The Pennsylvania State University
The Graduate School

ROBUST PARAMETER DESIGN: A PENALIZED LIKELIHOOD APPROACH

A Dissertation in
Statistics
by
Kwame A. Kankam

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2014

The dissertation of Kwame A. Kankam was reviewed and approved* by the following:

James L. Rosenberger
Professor of Statistics
Dissertation Advisor and Chair of Committee

Dennis Lin
Professor of Statistics

Matthew Reimherr
Assistant Professor of Statistics

Lingzhou Xue
Assistant Professor of Statistics

Enrique Del Castillo
Professor of Industrial & Manufacturing Engineering and Statistics

David Hunter
Professor of Statistics
Head of Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Robust Parameter Design (RPD) is an engineering methodology aimed at designing quality into industrial products and processes by optimizing a quality characteristic with respect to the controllable input variables. It involves designing a system to withstand unavoidable variation while meeting its intended goal. Proposed and popularized by Japanese engineer and quality expert, Genichi Taguchi, it has entered the statistical mainstream and several approaches have been proposed. In the dual response RPD, response surfaces for the mean and the variance are obtained and depending on the goal of the experiment an objective function is determined. This is then optimized with respect to the control factors in order to find the best levels at which industrial processes should be carried out in order to obtain the best products.

In this dissertation, we propose a penalized likelihood approach to the dual response RPD which we call Adaptive Penalized Likelihood Effects Selection (APLES). We begin with a heteroscedastic linear model and specify a parametric variance function (specifically the log-linear variance function). We maximize the loglikelihood subject to constraints on the L_1 norms of the mean and variance parameter vectors in order to obtain simultaneous variable selection and estimation of the mean and variance parameters. For fixed values of the variance parameters, the problem reduces to a weighted least squares regression with an adaptive Lasso penalty. For fixed values of the mean parameters, the problem is equivalent to a gamma-error generalized linear model (GLM) with log link and an adaptive Lasso penalty. By iterating between these two minimization problems, we obtain the final set of APLES estimates for the non-zero mean and variance parameters. We describe and utilize the cyclic coordinate descent (CCD) algorithm which is very fast and has good convergence properties for the data sets that typically arise in applications.

We apply APLES to RPD and show using simulations that it has good perfor-

mance in a wide variety of settings for the mean parameters, variance parameters, noise variables, and different choices of 'quality' objective functions. We also illustrate the use of APLES by analyzing some well-known data sets from the literature. Another advantage of our approach is that it can be used for identification of location and dispersion effects in screening experiments. Screening experiments are the initial experiments carried out on a new process in which a potentially large number of factors are studied. Unlike RPD, the aim is not to optimize the subsequent response surfaces, but to select important variables for subsequent experimentation. We show that APLES performs equally well in these situations compared to other methods.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I thank God my maker, the alpha and the omega, 'for in Him I live and move and have my being'.

I thank my family (Dad, Yaw, Akua, Kweku, Efe and Kofi) for their love, support and prayers through this journey.

I thank Professor James Rosenberger for his guidance and friendship over these years.

I thank the members of my dissertation committee (Dr. Dennis Lin, Dr. Enrique Del Castillo, Dr. Matthew Reimherr and Dr. Lingzhou Xue) for their time and helpful suggestions to improve my dissertation.

I thank the Penn State Statistics department for being a great home for an international student.

I thank Bill and Barb Saxton and all other members of my Life Group for being such wonderful friends.

Dedication

To the memory of love, such as only a mother can give.

Till we meet again, Esther Boadiwah Kankam, I am always comforted by Philippians 1:20:

“ According to my earnest expectation and hope that in nothing I shall be ashamed, but with all boldness, as always, so now also Christ will be magnified in my body, whether by life or by death. ”

Chapter 1

Introduction

The problem of unwanted variation in products and processes in industrial settings is an important one which has been attacked on several fronts by both engineers and statisticians. New ideas for solving this problem and producing greater quality are always welcome to management and much effort is put into research and development to achieve this. One approach, which combines sound statistical practice as well as engineering knowledge, is known as robust parameter design (RPD). RPD is the name given to a group of techniques ranging from experimental designs to methods of analysis that were originally popularized in the 1980s by Japanese quality control expert, Genichi Taguchi. It is based on the idea of designing systems that meet their stated goal while being insensitive or robust to unavoidable variation.

A quality product meets certain targets as well as has very little variation around this target. Experimental design principles are used to discover which factors associated with the production lead to control of the mean so that the target can be met. These same principles can be used to discover those factors which affect the variation of the response, so that they can be set at the optimal levels for minimum variance.

We first review the nature of the problem, and then present a survey of the key literature, both classical and recent. In Chapters 3 - 5, we propose several techniques for improving these methods. In Chapter 6, we summarize our contributions.

1.1 Notation, Model and Setting

Let $\mathbf{X}$ be an $n \times p$ matrix of regressors resulting from an experiment. $\mathbf{X}$ can be as general as possible and in Chapter 3, we develop our approach under this general setting. $\mathbf{X}$ may result from an RPD experiment, in which case the regressors will consist of control factors, noise factors as well as possible interactions between them. We assume this in Chapter 4. In Chapter 5, the experiments will correspond to a two-level orthogonal array, for example a fractional factorial design.

Let the $n \times 1$ vector $\mathbf{y} = (y_1, \ldots, y_n)'$ consist of observations of a performance characteristic corresponding to the n runs of the experiment. Let $\boldsymbol{\epsilon} = \epsilon_1, \ldots, \epsilon_n$ be an $n \times 1$ vector of unobserved errors. They are independent and have zero mean. They do not generally have the same variance however. Thus $\mathrm{Var}(\epsilon_i) = \sigma_i^2$. Generally, these will be a function of the design matrix $\mathbf{X}$.

The joint location and dispersion model that we assume for the observations is

$$y_i = \mu_i + \epsilon_i \qquad i = 1, ..., n$$

where $\epsilon_i \sim N(0, \sigma_i^2)$.
The mean submodel is

$$\mu_i = \mathbf{x}_i' \boldsymbol{\beta} \tag{1.1}$$

and so the mean of the response lies in the linear space spanned by the columns of $\mathbf{X}$. The elements $(\beta_0, \beta_1, ..., \beta_{p-1})$ of $\boldsymbol{\beta}$ will henceforth be called mean or location parameters.

In effect we have the usual heteroscedastic linear model, which reduces to the homoscedastic case when the σ_i are all equal. We will consider the following log-linear variance submodel:

$$\log(\sigma_i^2) = \mathbf{x}_i' \boldsymbol{\tau}$$

The elements $(\tau_0, \tau_1, \ldots \tau_{p-1})$ of $\boldsymbol{\tau}$ will henceforth be called variance or dispersion parameters.

We say that the jth regressor corresponding to column j of $\mathbf{X}$ has an active location effect if the mean of $\mathbf{y}$ is different for different values of that column. This

is equivalent to saying that $\beta_j \neq 0$. Columns without active location effects will be said to have nonactive or inert location effects.

We say that the jth regressor corresponding to column j of $\mathbf{X}$ has an active dispersion effect if the variance of $\mathbf{y}$ is different for different values of that column. This is equivalent to saying that $\tau_j \neq 0$. Columns without active dispersion effects will be said to have nonactive or inert dispersion effects.

Let us denote the matrix of column vectors of true location effects by $\mathbf{X}_{\mathcal{L}}$ and the corresponding mean parameters vector by $\boldsymbol{\beta}_{\mathcal{L}}$. $\boldsymbol{\beta}_{\mathcal{L}}$ is unknown and interest lies in identifying its members.

After identification, the model is given by

$$\mathbf{y} = \mathbf{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}} + \boldsymbol{\epsilon}$$

and interest lies in obtaining estimates $\hat{\boldsymbol{\beta}}_{\mathcal{L}}$ of $\boldsymbol{\beta}_{\mathcal{L}}$.

Similarly, denote the matrix of column vectors of true dispersion effects by $\mathbf{X}_{\mathcal{D}}$ and the corresponding dispersion parameters vector by $\boldsymbol{\tau}_{\mathcal{D}}$. $\boldsymbol{\tau}_{\mathcal{D}}$ is unknown and interest lies in identifying its members.

Methods for identification of location and dispersion effects rely heavily on the assumption of effect sparsity. This allows the procedures to seek a parsimonious model. For example, before identification of location effects, we begin with a vector $\boldsymbol{\beta}$ of possible location parameters. However, only a small subset of the elements of $\boldsymbol{\beta}$ will be non-zero. Similarly, before identification of dispersion effects, we begin with a vector $\boldsymbol{\tau}$ of possible dispersion parameters but only a small subset of its elements will be non-zero.

In all of our subsequent work, we assume that the variance (or heteroscedasticity) is not a function of the mean.

1.2 Dissertation Topics

1.2.1 Adaptive Penalized Likelihood Effects Selection (APLES)

We develop a novel approach to simultaneously identify and estimate the active location and dispersion effects with the aim of obtaining the dual response surface for implementation of RPD. Variable selection in the classical linear model has usually been performed by choosing the model whose parameter estimates minimize (or maximize) a specific criterion function. These methods work well when the number of potential models is modest. When the number of contending models is large, it is out of the question to consider searching the entire space and this renders them impractical.

Recent research in high-dimensional data analysis over the last two decades has led to the development of penalized regression techniques. These techniques are able to obtain a parsimonious and interpretable subset of the variables under consideration by continuously shrinking the sizes of the coefficients to zero (based on the value of a regularization or tuning parameter). Thus they achieve simultaneous variable selection and estimation. However, it is not enough to have a sparse model. A model may be sparse and yet exclude the active (or truly non-zero predictors) in which case it will have poor predictive accuracy. One very popular approach, known as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), is to minimize the residual sum of squares subject to bounds on the L_1 norm of $\boldsymbol{\beta}$. This method can produce a sparse model. However, the coefficient sizes may be shrunk to an extent which leads to biased estimates. The adaptive LASSO (Zou, 2006) applies adaptive (or weighted and data-dependent) penalties and has several desirable properties.

