- \sum_{r=q+1}^{p} \frac{1}{\ell_r^2}\right) \geq 1 - \sum_{r=q+1}^{p} \frac{1}{\ell_r^2}.
$$

When the dropped $\ell_r$'s are large, we can approximate $\xi$ with $1 - \sum_{r=q+1}^{p} 1/\ell_r^2$. Roughly speaking, the proportion explained is reduced by $1/\ell_r^2$ when we drop the $r$th variable.
3.5.3 Estimation of $\xi$

We define $\hat{\xi}$ as the plug in estimate using the estimates of hyperparameters $\hat{\ell}$ in the Gaussian process regression. For squared exponential covariance,

$$\hat{\xi} = \prod_{r=q+1}^{p} \left( 1 + \frac{2}{\ell_r^2} \right)^{-1/2}.$$  

(3.26)

The maximum likelihood estimators are standard choices for $\hat{\ell}$ in fitting the Gaussian process regression to $\{(x_i, f(x_i))\}_{i=1}^{n}$ with the covariance function in (3.21).

In practice, the regression procedure follows after the dimension reduction, so the estimates of $\ell_{q+1}, \ldots, \ell_p$ are not available. We discuss a heuristic way to estimate $\xi$ from the regression to dimensionally reduced data.

The unexplained proportion of variance $1 - \xi$ can be modeled as noise when fitting Gaussian process regression to the dimensionally reduced data $\{(Q_1^T x_i, f(x_i))\}_{i=1}^{n}$. Let us write $Q_1^T x = z$. We model the association between $Q_1^T x$ and $f(x)$ as a Gaussian process $g$ with mean $\beta$ and covariance

$$\text{Cov}_r(g(z), g(z')) = \sigma_{\text{proj}}^2 \prod_{r=1}^{q} \exp \left( -\frac{(z_r - z'_r)^2}{2\ell_r^2} \right) + \tau^2 I(z = z')$$  

(3.27)

where the correlation function comes from Proposition 10.

Suppose that the estimate of the variance $f_{\text{proj}}$ is $\hat{\sigma}_{\text{proj}}^2$ and the estimate of the variance of noise is $\hat{\tau}^2$. Solving

$$\hat{\sigma}_{\text{proj}}^2 = \tilde{\xi} \hat{\sigma}_K^2 \quad \text{and} \quad \hat{\tau}^2 = (1 - \tilde{\xi}) \hat{\sigma}_K^2$$

leads to

$$\hat{\sigma}_K^2 = \hat{\sigma}_{\text{proj}}^2 + \hat{\tau}^2 \quad \text{and} \quad \tilde{\xi} = \frac{\hat{\sigma}_{\text{proj}}^2}{\hat{\sigma}_{\text{proj}}^2 + \hat{\tau}^2}.$$
We use $\tilde{\xi}$ and $\tilde{\sigma}^2_K$ to differentiate from the plug-in estimates in the full Gaussian process regression.

In practice, $\tilde{\xi}$ overestimates $\xi$ in many cases. In the extreme case, Gaussian process regression can interpolate the data and then the estimate $\hat{\tau}^2$ can be nearly zero even if the dimensionally reduced inputs do not explain the variability of the function. Thus, $\tilde{\xi}$ should be used only as a heuristic guide. We will compare $\xi$ and $\tilde{\xi}$ in a numerical example in Section 3.5.5.

### 3.5.4 Extensions to general Matérn kernels

We can extend our analysis to the Matérn covariance case. Matérn covariance kernels (Matérn, 1986) are defined as

$$K^\nu_\ell(d) = \frac{(\sqrt{2\nu d}/\ell)^\nu}{\Gamma(\nu)2^{\nu-1}} K_\nu\left(\sqrt{2\nu d}/\ell\right)$$

where $K_\nu : \mathbb{R} \to \mathbb{R}$ is the modified Bessel function of the second kind. Matérn covariance kernels are commonly used in practice because the smoothness of the sample path, defined in terms of its mean square differentiability, can be parametrized through $\nu$. A Gaussian process with Matérn covariance has sample paths that are $k$ times mean square differentiable if and only if $k < \nu$ (Rasmussen, 2006). We consider kernels with $\nu > 1$.

The Matérn covariance function has a simple closed form when $\nu$ is a form of $q + 1/2$, where $q$ is a non-negative integer. In this case, the covariance function is a product of an exponential and a polynomial of order $q$. The most popular cases are $\nu = 3/2$ and $\nu = 5/2$, with the corresponding kernels

$$K^{3/2}_\ell(d) = \left(1 + \frac{\sqrt{3}d}{\ell}\right) \exp\left(-\frac{\sqrt{3}d}{\ell}\right) \quad \text{and}$$

$$K^{5/2}_\ell(d) = \left(1 + \frac{\sqrt{5}d}{\ell} + \frac{5d^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}d}{\ell}\right).$$

As $\nu \to \infty$, the Matérn covariance converges to the squared exponential kernel $K(d) =$
\( \exp(-d^2/\ell^2) \).