Penalization can also be applied to likelihood functions in order to obtain penalized likelihood estimators. These are usually done in the case of non-normal models, specifically Generalized Linear Models (GLMs).

Our first goal in this dissertation is the proposal and implementation of a method we call Adaptive Penalized Likelihood Effects Selection (APLES). This maximizes the likelihood of the data subject to the application of adaptive LASSO penalties on the mean parameters as well as the variance parameters. We obtain the coefficient estimates by alternating between two cyclic coordinate descent

(CCD) sub-algorithms. We specify the theoretical and sampling properties of our proposed approach.

Our second goal is to justify the use of the proposed APLES estimation procedure to the analysis of designed experiments. To achieve this goal, we apply APLES to experimental data arising from the RPD context as well as to unreplicated fractional factorial experiments. In applying APLES to RPD, we show that it performs well especially in the heteroscedastic RPD framework. For unreplicated fractional experiments, we show that APLES performs well in screening for the active location and dispersion effects.

1.2.2 Analysis of Robust Parameter Design Experiments using APLES

The unconditional variance of the response from an RPD experiment with noise factors $\mathbf{z}$ is $\text{Var}\ (y_i) = \text{Var}\ [\ \text{E}\ (Y_i|z_i)] + \text{E}\ [\ \text{Var}\ (Y_i|z_i)]$. The first term is the variance of the conditional mean. This term is important when there are significant interactions between some of the noise factors and control factors. It may be possible in that case to set these control factors to certain levels that minimize the variance of the response, subject to meeting the mean target. If we can identify important noise factors for which we know the distribution up to second order terms, then by applying the variance operator to the conditional mean we may achieve variance reduction through the first term. The second term is the expectation of the conditional variance. Most approaches to dual RPD assume that the conditional variance is constant and so they estimate it using the mean square error of the least squares regression fit. However if there are important noise factors that are not included in the experiment, it may happen that there still remains substantial heteroscedasticity.

We apply the newly developed technique (APLES) of Chapter 3 to analyze RPD experiments. We consider the case where there are no noise variables first. Next we consider the situation with noise variables. In both situations, our method is superior in terms of the power to detect true location and dispersion effects, as well as the mean square error of the estimated coefficient vectors. We also show that when the prior estimates of the variance of the noise variables obtained

from previous experiments or process expertise are incorrect, our method is again superior.

1.2.3 Analysis of Unreplicated Fractional Factorials using APLES

Identification of location and dispersion effects in orthogonal array experiments is known as screening and has a long and well established history in industry. We compare our proposed method to the vast array of methods in the literature: Box and Meyer (1986b), Bergman and Hynen (1997), Wang (1989), Harvey (1976), McGrath and Lin (2001b), Brenneman and Nair (2001). We show that our method has a better performance in terms of the statistical power to detect effects when they are present while controlling the type I error rate.

1.3 Dissertation Research Objectives

One key aspect of the dual response RPD is how closely the estimated response surfaces approximate to the true response surfaces. Therefore, the issue of which control factors to include in the model is very important. This makes variable selection a very important step. Misspecification of the model will lead to a poor approximation to the true mean and variance models and thus we will be optimizing the wrong function.

An interesting aspect is that with two groups of models how one selects the parameters of the mean model affects the variance model that is selected, and vice versa. Thus, when estimating the mean model if we use an incorrect or inconsistent variance model the mean model will be inefficient. On the other hand if we incorrectly specify the mean model, it will have adverse effects on the variance model because the residuals will be incorrect.

In this dissertation, our primary focus is to first describe a method for simultaneously selecting the location and dispersion models as well as estimating them. We show that our method is both variable selection consistent as well as estimation consistent. It possesses the oracle property (Fan and Li, 2001) in both the location as well as the variance models. We call our method adaptive penalized likelihood effects selection (APLES). It is based on the penalized likelihood and the specific form of the penalization is the adaptive LASSO.

1.4 Dissertation Outline

This dissertation is organized as follows. Chapter 2 is a review of the literature. It is made up of three main parts. RPD provides the source of our problem. Given a heteroscedastic response function how to identify the mean and variance response functions as well as optimize these to select the best levels of the control factors. Modern penalized regression and penalized likelihood techniques (primarily developed for high dimensional situations) are the area from which we draw to solve the RPD problem. Lastly, we review the literature on identification of effects in screening experiments because these also provide a fertile area of application for our proposed method.

In Chapter 3, we first describe our method for the general regression context without any special properties of the design matrix. We motivate and describe the APLES method, and then describe algorithms for obtaining the APLES estimates. Finally, we discuss the theoretical and sampling properties of the APLES procedure.

In Chapter 4, we apply the APLES method specifically to RPD and show where it applies to the general RPD framework. We compare its performance when important noise variables are included in the experiment and when they are excluded, by simulation.

In Chapter 5, we apply APLES to the analysis of screening experiments, specifically unreplicated fractional factorials and show that it performs favorably in comparison to current methods that exist in the literature.

Finally, in Chapter 6, we discuss our contributions as well as some potential future directions.

Chapter 2

Literature Review

2.1 Overview

The following literature review covers three broad areas of Statistics. Section 2.2 generally covers the statistical methods that have been developed for reducing the variance of a polynomial response where the response is a quality characteristic of interest from an industrial process, and specifically it discusses Robust Parameter Design (RPD). Section 2.3 covers location and dispersion effects in screening experiments. Section 2.4 reviews some recent approaches to variable selection through penalization techniques. In chapter 3, we propose a method for identifying and estimating location and dispersion effects which relies on these variable selection techniques.

2.2 Robust Parameter Design (RPD)

Industrial experiments are performed to identify the effects that certain variables or factors have on the mean and variance of a quality characteristic. A quality characteristic is any response of interest to the engineer or scientist which measures the quality of a product or process. Let y be a random variable that represents the response from the process at given settings of various factors. There is usually a target value for y which will be denoted T. Even if the expected value of y happens to equal the target T, y will not always be equal to this target but will vary randomly around it, due to the natural variability in the response y. While meeting the target on average is very important, it is equally important that the response has as little variability about the target as possible.

Different levels of the input variables or factors result in different means and different variances for the response. Procedures for bringing the process on target have been studied extensively and they usually involve experiments in which a model for the mean of the response is related to the variables in the manufacturing process. This way we are able to find the settings of the variables in order to attain the optimum mean response.

An engineering or industrial process is a system where the factors are the inputs to the system and the response is the output from the system. An experiment is performed in which the levels of these factors are varied. There are several ways to characterize the factors in an industrial experiment. Taguchi (1986) identified two kinds of factors, **control** factors and **noise** factors. The difference between these is not just a statistical issue, but primarily it requires engineering or process knowledge as well.

Noise factors are impossible or too expensive to control during ongoing production, but they can be set at fixed levels in an experiment. Thus, the noise factors are modeled as random variables and their distributions are usually assumed known at least up to the the second moments. These moments are usually determined by engineers or scientists with knowledge of the process. Thus in the experiment, the observed values of the response are conditional on the values of the noise variables. The control factors are those whose levels can be modified both in the experiment and during the actual process. **Target control factors** are those factors whose settings can be modified to bring the mean of the process response on target, and **variability control factors** are those factors whose settings are modified to reduce the variance of the process response.

RPD is based on a set of principles for reducing variation in industrial experiments which exploits the fact that most of the variation may be due to one or more uncontrollable noise factors. Even though the noise factors cannot be controlled when the process is online, it is assumed that they can be controlled in an experiment. An **RPD experiment** varies the levels of some or all of the known noise factors together with the other factors. The goal of an RPD experiment is to seek the settings of the control factors which result in the minimum variability in the response, while the mean of the response is as close as possible to the target in a clearly defined sense. Taguchi (1987) first formulated this problem and suggested

a loss model approach to solving it. Robinson et al. (2004) provides an extensive review of RPD.

More generally, there are industrial experiments which do not contain noise factors that can be controlled in an experimental setting. It is still possible to use statistical methods to identify optimum or ideal settings of the control factors to ensure minimal variation and closeness to target. Strictly speaking, these are not RPD experiments in the sense of Taguchi (1987).

Suppose $\mathbf{z}$ denotes the noise factors and $\mathbf{x}$ denotes the control factors in an RPD experiment. The mean and variance responses resulting from an RPD experiment are called the conditional mean and conditional variance respectively and are written as: $E(y|\mathbf{x}, \mathbf{z})$ and $Var(y|\mathbf{x}, \mathbf{z})$. Assuming $\mathbf{z}$ varies randomly after running the experiment, we obtain the unconditional responses by using the expressions:

$$E(Y) = E_{\mathbf{z}}(E(y|\mathbf{x}, \mathbf{z})) \tag{2.1}$$

and

$$Var(y) = Var_{\mathbf{z}}(E(y|\mathbf{x}, \mathbf{z})) + E_{\mathbf{z}}(Var(y|\mathbf{x}, \mathbf{z})). \tag{2.2}$$

The various procedures for reducing variation in the literature may be broadly placed in three categories and equation (2.2) illustrates these. If we assume that $Var(y|\mathbf{x}, \mathbf{z})$ is constant, then the only opportunity for variance reduction is by means of the noise variables. Taguchi methods as well as the response surface approach (in the form originally proposed, e.g. Myers et al. (1992), Vining and Myers (1990)) make this assumption. We review Taguchi methods as well as the response surface approach in section 2.2. Using these approaches the control factors that are found to provide an opportunity for reducing the variance of the response are called the variability control factors.

If there are no noise variables present in the experiment, the only opportunity we have to determine the factors that affect the variance is through factors with **dispersion effects**. This second broad approach is called variance modeling (or dispersion effects modeling). In this approach we write a model for the variance of the response as a function of the control factors in order to find the factors that explain the heteroscedasticity in the response. Harvey (1976) and others

have previously written about this in the Econometrics literature. Box and Meyer (1986b) also proposed a method that applies specifically to orthogonal arrays lacking replication. Subsequently, several methods for identifying dispersion effects in unreplicated fractional factorial designs have been proposed. A review of methods for the identification of dispersion effects in unreplicated fractional factorial experiments is provided by Brenneman and Nair (2001) and discussed in section 2.4. Davidian and Carroll (1987) is also a very important paper that presents a unified view of variance function estimation in linear models in the general case, i.e. not restricted to orthogonal arrays.

This variance modeling approach is independent of RPD in the sense that for an experiment to be analyzed using this approach, we do not require any of the factors in the experiment to be noise variables. However, in RPD as originally formulated by Taguchi, the key principle is the use of noise variables.

The common theme in both these approaches is the fact that we aim to ascribe the heteroscedasticity in the model to various control factors, so that we can set them to their optimal levels in order to reduce variation in the response. The Taguchi approach and the response surface approach (as originally proposed) assumed homoscesdastic residual errors and relied on the fact that heteroscedasticity would manifest itself when analyzing the signal-to-noise ratio (SNR) or by the presence of interactions between noise and control factors in the conditional model. Thus a key to the success of these approaches is that all the major noise factors be included in the experiment. The variance modeling approach however does not require the pre-identification of noise variables, but explicitly postulates a model for the variance in terms of the factors (control) in the experiment. It is hoped that by obtaining a model for the variance, the few factors (assuming sparsity) that appear in this model will be the ones that do affect the variability and so we may set them at their optimal levels in order to obtain minimum variance.

The third broad approach is found in Engel and Huele (1996b) who combine the above two approaches. In their model, we have an RPD experiment with noise factors and so we proceed similarly as in the response model approach. However, we do not assume that $Var(y|\mathbf{x}, \mathbf{z})$ is a constant, but rather assume that it is dependent on the levels of the control factors. We discuss this approach in section 2.2.3.

2.2.1 Taguchi's Signal-to-Noise Ratio (SNR)

Taguchi's contributions (Taguchi 1980; 1986) to RPD include proposing techniques of experimental design as well as techniques of analysis for the data arising from the experiments. His designs are product or cross arrays where we cross an orthogonal array called the inner or control array with another orthogonal array called the outer or noise array.

As an example of a cross array, Wu and Hamada (2000) (page 437) describe a silicon layer growth experiment with eight control factors (A-H) and two noise factors (L and M) which was first reported by Kackar and Shoemaker (1986). The growth of uniform layers of silicon on top of silicon wafers is an early step in the fabrication of integrated-circuits. The process is carried out on a device called a susceptor which has a certain number of facets (or sides), each with a top and bottom position. The two noise factors considered are the location (top or bottom) of the wafer on the susceptor, and the facet (there were four of these). The experiment crosses a 2^{8-4}_{IV} design for the control factors with a 2×4 design for the noise factors, resulting in eight observations for each of the 16 control factor settings. The experimental layout is shown in Table 2.1.

Table 2.1. Example of Cross Array, where factors A, B, C, E create a full factorial, and generators D=ABC, F=ABE, G=ACE, and H=BCE complete the 2^{8-4} design in the control factors.

								Noise Factor							
Control Factor								L-Bottom				L-top			
A	B	C	D	E	F	G	H	M-1	M-2	M-3	M-4	M-1	M-2	M-3	M-4
-	-	-	+	-	-	-	-	14.2908	14.1924	14.2714	14.1876	15.3182	15.4279	15.2657	15.4056
-	-	-	+	+	+	+	+	14.8030	14.7193	14.6960	14.7635	14.9306	14.8954	14.9210	15.1349
-	-	+	-	-	-	+	+	13.8793	13.9213	13.8532	14.0849	14.0121	13.9386	14.2118	14.0789
-	-	+	-	+	+	-	-	13.4504	13.4788	13.5878	13.5167	14.2444	14.2573	14.3951	14.3724
-	+	-	-	-	+	-	+	14.1736	14.0306	14.1398	14.0796	14.1492	14.1654	14.1487	14.2765
-	+	-	-	+	-	+	-	13.2539	13.3338	13.1920	13.4430	14.2204	14.3028	14.2689	14.4104
-	+	+	+	-	+	+	-	14.0623	14.0888	14.1766	14.0528	15.2969	15.5209	15.4200	15.2077
-	+	+	+	+	-	-	+	14.3068	14.4055	14.6780	14.5811	15.0100	15.0618	15.5724	15.4668
+	-	-	-	-	+	+	-	13.7259	13.2934	12.6502	13.2666	14.9039	14.7952	14.1886	14.6254
+	-	-	-	+	-	-	+	13.8953	14.5597	14.4492	13.7064	13.7546	14.3229	14.2224	13.8209
+	-	+	+	-	+	-	+	14.2201	14.3974	15.2757	15.0363	14.1936	14.4295	15.5537	15.2200
+	-	+	+	+	-	+	-	13.5228	13.5828	14.2822	13.8449	14.5640	14.4670	15.2293	15.1099
+	+	-	+	-	-	+	+	14.5335	14.2492	14.6701	15.2799	14.7437	14.1827	14.9695	15.5484
+	+	-	+	+	+	-	-	14.5676	14.0310	13.7099	14.6375	15.8717	15.2239	14.9700	16.0001
+	+	+	-	-	-	-	-	12.9012	12.7071	13.1484	13.8940	14.2537	13.8368	14.1332	15.1681
+	+	+	-	+	+	+	+	13.9532	14.0830	14.1119	13.5963	13.8136	14.0745	14.4313	13.6862

For analysis, Taguchi combines the mean and variance into a single performance measure, the SNR. The SNR is chosen based on the type of RPD problem. He distinguishes three types of RPD problems based on the goal of the experiment. These are Nominal-the-Best, Larger-the-Better and Smaller-the-Better problems. The analysis involves relating the SNR to main effects of the control variables. The performance measure, SNR, is chosen with the hope that its relationship with the control factors has minimal control factor interactions. i.e. it should have a monotonic relationship with the control variables.

Taguchi assumes that the set of control factors can be written as: $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ where $\mathbf{x}_1$ are those control factors that affect both the mean and the variance, whereas $\mathbf{x}_2$ are those that affect only the mean. This second group of factors that are found not to affect the SNR are called adjustment factors. He proposes a two-step procedure (Phadke 1989; Phadke and Taguchi 1987) to minimize the quadratic loss function:

$$E(y - T)^2 = Var(y(\mathbf{x}_1, \mathbf{z})) + [E(y(\mathbf{x}_1, \mathbf{x}_2)) - T)]^2.$$

First use the set of factors $\mathbf{x}_1$ to minimize the variance and then the adjustment factors $\mathbf{x}_2$ can be used to adjust the quality characteristic to its target value. The assumption is that this second step can be performed without affecting the variability.

Criticism of Taguchi's approach focuses on the use of SNR as a summary measure, the methods of analysis as well as the choice of experimental design. Leon et al. (1987) show that there are situations where the relationship between the response and the factors is such that SNR is not an appropriate performance measure. They define the concept of performance measures independent of adjustment (PerMIA). However, several authors have noted problems with this approach and there are instances where the structure of the mean-variance relationship implies that the goal cannot be achieved. Instead, other approaches are suggested such as data transformations. Criticism of Taguchi's methods are contained in several papers including Nair and Pregibon (1988) and Box (1988). Berube and Nair (1998) perform extensive studies to determine the situations in which the SNR is

an appropriate measure and situations when it performs poorly. Nair et al. (1992) organized a panel of quality and statistical experts.

The remaining sections contain other methods and approaches that have been suggested as alternatives to the Taguchi approach and do not suffer from the same criticisms.

2.2.2 Response Surface Approach to RPD

Response Surface Methodology (RSM) (Box and Wilson, 1951) has been applied to the analysis of RPD experiments. The purpose of RSM is to understand the relationship between a response and the various input factors by performing sequential experiments with the aim of approximating the response surface relationship with a fitted low order regression model (usually this regression model is at most of second order). Once a fitted response surface is obtained it can be optimized with respect to the input factors. Several experimental designs with very good properties have been developed specifically for this purpose and these are called response surface designs.

In the dual response model approach to RPD, we fit a mean model and a variance model. This approach was proposed by Vining and Myers (1990) using the Dual Response Surface method of Myers and Carter (1973). One of the responses is designated the secondary response and the other response is called the primary response. These decisions are made based on the goal of the RPD experiment. Estimated response surfaces $\widehat{E(y)}$ and $\widehat{Var(y)}$ are fitted for the mean and the variance respectively. The primary response is minimized with respect to $\mathbf{x}$ subject to certain constraints on the secondary response. For example, where we wish to minimize the variance subject to meeting the target, we would minimize $\widehat{Var(y)}$ subject to $\widehat{E(y)} = T$. In Vining and Myers (1990), it is assumed that the experiment is from a Taguchi cross array and they obtain the fitted mean from the sample means at each control array setting and the fitted variance from the sample variance at each control array setting. Del Castillo and Montgomery (1993), Lin and Tu (1995), Kim and Lin (1998), Del Castillo et al. (1997), Tang and Xu (2002), Koksoy and Doganaksoy (2003) have proposed improvements to the optimization step of this analysis.

Several authors observed that the Taguchi cross-arrays are expensive because they require a high number of experimental runs. An alternative is to use a single array containing both the control and noise factors. This is called a **combined** array. Shoemaker et al. (1991) investigated the savings in the number of runs that can be obtained by using the combined array instead of the product array. Wu and Hamada (2000) provide guidelines for choosing between the two types of experimental format based on estimation capacity and effect ordering. They define estimation capacity as the overall performance of an array in estimating different effects such as the number of effects that are eligible, clear or strongly clear.

RSM can be used to analyze the results of an experiment from a combined array as well. The ideas for this are present in Welch et al. (1990) and formalized in Myers et al. (1992). The remaining portion of this subsection follows the discussion in chapter 9 of Del Castillo (2007)

Consider a combined array RPD experiment with k controllable factors and r noise factors, which are given by

$$\mathbf{x}' = (x_1, x_2, ..., x_k) \text{ and}$$
$$\mathbf{z}' = (z_1, z_2, ..., z_r) \text{ respectively.}$$

We fit the usual response model given by:

$$y_i(\mathbf{x}_i, \mathbf{z}_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i + \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\Delta}\mathbf{z}_i + \epsilon_i \tag{2.3}$$

for $i = 1, ..., n$, where:

$y_i(\mathbf{x}_i, \mathbf{z}_i)$ is the observed response for a fixed setting of the control variables $\mathbf{x}_i$ and a fixed setting of the noise variables $\mathbf{z}_i$.