Suppose now that the correlation function in (3.21) is a form of

\[ K(z) = \prod_{i=1}^{p} K_{\nu}(z_i). \]

All of the arguments still hold except the formula for \( \xi \) in Proposition 10. We need to compute

\[ \mathbb{E}_x [K'_{\ell_r}((Q_1^\top x - Q_1^\top x')_r)] \]

where \((Q_1^\top x - Q_1^\top x')_r \sim N(0, 2)\). We can use the moment generating function of the half-normal distribution (Tsagris et al., 2014) to find a closed form for \( \xi \). Let \( Z \sim N(0, 1) \), then

\[ \mathbb{E}[\exp(t|Z|)] = 2 \exp \left( \frac{t^2}{2} \right) \Phi(t) \]

where \( \Phi(t) \) is the cumulative distribution of \( N(0, 1) \). Taking the derivatives with respect to \( t \) leads us to

\[
\mathbb{E}[|Z| \exp(t|Z|)] = 2t \exp \left( \frac{t^2}{2} \right) \Phi(t) + \frac{\sqrt{2}}{\sqrt{\pi}} \\
\mathbb{E}[|Z|^2 \exp(t|Z|)] = 2 \exp \left( \frac{t^2}{2} \right) \Phi(t) + 2t^2 \exp \left( \frac{t^2}{2} \right) \Phi(t) + \frac{\sqrt{2}}{\sqrt{\pi}} t.
\]

Thus, if \( \nu = 3/2 \),

\[
\mathbb{E}_x [K'_{\ell_r}((Q_1^\top x - Q_1^\top x')_r)] = \mathbb{E}_x \left[ \left( 1 + \frac{\sqrt{6}|W|}{\ell_r} \right) \exp \left( -\frac{\sqrt{6}|W|}{\ell_r} \right) \right] \\
= 2 \exp \left( \frac{3}{\ell_r} \right) \Phi \left( -\frac{\sqrt{6}}{\ell_r} \right) + \frac{\sqrt{6}}{\ell_r} \left[ -2\sqrt{6} \exp \left( \frac{3}{\ell_r^2} \right) \Phi \left( -\frac{\sqrt{6}}{\ell_r} \right) + \frac{\sqrt{2}}{\sqrt{\pi}} \right] \\
= \frac{2\sqrt{3}}{\sqrt{\pi} \ell_r} + \left( 2 - \frac{12}{\ell_r^2} \right) \exp \left( \frac{3}{\ell_r^2} \right) \Phi \left( -\frac{\sqrt{6}}{\ell_r} \right).
\]
and

\[ \xi = \prod_{r=q+1}^{p} \left( \frac{2\sqrt{3}}{\sqrt{\pi} \ell_r} + \left( 2 - \frac{12}{\ell_r^2} \right) \exp \left( \frac{3}{\ell_r^2} \right) \Phi \left( -\frac{\sqrt{6}}{\ell_r} \right) \right). \]

Similarly, if \( \nu = 5/2 \),

\[
\mathbb{E}_x \left[ K_{\ell}^\nu ((Q_1^\top x - Q_1^\top x')_r) \right] = \frac{6\sqrt{5}\ell_r^2 - 20\sqrt{5}}{3\sqrt{\pi} \ell_r^3} + \left( 2 - \frac{40}{3\ell_r^2} + \frac{200}{3\ell_r^4} \right) \exp \left( \frac{5}{\ell_r^2} \right) \Phi \left( -\frac{\sqrt{10}}{\ell_r} \right)
\]

and

\[ \xi = \prod_{r=q+1}^{p} \left( \frac{6\sqrt{5}\ell_r^2 - 20\sqrt{5}}{3\sqrt{\pi} \ell_r^3} + \left( 2 - \frac{40}{3\ell_r^2} + \frac{200}{3\ell_r^4} \right) \exp \left( \frac{5}{\ell_r^2} \right) \Phi \left( -\frac{\sqrt{10}}{\ell_r} \right) \right). \]

### 3.5.5 Example: borehole function

We revisit the borehole function described in Subsection 1.2.1 to illustrate the application of our quantification. We first fit a Gaussian process regression to the borehole function with all the eight input variables, using multiplicative one-dimensional covariance kernels. We transform the inputs to follow the standard Gaussian distribution, by applying the cumulative distribution function of the inputs and the inverse cumulative distribution function of the Gaussian distribution. Table 3.6 shows the maximum likelihood estimators of the length-scale hyperparameters that result from fitting the regression to 1000 data points. We observe that \( r_w, H_u, H_l, L, \) and \( K_w \) are the important variables in the borehole function, which matches with the sensitivity analysis result in Harper and Gupta (1983).

Let us consider dimension reduction through variable selection in this example. In this case, \( Q \) is a \( 8 \times 8 \) permutation matrix, and \( Q_1 \) consists of the first \( q \) columns of \( Q \). Suppose that an experimenter wants to fit a Gaussian process regression only on a set of selected input variables. Table 3.7 shows estimates of \( \xi \) for different sets of selected variables.
Table 3.6: Estimated length-scale hyperparameters $\ell$ in fitting Gaussian process regression to the borehole function, using 1000 function evaluations.

<table>
<thead>
<tr>
<th>Covariance</th>
<th>$r_w$</th>
<th>$H_u$</th>
<th>$H_l$</th>
<th>$L$</th>
<th>$K_w$</th>
<th>$r$</th>
<th>$T_u$</th>
<th>$T_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matérn 3/2</td>
<td>6.39</td>
<td>10.27</td>
<td>10.25</td>
<td>9.20</td>
<td>&gt;15</td>
<td>&gt;15</td>
<td>&gt;15</td>
<td>&gt;15</td>
</tr>
<tr>
<td>Matérn 5/2</td>
<td>6.09</td>
<td>6.88</td>
<td>6.88</td>
<td>6.34</td>
<td>8.99</td>
<td>&gt;15</td>
<td>&gt;15</td>
<td>&gt;15</td>
</tr>
<tr>
<td>Squared exponential</td>
<td>8.28</td>
<td>2.64</td>
<td>2.75</td>
<td>2.76</td>
<td>3.77</td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;10</td>
</tr>
</tbody>
</table>

We compare the heuristic estimate $\tilde{\xi}$ with the plug-in estimate $\hat{\xi}$ from the full Gaussian process model. We observe that the estimated proportion of variance in the Gaussian process varies over different covariance kernels. The proportion depends on the mean function and covariance function specified in the Gaussian process model. For example, based on $\hat{\xi}$, we can say that $r_w$ and $H_l$ explain 95.6\% of the variance in the Gaussian process with Matérn covariance $\nu = 3/2$.

The estimate of $\xi$ also varies over the type of the estimators $\hat{\xi}$ and $\tilde{\xi}$. As discussed in Section 3.5.3, $\hat{\xi}$ tends to overestimate $\xi$. Also, variance of the hyperparameter estimates used in $\hat{\xi}$ and $\tilde{\xi}$ can be very large; it has been shown that the likelihoods of the hyperparameters in the Gaussian process regression are flat. There are approaches in the literature (Li and Sudjianto, 2005) to mitigate the issue. Reducing the variance of the estimators of the hyperparameters can improve estimates of $\xi$. Table 3.7 shows the estimates of $\xi$ from Gaussian process regression with Matérn covariance illustrated in Figure 3.5.

Table 3.7: Proportional variance explained in the Gaussian process by variable selection in borehole function. $\hat{\xi}$ is the plug-in estimate from the original data, and $\tilde{\xi}$ is the heuristic estimate from the dimensionally reduced data.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Matérn 3/2</th>
<th>Matérn 5/2</th>
<th>Squared exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_w$, $H_l$</td>
<td>0.980</td>
<td>0.956</td>
<td>0.982</td>
</tr>
<tr>
<td>$r_w$, $H_u$, $H_l$</td>
<td>0.983</td>
<td>0.969</td>
<td>0.984</td>
</tr>
<tr>
<td>$r_w$, $H_u$, $H_l$, $L$</td>
<td>0.996</td>
<td>0.989</td>
<td>0.994</td>
</tr>
<tr>
<td>$r_w$, $H_u$, $H_l$, $L$, $K_w$</td>
<td>$\approx 1$</td>
<td>$\approx 1$</td>
<td>$\approx 1$</td>
</tr>
</tbody>
</table>
3.5.6 Discussion

We have translated the length-scale parameters in the Gaussian process regression to the importance of the variables, using the proportion of variance in the Gaussian process explained by the dimension reduction, $\xi$. We have also shown that modeling $f_{\text{proj}}$ as a Gaussian process with a multiplicative squared exponential covariance matches with our Gaussian process model on the original function.

In this work, we have focused on commonly used kernels, but one can compute $\xi$ for other kernels as well. Our analysis can be applied to any multiplicative covariance function.

Our analysis is developed for Gaussian inputs. The Gaussian distribution guarantees $Q_1^T \mathbf{x}$ and $Q_2^T \mathbf{x}$ to be independent, and also helps us quantify $\xi$ in a closed form. But our analysis can be applied to other input distributions. Especially in a variable selection case, when input variables are independent, then $Q_1^T \mathbf{x}$ and $Q_2^T \mathbf{x}$ are independent, and thus $\xi$ can be calculated based on the given input distribution.
Bibliography