$\beta_0, \boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\gamma}$, and $\boldsymbol{\Delta}$ are the model parameters.

$\{\epsilon_i\}$ is a set of independently and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ_ϵ^2, where σ_ϵ^2 is a constant.

Specifically,

β_0 is the intercept

$\boldsymbol{\beta}$ is the $k \times 1$ vector of control factor coefficients

$\mathbf{B}$ is the $k \times k$ matrix of second order coefficients

$\boldsymbol{\gamma}$ is the $r \times 1$ vector of noise factor coefficients

$\boldsymbol{\Delta}$ is the $k \times r$ matrix of control by noise interaction coefficients

There are no second order terms involving the noise factors in model (2.3). However, this extension is considered in Box and Jones (1990), Myers (1991), Box and Jones (1992) and Engel and Huele (1996a).

For the noise variables, it is assumed that $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{z}})$ where,

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \begin{pmatrix} \sigma^2_{Z_1} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & \sigma^2_{Z_r} \end{pmatrix}$$

is the (diagonal) covariance matrix of $\mathbf{z}$, which is assumed to be known.
Conditional on the the noise variables, the mean and variance are given by:

$$E_\epsilon(y(\mathbf{x}, \mathbf{z})|\mathbf{z}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{z}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\Delta}\mathbf{z} \tag{2.4}$$

$$Var_\epsilon(y(\mathbf{x}, \mathbf{z})|\mathbf{z}) = \sigma^2_\epsilon \tag{2.5}$$

The unconditional mean and variance are given by:

$$E_{\mathbf{z},\epsilon}(y(\mathbf{x}, \mathbf{z})) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} \tag{2.6}$$

$$Var_{\mathbf{z},\epsilon}(Y(\mathbf{x}, \mathbf{z})) = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}_i)'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}_i) + \sigma^2_\epsilon \tag{2.7}$$

$\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}_i$ is the vector of partial derivatives of $Y(\mathbf{x}, \mathbf{z})$ with respect to the noise variables and it plays a very important role in RPD.

The fitted model by least squares is:

$$\hat{y}\,(\mathbf{x}, \mathbf{z}) = \hat{\beta}_0 + \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{x}_i'\hat{\mathbf{B}}\mathbf{x} + \mathbf{z}'\hat{\boldsymbol{\gamma}} + \mathbf{x}'\hat{\boldsymbol{\Delta}}\mathbf{z} \tag{2.8}$$

By plugging the least squares estimates of the parameters from (2.8) into (2.6),

the estimated mean response surface is

$$\widehat{E}_{\mathbf{z},\epsilon}\left[Y(\mathbf{x},\mathbf{z})\right] = \hat{\beta}_0 + \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{x}_i'\hat{\mathbf{B}}\mathbf{x} \tag{2.9}$$

Similarly, by plugging the least squares estimates of the parameters from (2.8) into (2.7), the estimated variance response surface is

$$\widehat{Var}_{\mathbf{z},\epsilon}\left[y(\mathbf{x},\mathbf{z})\right] = (\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}_i)'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}_i) + \hat{\sigma}_\epsilon^2 \tag{2.10}$$

where $\hat{\sigma}_\epsilon^2$ is the residual mean square error from the fitted response.

Thus, we see that even though we began with a single model, we have obtained two response surfaces and we can apply dual response approach to obtain optimum settings of the control factors. The goal of RPD would then be, for example, to minimize the response variance while keeping the mean on a target, T. We could solve

$$\begin{aligned} &\min_{\mathbf{x}} \widehat{\mathrm{Var}}_{\mathbf{z},\epsilon} Y(\mathbf{x},\mathbf{z}) \\ &\text{subject to } a \leq \widehat{E}_{\mathbf{z},\epsilon}\left[Y(\mathbf{x},\mathbf{z})\right] \leq b \end{aligned}$$

for suitable bounds on the mean response $a \leq T \leq b$.

Borror et al. (2002) provide a systematic way of fitting the response.

Myers and Montgomery (2002) show that the estimated variance response surface given by (2.10) is biased. However, the unbiased estimates may produce negative estimates of variance. Miro-Quesada and Del Castillo (2004) observe that minimizing the estimated variance response surface, $\widehat{Var}_{\mathbf{z},\epsilon}\left[Y(\mathbf{x},\mathbf{z})\right]$, ignores the additional variance component due to the uncertainty in parameter estimates. Thus the need for an objective function that combines the noise factor variance with the parameter estimation variability. They propose that the objective function be the variance of the predicted response where the variance is taken with respect to the parameter estimates as well as the noise factors. They propose an unbiased estimate of the variance of the predictions which does not produce negative estimates of the variance.

2.2.3 Other Approaches

Engel and Huele (1996b) employ generalized linear models (GLMs) in the analysis of RPD experiments. We have data Y, k control factors and r noise factors. We write the single response model conditional on the levels of the noise factor exactly as in equation (2.3).

$$y_i\left(\mathbf{x}_i, \mathbf{z}_i\right) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i + \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\Delta}\mathbf{z}_i + \epsilon_i$$

The difference is that $\epsilon_i \sim N(0, \sigma_i^2)$, but σ_i^2 is not a constant. They model it as

$$\sigma_i^2 = exp(\mathbf{x}_i'\boldsymbol{\tau}) \tag{2.11}$$

But this is equivalent to saying that σ_i^2 is log-linearly related to the control factors. Note that σ_i^2 =Var$(Y_i|\mathbf{z})$, and therefore the Engel and Huele (1996b) approach to RPD can be described as follows. It is a dual response surface approach in which we do not assume that conditional variance is constant but rather model it as being log-linear in the control factors.

Thus the unconditional response surface for the variance is

$$Var(y_i) = Var(E(y_i|\mathbf{z}) + E(Var(y_i|\mathbf{z}) = (\delta' + g'(\mathbf{x}_i)\boldsymbol{\Delta})\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma}' + g'(\mathbf{x}_i)' + \sigma_i^2 \tag{2.12}$$

Engel and Huele (1996b) estimate the parameters $\boldsymbol{\tau}$ by fitting a GLM with gamma errors and log link, i.e. using the approach suggested by Aitkin (1987).

An alternative GLM approach to RPD is suggested in Nelder and Lee (1991), Lee and Nelder (1998) and Lee and Nelder (2003). However,this approach is not used further in this proposal.

Vining and Bohn (1998) provide a nonparametric approach and Pickle et al. (2008) introduce a semi-parametric approach to RPD. Simpson et al. (1998) compares second-order response surface models and kriging models for approximating non-random, deterministic computer analyses.

2.3 Location and Dispersion Effects in Unreplicated Fractional Factorial Experiments

2.3.1 Identification methods

In this section, we consider the identification of location and dispersion effects in screening experiments. Screening experiments in industry are carried out when a large number of factors are being considered as explanatory variables for a response. However, it is assumed that only a small subset of these factors actually affects the mean of the response. This assumption is called location effect sparsity Box and Meyer (1986a). In screening experiments with heteroscedasticity, we may also assume that the variance is only affected by a small subset of factors. This assumption is called dispersion effect sparsity (Box and Meyer, 1986b).

The 2^{s-q} fractional factorial design, consisting of a $1/2^q$ fraction of a full 2^s design, lends itself very well for screening experiments because a potentially large number of factors (with two levels) can be varied and their effects on a response studied. Sixteen-run fractional factorials are fairly common and several examples of experiments carried out abound in the literature. By their nature, they are very efficient and save costs in experimentation. However, their analysis usually requires a combination of statistical tools as well as process expertise. Confirmation and follow-up experiments should be carried out.

The issues pertaining to analyzing an unreplicated 2^{s-q} design are three-fold:

1. Identify the important (or active) location factors.

2. Identify the important (or active) dispersion factors.

3. Estimate the identified location and dispersion effects.

The classical approach to identifying active location effects is the probability plotting approach of Daniel (1959). Box and Meyer (1986a) proposed a Bayesian analysis which computes the posterior probability that a particular effect is active. Though computationally expensive, and requiring extensive prior knowledge of the particular process, it incited a spate of publications on frequentist approaches to the problem. Most of these are reviewed in Hamada and Balakrishnan (1998).

Among these, the one that is most frequently recommended is by Lenth (1989). This is because it is robust and adaptive. Voss and Wang (2006) present other robust and adaptive approaches to identifying active location effects.

Even though the literature on variance modeling in designed experiments dates back to Bartlett and Kendall (1946), Box and Meyer (1986b) were the first to consider the specific case of modeling variance in unreplicated fractional factorial designs. For unreplicated fractional factorials, there is no estimate of variation available at each design setting. In screening experiments, a large number of factors is considered and if we require replication for the analysis, then a very large number of runs will be required. This makes the development of methods for the unreplicated case very crucial. A common theme in all the methods proposed to deal with the unreplicated situation is their reliance on the residuals that are obtained after removal of location effects.

There are several challenges involved in the analysis of dispersion effects in unreplicated fractional factorial experiments. The residuals are not independent and are correlated depending on the number of location effects. In addition, the residuals are a function of the factors that we include in the location model. The residuals are exactly 0 if the model is saturated. Nair and Pregibon (1988) give a review of methods for identifying dispersion effects when we have replication. Reviews of methods for unreplicated experiments can be found in Brenneman and Nair (2001) and Bursztyn and Steinberg (2006). Several methods are well-known for the general regression situation and discussed in Davidian and Carroll (1987).

We describe next some of the important methods that have been proposed in the literature for modeling dispersion effects in unreplicated fractional factorials. Let x_{ij} denote the element in the $(i,j)th$ position of the design matrix. Let $r_i, i = 1 \ldots n$ denote the residuals from the fitted location model. Let $S_{j+} = \sum_{i:x_{ij}=+1} r_i^2$ be the sum of the squared residuals at all observation points for which the jth column of the design matrix is $+1$ and $S_{j-} = \sum_{i:x_{ij}=-1} r_i^2$ be the sum of the squared residuals at all observation points for which the jth column is -1.

Box and Meyer (1986b) proposed the statistic

$$D_j^{BM} = 0.5 \log \frac{S_{j+}}{S_{j-}} \tag{2.13}$$

for identifying a dispersion effect of column j. They did not provide the distribution of this statistic and in practice it is referred to a normal probability plot to determine which dispersion effects stand out. We will refer to the use of this statistic as the BM method.

Bergman and Hynen (1997) pointed out that the residuals obtained from estimating the location effects are dependent and this makes it hard to obtain the distribution of test statistics like D_j^{BM}. They proposed a new approach which we denote as the BH method. The BH method avoids the dependence in the residuals by fitting an expanded location model before forming the test statistic. Let $\tilde{r}_i, i = 1, \ldots n$ be the residuals from the expanded location model. Define $\tilde{S}_{j+} = \sum_{i:x_{ij}=+1} \tilde{r}_i^2$ and $\tilde{S}_{j-} = \sum_{i:x_{ij}=-1} \tilde{r}_i^2$. The Bergman-Hynen statistic for factor j is given by

$$D_j^{BH} = \frac{\tilde{S}_{j+}}{\tilde{S}_{j-}}$$

It has an exact reference distribution which is $F(\nu, \nu)$ with ν defined in Bergman and Hynen (1997). Brenneman and Nair (2001) showed that this exact reference distribution does not hold in general. The F distribution holds only under the restrictive condition that there are no dispersion effects in all columns of the design matrix, not just under the null hypothesis of no dispersion effect in the jth column being tested. Brenneman and Nair (2001) use the Satterwaithe approach to approximate the distribution of D_j^{BH} when this condition does not hold.

Wang (1989) proposed using the score test statistic derived by Cook and Weisberg (1983). The Wang statistic Wang (1989) for factor j is given by

$$D_j^W = \frac{1}{2n\hat{\sigma}^2} \left(S_{j+} - S_{j-}\right)^2 = \frac{1}{2n} \left(\frac{S_{j+} - S_{j-}}{S_{j+} + S_{j-}}\right)^2$$

where $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} r_i^2$

The Wang method (which we shall call the W method) was developed under a log-linear model for the variance. The test statistic D_j^W has a χ_1^2 (Chi-squared with 1 degree of freedom) reference distribution under the restrictive null hypothesis where all dispersion effects are assumed zero. Thus the actual type 1 error rates

may differ from the advertised ones (Brenneman and Nair, 2001).

Harvey (1976) also assumes a log-linear dispersion model and proposes

$$D_j^H = \frac{1}{n}\log\left(\frac{\prod_{i:x_{ij}=+1} r_i^2}{\prod_{i:x_{ij}=-1} r_i^2}\right)^{1/n}$$

as an estimator of the dispersion effect of column j. Brenneman and Nair (2001) propose a modified Harvey method because the Harvey method can give severely biased estimates for some dispersion effects. However, the modified Harvey method can also be biased when it coincides with the Harvey method. The modified Harvey method uses

$$D_j^{MH} = \frac{1}{n}\left(\frac{\prod_{i:x_{ij}=+1} \tilde{r}_i^2}{\prod_{i:x_{ij}=+1} \tilde{r}_i^2}\right)^{1/n}$$

as an estimator of the dispersion effect of column j. We will refer to the Harvey method as the H method and the modified Harvey as the MH method.

McGrath and Lin (2001b) proposed two methods, a parametric method and a non-parametric method. The McGrath and Lin (2001b) parametric method uses a test statistic which we shall denote D_j^{ML}. We will refer to this statistic as the ML method.

Liao (2000) (L method) considered the likelihood ratio test statistic given by:

$$D_j^L = \frac{n}{2}\,\log\,\left(\frac{1}{4S_{j+}}\cdot\frac{1}{4S_{j-}}\right)$$

Brenneman and Nair (2001) also propose a test statistic for dispersion effects under a linear structure.

Finally, it has been shown by McGrath and Lin (2001a) and Pan (1999) that if some active location effects are missed in the location model from which the residuals are computed, then all of these methods may spuriously detect some dispersion effects that do not actually exist. Thus the first step of identifying location effects is very important.

2.3.2 Estimation Methods

After using the assumptions of location and dispersion effect sparsity and an appropriate method to identify the active effects, the next step is to estimate them. The two methods of estimation recommended are maximum likelihood estimation (MLE) and **re**sidual (or **re**stricted) **m**aximum **l**ikelihood (REML).

Harvey (1976) derives the MLEs for the multiplicative or log-linear variance model using the Fisher scoring algorithm. The regression model with log-linear heteroscedasticity has been studied by several authors in Econometrics. Park (1968) first suggested a generalization of the practice of modeling the variance of the error term as proportional to the square of the independent variable, that one should rather assume a structure for the variance. He then suggested a two-step estimator for the parameters of the variance function.

Wang (1989) obtained an alternative way to compute the MLE when the design matrix is an orthogonal array, which uses the special orthogonal structure and reduces to fitting non-linear least squares.

Aitkin (1987) observes that a Fisher scoring algorithm for simultaneous estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ reduces to two separate algorithms for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. Using this, he shows how the MLEs can be obtained by iterating between models that are easily fitted by statistical software. For fixed $\boldsymbol{\tau}$, $\hat{\boldsymbol{\beta}}$ is a weighted least squares estimate and for fixed $\boldsymbol{\beta}$, $\hat{\boldsymbol{\tau}}$ is the MLE from a generalized linear model with gamma distributed errors and a log-link for the gamma mean.

Verbyla (1993) obtain REML estimators and compares them to the MLEs.

2.4 Penalized Likelihood for Variable Selection

Recently, in high-dimensional data analysis, methods have been developed for utilizing the sparse representation of the true vector of regression parameters. Under the constant variance assumption, variable selection by appropriate penalization of the likelihood has been found to have several desirable properties and a large number of such techniques exist. These include the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and the adaptive LASSO (Zou, 2006). The LASSO is a special case of the more general L_q-penalized or bridge estimators (Frank and Friedman, 1993). The SCAD, the adaptive LASSO and the bridge

estimators (for $0 < q < 1$) have been shown to possess the oracle property. This is a very desirable property to have because it ensures that we can correctly identify the truly non-zero parameters as well as estimate them with the optimal rate for large sample size. Thus it ensures that we can perform as well as if we knew the true sparse set.

A key aspect of the dual response RPD is how well the estimated response surfaces approximate the unknown response surfaces. An estimated response surface which is a poor local approximation to the true response surface will lead to incorrect settings for the control factors which do not meet the target or have excessive variability. There are two measures for assessing the closeness of the estimated response surfaces to the true ones. One measure is the predictive ability of estimated response surfaces. The other is the ability to correctly identify the true subset of the potentially large number of control factors which should be included in the model. This second measure is particularly important in screening experiments where the aim is mainly to identify the important factors for subsequent experimentation. In this situation, we are not primarily interested in optimizing the estimated response functions (mean and variance).

Traditional methods for variable selection include forward selection, backward elimination, and best subset selection. There is an extensive literature on these approaches. Comprehensive reviews can be found in Linhart and Zucchini (1986) and Miller (2002). These variable selection methods are used in conjunction with model selection criteria such as Mallow's Cp (Mallows, 1973), Akaike's Information Criterion (Akaike, 1974), and Bayesian Information Criterion (Schwarz, 1978).

The computational burden arising from these approaches make them impractical to use most of the time. To address these shortcomings, penalized regression was introduced.

2.4.1 Some Penalized Regression Procedures

Consider the linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$ with i.i.d. errors ϵ_i. Unlike ordinary least squares which performs an unconstrained minimization of the residual sum of squares, penalized regression methods minimize the residual sum of squares subject to bounds on the "size" of the coefficients. The type of penalty

function $p_\lambda(|\beta_j|)$ determines the nature and properties of the resulting estimator. Generally, a penalized regression estimator is given by

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{k} p_\lambda(|\beta_j|) \tag{2.14}$$

Suppose $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ has the sparse representation $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{A}^c})$. Here $\mathcal{A} = \{j : \beta_j \neq 0\}$ and the cardinality of $\mathcal{A}$ is smaller than p. An oracle, knowing the true support set $\mathcal{A}$ would solve

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\mathcal{A}}'\boldsymbol{\beta}_{\mathcal{A}})^2$$

An estimator is said to have the oracle property if it has the same asymptotic distribution as the oracle estimator (Fan and Li, 2001; Fan et al., 2004). Moreover, an estimator is said to have the strong oracle property if the estimator equals the oracle estimator with overwhelming probability (Fan and Lv, 2011).

Next, before we describe some popular penalized regression estimators, let us define the soft thresholding operator by

$$S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+ = \begin{cases} z - \lambda & \text{if } z > \lambda \\ z + \lambda & \text{if } z < \lambda \\ 0 & \text{if } \lambda \geq |z| \end{cases}$$

In words: Given z and λ, if z is small in absolute value relative to λ set it to 0. If it is large in absolute value then 'penalize' it by λ.

We now describe various penalty functions that have been proposed in the literature.

1. The LASSO estimator (Tibshirani, 1996), imposes an L_1 penalty ($p_\lambda(|\beta_j|) = \lambda \sum_{j=1}^{k} |\beta_j|$) on the size of regression coefficients. It sets small coefficients exactly to zero and shrinks large coefficients with the level of shrinkage determined by the tuning parameter, λ. Hence it selects variables and estimate coefficients simultaneously. The LASSO solves a convex optimization problem and has shown good empirical performance in various contexts, especially

for orthogonal predictors. However, it has less than optimal estimation properties because it can be biased for truly large coefficients and it does not possess variable selection consistency for certain design matrix structures (Zou, 2006). Variable selection consistency means that as the sample size gets large, the set of non-zero coefficients that are selected approaches the true set of non-zero coefficients with probability approaching 1.

For uncorrelated predictors, and using the LASSO penalty, the solution to (2.14) is given by soft thresholding the ordinary least squares (OLS) estimate

$$\hat{\beta}_j^{LASSO}(\lambda) = S(\hat{\beta}_j^{OLS}, \lambda) = \text{sign}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda)_+ \tag{2.15}$$

where $\hat{\beta}_j^{OLS}$ is the ordinary least squares estimate for β_j. This formula applies for design matrices with orthogonal columns.

2. The adaptive LASSO (Zou, 2006) imposes a weighted L_1 penalty on the regression coefficients. The adaptive LASSO estimator possesses theoretical properties like model selection consistency and oracle properties (Fan and Li, 2001), and its convex formulation makes the solution straightforward. The LASSO applies the same penalty to each coefficient whereas the adaptive LASSO allows the possibility of different weights. By a careful choice of the weights, it is able to achieve the oracle properties. The adaptive LASSO is given by

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} w_j|\beta_j|$$

The weights are chosen as $w_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$ where $\gamma > 0$ and $\tilde{\beta}$ is a pre-determined $\sqrt{n}$- consistent estimator of β, such as the least squares estimator. For the rest of our discussion, we shall take $\gamma = 1$.

3. The SCAD estimator proposed by Fan and Li (2001). Unlike the LASSO and adaptive LASSO, it does not solve a convex minimization problem. It is shown in Fan and Li (2001) that the SCAD possesses the oracle property.

The SCAD penalty function is given by

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda, \\ \frac{-(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda, \end{cases}$$

It is non-convex and non-differentiable at 0. The derivative is given by:

$$p'_\lambda(|\beta_j|) = \text{ sign } (|\beta_j|) \left(\lambda I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{a-1} I(|\beta_j| > \lambda) \right)$$

In the above two expressions, the value of a is usually taken to be 3.7.

4. The L_0 penalty is defined as

$$p_\lambda(u) = \lambda I(|\beta_j| \neq 0) = \begin{cases} \lambda & \text{if } |\beta_j| \neq 0, \\ 0 & \text{if } |\beta_j| = 0, \end{cases}$$

The L_0 penalty directly penalize the non-zero parameter estimates. Most classical variable selection procedures such as C_p (Mallows, 1973), AIC (Akaike, 1974) and BIC (Schwarz, 1978) correspond to using the L_0 penalty. The drawbacks to using this penalty are computational expense and high instability of the estimators due to the discontinuity of the L_0 penalty at 0. The action of the L_0 penalty is known as hard thresholding in the literature.

Each of these penalties is plotted in figure 2.1 with the exception of the adaptive LASSO penalty. The penalty varies with the weights w_j and for each fixed w_j it is similar to the LASSO penalty with a different slope. In each of these plots, the horizontal axis shows values of the parameter (β) and the vertical axis shows values of the penalty function ($p_\lambda(\beta)$).

Each of the four estimators resulting from the four penalties described above can be thought of as a "thresholding rule". In the case of a single regression coefficient, the LASSO estimator can be written as a function of the (unpenalized) least squares estimator. The action of the LASSO operator on the least squares estimator is known as soft thresholding and is illustrated in the upper left plot of figure 2.2. The thresholding action of the adaptive LASSO, SCAD and L_0

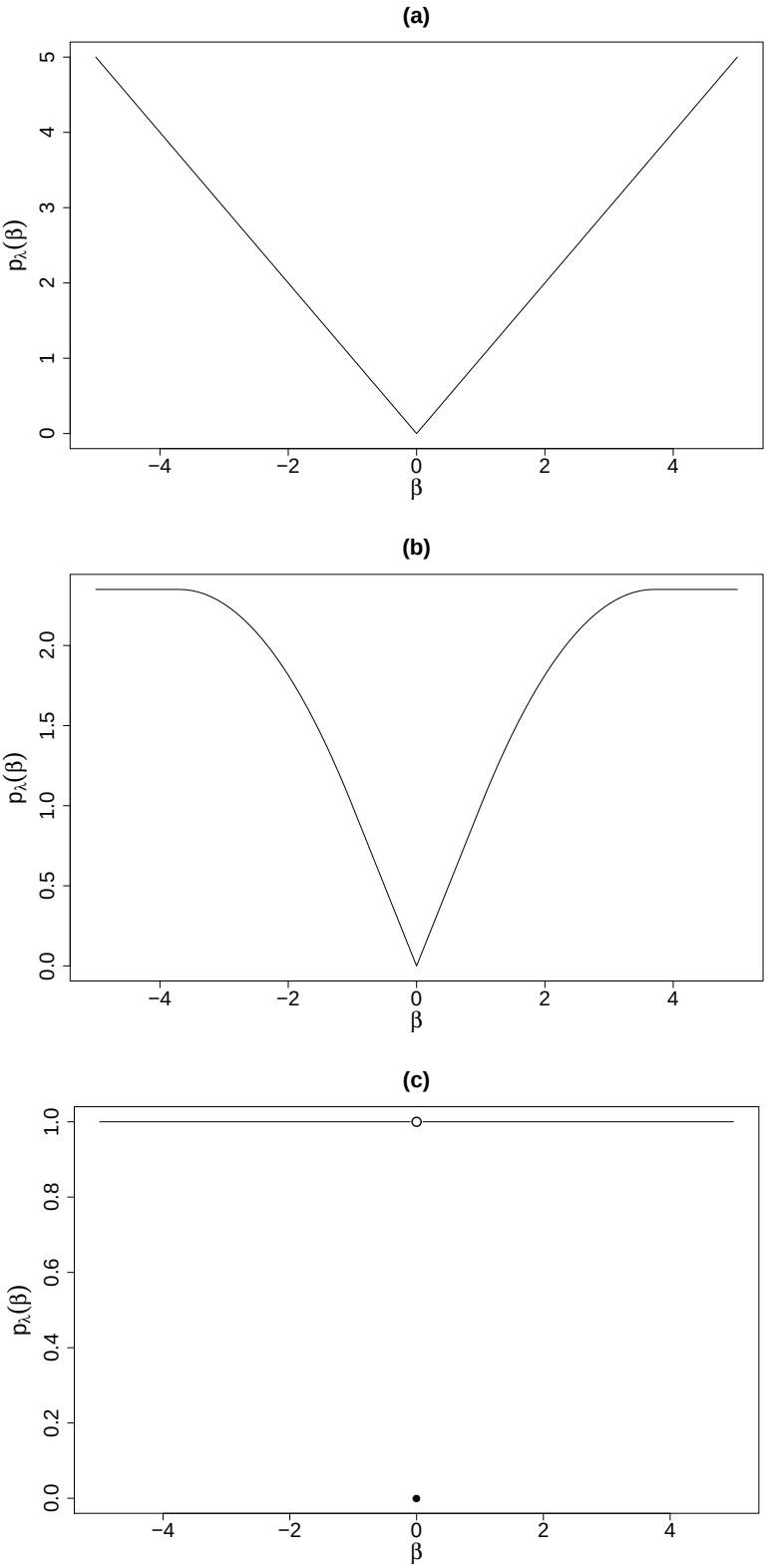

Figure 2.1. Plots of penalty functions with $\lambda = 1$: (a) The LASSO penalty; (b) The SCAD penalty; (c) The L_0 penalty

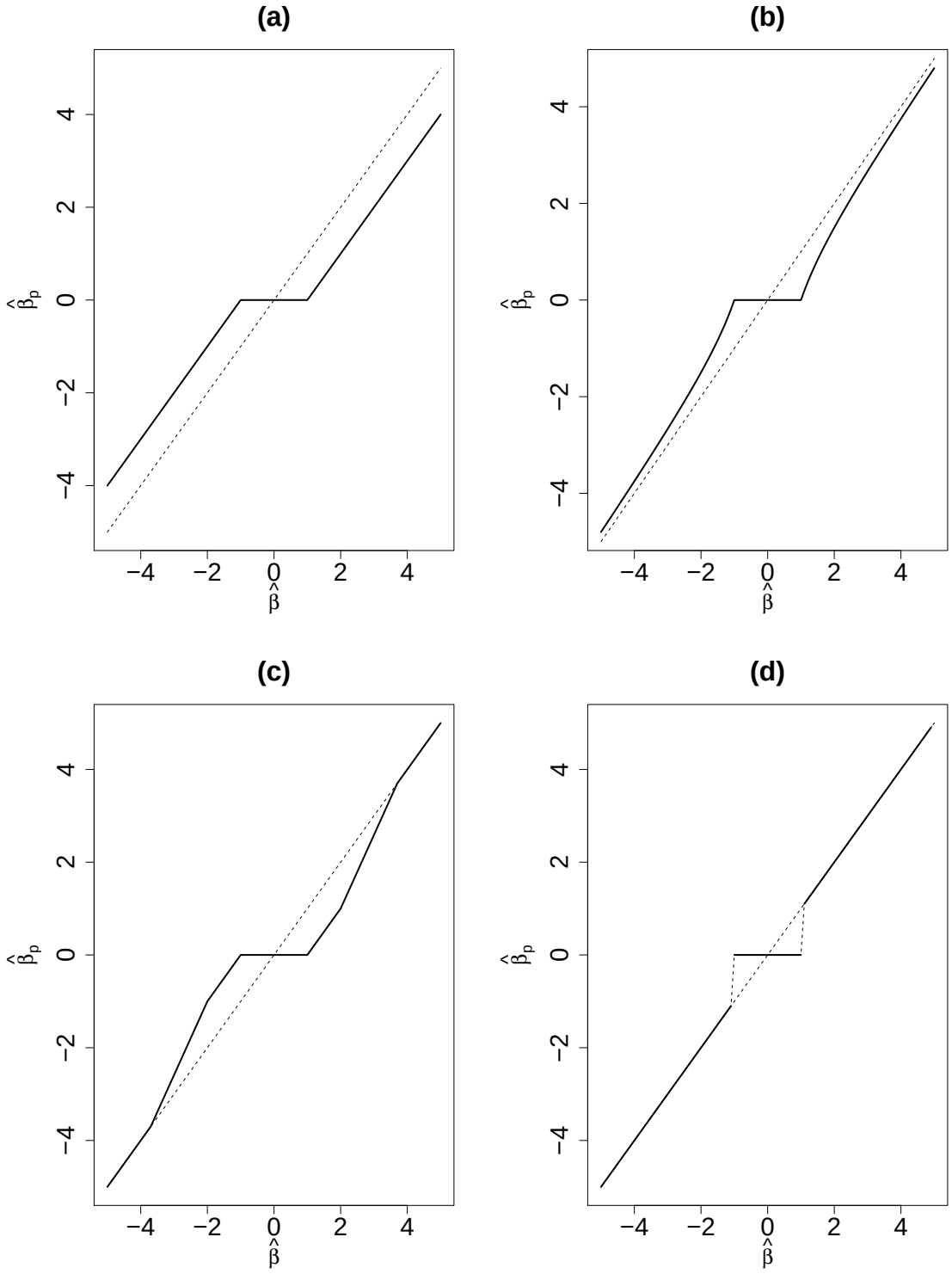

Figure 2.2. The Action of Thresholding Functions with $\lambda = 1$: (a) The LASSO (Soft Thresholding); (b) The adaptive LASSO; (c) The SCAD; (d) Hard Thresholding

estimation procedures are shown in the upper right, lower left and lower right plots of figure 2.2 respectively. In each of these plots, the horizontal axis shows values of the unpenalized estimate ($\hat{\beta}$) and the vertical axis shows values of the penalized estimate ($\hat{\beta}_p$) after application of the thresholding rule.

Wagener and Dette (2012) studied the performance of bridge, LASSO and adaptive LASSO estimators in linear regression models with heteroscedastic error structure. They found that while these methods retain their basic properties with regard to model selection and parameter estimation, they cease to possess the oracle property if the oracle is based only on weighted least squares. This led them to propose weighted penalized least squares methods.

2.4.2 Some Penalized Likelihood Procedures

Apart from the residual sum of squares, penalties can be applied to other loss functions in order to achieve simultaneous estimation of coefficients and variable selection. When applied to the log-likelihood function of a GLM, we obtain a penalized likelihood estimator.

We recall that there are three components to a GLM: the linear predictor ($\mathbf{x}_i'\boldsymbol{\beta}$), the response, y_i, from an exponential family and the link function $g(\cdot)$ which links the mean of the response to the linear predictor. If y_i has mean $E(y_i) = \mu_i$, then the link function g satisfies $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$

As an example of a penalized likelihood procedure, one may consider a penalized likelihood problem with an adaptive LASSO penalty. Maximizing the penalized likelihood function is then equivalent to minimizing

$$-l(\boldsymbol{\beta}) + n\lambda \sum_{j=1}^{k} w_j |\beta_j|$$

where $l(\boldsymbol{\beta})$ is the likelihood.

2.4.3 Cyclic Coordinate Descent (CCD) Algorithms

2.4.3.1 CCD for Penalized Regression

The popularity of the LASSO and the adaptive LASSO stems from the availability of fast and efficient algorithms for obtaining the estimators. "One at a time" or cyclical coordinate descent (CCD) algorithms for the LASSO are described by Friedman et al. (2007) and Wu and Lange (2008) and are readily applicable to the adaptive LASSO. Breheny and Huang (2011) also applies CCD algorithms to obtain the SCAD estimators.

We describe next the use of CCD to obtain adaptive LASSO estimators. The coordinate descent step for solving the jth adaptive LASSO coefficient estimate when we have uncorrelated predictors with coefficients $(\beta_1, \ldots, \beta_k)$ is simple. The solution, $\hat{\beta}_j^{ALASSO}(\lambda)$ is a soft thresholded version of the ordinary least squares estimate

$$\hat{\beta}_j^{ALASSO}(\lambda) = S(\hat{\beta}_j^{OLS}, \lambda^*) = \text{sign}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda^*)_+ \tag{2.16}$$

where $\hat{\beta}_j$ is the ordinary least squares estimate for β_j, and $\lambda^* = \lambda w_j$. This formula applies for design matrices with orthogonal columns.

For the general case of correlated predictors, the coordinate-wise update is not so simple. Friedman et al. (2007) present an iterative algorithm that applies soft thresholding with a "partial residual" as a response variable. Write the adaptive LASSO objective function as

$$\sum_{i=1}^{n}(y_i - \sum_{l \neq j} x_{il}\beta_l - x_{ij}\beta_j)^2 + \lambda \sum_{l \neq j} |\beta_j| + \lambda|\beta_j|$$

where all the values of β_k for $k \neq j$ are held fixed at values $\tilde{\beta}_k(\lambda)$. Minimizing with respect to β_j, we obtain

$$\tilde{\beta}_j(\lambda) = S\left(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda^*\right)$$

where $\tilde{y}_i^{(j)}) = \sum_{l \neq j} x_{il}\tilde{\beta}_l(\lambda)$ is the fitted value excluding the contribution from x_{ij} and $(y_i - \tilde{y}_i^{(j)})$ is the partial residual for fitting β_j. The update is repeated $j = 1, 2, \ldots, k, 1, 2 \ldots$ until convergence to the solution $\hat{\beta}_j^{ALASSO}(\lambda)$. Thus, on each iteration, for each coordinate we compute the least squares coefficient on the partial residual and then apply soft thresholding.

The adaptive LASSO penalty can be applied to a weighted least squares problem when known weights ν_i are associated with each observation. In this situation the objective function is

$$\sum_{i=1}^{n} \nu_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} w_j |\beta_j|$$

The solution, in the case of uncorrelated predictors, $\hat{\beta}_j^{ALASSO}(\lambda)$ is a soft thresholded version of the weighted least squares estimate

$$\hat{\beta}_j^{ALASSO}(\lambda) = S(\hat{\beta}_j^{WLS}, \lambda^*) = \text{sign}(\hat{\beta}_j^{WLS})(|\hat{\beta}_j^{WLS}| - \lambda^*)_+ \tag{2.17}$$

This can be extended to the correlated predictors as above.

2.4.3.2 CCD for Penalized Likelihood

The wide applicability of GLMs is due to a straightforward procedure for obtaining the maximum likelihood estimator (MLE) of the structural parameter $\boldsymbol{\beta}$ (McCullagh and Nelder, 1989). To obtain the MLE, first form the working response which is a linearized form of the link function

$$z_i = \mathbf{x}_i'\boldsymbol{\beta} + (y_i - \mu_i)g'(\mu_i)$$

Next, fit a weighted least squares regression of $\mathbf{z}$ on $\mathbf{X}$ using $W_i = [Var(z_i)]^{-1}$ as weights and iterate until convergence. This procedure is well known and is implemented in almost all statistical software. It works because we have obtained a local quadratic approximation to the log-likelihood and the procedure of regressing the working response on $\mathbf{X}$ with the indicated weights is actually the solution to the problem of minimizing this local quadratic approximation $l_Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} W_i(z_i - x_i'\boldsymbol{\beta})^2$.

Friedman et al. (2010) introduce fast cyclical coordinate descent algorithms for GLMs with convex penalties. They implement these for logistic regression with L_1 penalties. The procedure is as follows:

1. Given current parameter estimates, $\tilde{\boldsymbol{\beta}}$, form the quadratic approximation about the current parameter estimates. i.e. form the working response.

2. Use coordinate descent to solve the penalized weighted least squares problem:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum W_i(z_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{n-1} w_j|\beta|$$

3. Iterate between these steps until convergence.

Thus we see that for fixed λ, the problem has been reduced to successive iterations between forming the working response and solving a weighted least squares problem with an adaptive LASSO penalty. Thus it is exactly like the procedure for obtaining MLEs for a GLM with the added step of shrinking the coefficients at each iteration by soft thresholding.

Chapter 3

Adaptive Penalized Likelihood Effects Selection (APLES)

3.1 Introduction

This chapter focuses on the problem of simultaneous variable selection and estimation in the heteroscedastic linear regression model. In analysis of real datasets in biology, econometrics and engineering, one encounters heteroscedasticity. When there exist predictors that can potentially explain the non-constant variance, it is customary to model the variance of the response as a function of these predictors and subsequently re-use the estimated variance in a weighted least squares (WLS) framework. It is known that if the estimated variances are close to the true but unknown ones, then the resulting WLS estimators are more efficient than the OLS ones.

Doubly penalized likelihood methods which place penalties on both the mean parameters and the variance parameters have been proposed in the literature for heteroscedastic linear models. Xu and Zhang (2011) developed a regularized restricted maximum likelihood (REML) procedure in which they applied two smoothly clipped absolute deviation (SCAD) penalty terms. Daye et al. (2012) developed a similar procedure with two least absolute shrinkage and selection operator (LASSO) penalty terms called the high-dimensional heteroscedastic regression (HHR). Kolar and Sharpnack (2012) note that the HHR does not satisfy the oracle property because it uses the L_1 penalty for both the mean and variance parameters.

Kolar and Sharpnack (2012) developed a three-step procedure using two SCAD penalty terms called the heteroscedastic iterative penalized pseudo-likelihood operator (HIPPO). They proved various theorems regarding the attainment of the oracle property for their procedure. However, they noted that the algorithm can be run for more than three steps for which the theorems were proved. Wu and Huiqiong (2012) showed the generality of the doubly penalized procedure by applying it to mean and variance model selection and estimation for the inverse Gaussian distribution. All of these methods used the iterative ridge algorithm via a local quadratic approximation (LQA) proposed by Fan and Li (2001), with the exception of Daye et al. (2012) who used a cyclic coordinate descent (CCD) algorithm. Table 3.1 summarizes the characteristics of these methods.

Authors	Mean Penalty	Variance Penalty	Algorithm	Oracle	Non-normal Data
Xu and Zhang (2011)	SCAD	SCAD	LQA	Yes	No
Wu and Huiqiong (2012)	SCAD	SCAD	LQA	Yes	Yes
Daye et al. (2012)	LASSO	LASSO	CCD	No	No
Kolar and Sharpnack (2012)	SCAD	SCAD	LQA	Yes	No

Table 3.1. Summary of Methods Used by Authors of Recent Similar Work

This chapter follows a similar framework and provides an additional resource for researchers in using doubly penalized likelihood estimation for heteroscedastic linear models. We propose in this chapter a method of analysis for the heteroscedastic linear model in which both penalties are adaptive LASSO penalties. The doubly penalized likelihood problem reduces to two sets of problems, each of which is a penalized GLM. The HHR iterates between two penalized L_1 sub-problems each of which is convex. The HIPPO iterates between two SCAD-penalized problems neither of which is convex. Thus, even though the HIPPO does enjoy the oracle property it does not seek the global optimal solution at each step. The adaptive LASSO (Zou, 2006) was proposed mainly to provide a version of the LASSO which still retained convexity of the optimization problem but could attain the oracle property. Like the adaptive LASSO, our proposed procedure, which we call the Adaptive Penalized Likelihood Effects Selection (APLES) requires the choice of a preliminary $\sqrt{n}-$consistent estimator for use as weights in the adaptive procedure.

Penalized likelihood methods present a unique computational challenge and

new proposals have to be judged by their computational cost in addition to their intrinsic properties. This is because the tuning parameter(s) correspond to an entire continuum of solutions, each of which corresponds to a potential model. To select the best model based on a criterion, one first needs to obtain the entire tuning path. It is for this reason that the least angle regression (LARS) algorithm (Efron et al., 2004) for obtaining the entire LASSO solution path acquired such wide popularity. Two popular tuning path algorithms for the generalized linear model (GLM) have been proposed. Friedman et al. (2010) utilize a CCD algorithm and Park and Hastie (2007) propose a predictor-corrector algorithm, both of which which take special advantage of the structure of GLMs. The predictor-corrector algorithm is particularly effective and can be used in conjunction with any method for obtaining point solutions for a specific value(s) of the tuning parameter(s). We employ the CCD algorithm for GLMs proposed by Friedman et al. (2010) for the solution of the doubly penalized likelihood objective function. For obtaining best solutions according to some information criterion (AIC, AICc and BIC, in this dissertation), it is computationally expensive to search for solutions over an entire two-dimensional grid if the grid is too fine. We provide recommendations for reducing the computational burden.

The rest of this chapter is organized as follows. In the next section, 3.2, we formalize the model and notation. In section 3.3, we provide the motivation for this work. In section 3.4, we describe the proposed methodology. In section 3.5 we present the CCD algorithm for our framework and its implementation. Section 3.6 presents extensive simulation results under four combinations of conditions: correctly specified model with varying sample size; data containing outliers with increasing extremeness of the outliers; non-normal response; and lastly, a misspecification of the variance submodel. In section 3.7, we briefly illustrate how the procedure can be extended to alternative variance submodels by considering the situation when the variance is linear (rather than log-linear) in the explanatory variables. Section 3.8 presents our conclusions.

3.2 Notation and Model

We consider the usual heteroscedastic linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i, \qquad i = 1, \ldots, n \tag{3.1}$$

where y_i is the response with mean $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}^*$, $\mathbf{x}_i$ is the $p \times 1$ vector of explanatory variables for the mean response, and ϵ_i is the normally distributed random error with mean 0 and variance σ_i^2.

We assume the log-linear variance function

$$\sigma_i^2 = h(\mathbf{x}_i, \boldsymbol{\tau}^*) = \exp(\mathbf{x}_i'\boldsymbol{\tau}^*) \tag{3.2}$$

where $\mathbf{x}_i$ is the $p \times 1$ vector of explanatory variables for the variance response. In section 3.7, we extend this to the linear variance model with $h(\mathbf{x}_i, \boldsymbol{\tau}) = \mathbf{x}_i'\boldsymbol{\tau}$.

We let $\mathbf{X}$ denote the $n \times p$ matrix of explanatory variables. Our notation is general enough to include the situation where the mean and variance responses do not depend on the same set of predictors. For example, suppose $\mu_i = x_{i1}\beta_1^* + x_{i2}\beta_2^* + x_{i4}\beta_4^*$ and $\sigma_i^2 = \exp(x_{i2}\tau_2^* + x_{i3}\tau_3^* + x_{i5}\tau_5^*)$. Then we set $\beta_3^* = \beta_5^* = 0$ and $\tau_1^* = \tau_4^* = 0$, and we have $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$

The vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ are coefficients corresponding to mean and variance effects respectively. We assume that $\boldsymbol{\beta}^*$ has the representation $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ where $\boldsymbol{\beta}_2^* = (0, \ldots, 0)$. The dimensions of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are k_1 and k_2 respectively with $k_1 + k_2 = p$. Similarly, we assume that $\boldsymbol{\tau}^*$ has the representation $\boldsymbol{\tau}^* = (\boldsymbol{\tau}_1^*, \boldsymbol{\tau}_2^*)$ where $\boldsymbol{\tau}_2^* = (0, \ldots, 0)$. The dimensions of $\boldsymbol{\tau}_1^*$ and $\boldsymbol{\tau}_2^*$ are l_1 and l_2 respectively with $l_1 + l_2 = p$.

Let $\mathcal{L} = \{j : \beta_j^* \neq 0\}$ and $\mathcal{L}^c = \{j : \beta_j^* = 0\}$. Consider an estimation procedure which produces estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ and let $\mathcal{L}_n = \left\{j : \hat{\beta}_j \neq 0\right\}$. Such a procedure is said to possess mean model selection consistency if $\lim_n P(\mathcal{L}_n = \mathcal{L}) = 1$. Thus, a procedure is consistent for mean model selection if, with probability arbitrarily close to 1, it correctly identifies the set of non-zero mean coefficients when n is large.

Let $\mathcal{D} = \{j : \tau_j^* \neq 0\}$ and $\mathcal{D}^c = \{j : \tau_j^* = 0\}$. Consider an estimation procedure which produces estimates $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}^*$ and let $\mathcal{D}_n = \{j : \hat{\tau}_j \neq 0\}$. It is said to possess

variance model selection consistency if $\lim_n P(\mathcal{D}_n = \mathcal{D}) = 1$. Thus, a procedure is consistent for variance model selection if, with probability arbitrarily close to 1, it correctly identifies the set of non-zero variance coefficients when n is large.

Finally, any estimator $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}})$ is said to be consistent for model selection for models (3.1) and (3.2) if $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}$ are respectively consistent for the mean and variance models.

3.3 Motivation for Proposed Methodology

Penalized likelihood methods have primarily been proposed for the analysis of high dimensional datasets. In the framework of our joint models ((3.1) and (3.2)), this implies that $p \gg n$. In words, the cardinality of the mean parameters vector $\boldsymbol{\beta}^*$ and the variance parameters vector $\boldsymbol{\tau}^*$ is much greater than the sample size. In these high dimensional settings, the sparsity assumption is invoked to make data analysis possible. This is equivalent to assuming $k_1 \ll p$ and $l_1 \ll p$ in our notation.

In classical experimental designs, we generally do not have high dimensionality of the mean and variance parameters. Thus, for example, when $2p < n$ one does not need the sparsity assumption to estimate all of the parameters in the model. Indeed in response surface methodology (RSM) and the application to robust parameter design (RPD), most of the authors in the literature we reviewed have not found it necessary to use the sparsity assumption directly. Even though variable selection methods including half normal plots, hypothesis testing and subset selection are used to select a good model, the focus has always been on obtaining a model with good predictive abilities. Therefore, a relevant question is why we propose to use a penalized likelihood approach in this dissertation when our ultimate goal is the analysis of designed experiments.

We first note that due to the interaction terms that are of great importance in RPD, it is always possible to have the number of mean parameters being larger than the sample size. It is well-known that the noise factor by control factor interactions are crucial to the success of RPD. In the case where there are not enough observations to estimate all of these, the practice has been to throw some of the terms out of the model based on prior information. We believe that in these situations, the penalized likelihood approaches with sparsity-inducing penalties

can help to reduce the subjectivity in deciding which interaction terms to throw out. Another advantage of our method is that it can be extended to the use of group penalties, which ensure that groups of parameters are included or excluded from the model together (Yuan and Lin, 2006). This is a helpful way to build, for example, the Effect Heredity Principle (Wu and Hamada, 2000) into our model. We do not consider this extension in this dissertation, but consider it as very important future work.

Also, in our framework we consider the heteroscedastic RPD model where the response variance can be modeled directly in terms of the control factors through the variance parameters vector $\boldsymbol{\tau}$. This means that there is an extra set of parameters to be estimated which can easily lead to the total number of parameters being larger than the sample size.

Lastly, screening experiments using unreplicated fractional designs and supersaturated designs are one area where our methods are directly applicable. In these settings the total number of factors that are being considered are much larger than the number of observations. Indeed some recent works have considered screening experiments under the constant variance assumption. Li and Lin (2009) uses the SCAD penalty and Phoa (2009) uses the Dantzig selector.

Screening experiments with dispersion effects have traditionally proven to be hard to analyze as exemplified by the huge number of different methods that have been proposed. These were discussed in the literature review (section 2.3). We utilize the proposal developed in this chapter for the analysis of unreplicated fractional factorial designs in chapter 5. Our method can be extended to supersaturated designs which will be discussed in our future work in section 6.2.

3.4 The Proposed Methodology

Let $L(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*; \mathbf{y})$ be the likelihood for the joint models (3.1) and (3.2). Assuming the random errors are normal and independently distributed, the log-likelihood is (ignoring constant terms) given by:

$$l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*; \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau}^* - \frac{1}{2}\sum_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\tau}^*}(y_i - \mathbf{x}_i'\boldsymbol{\beta}^*)^2 \tag{3.3}$$

We propose to maximize l subject to bounds on the weighted L_1 norms of the mean parameters vector and the variance parameters vector. Thus we define the APLES estimator denoted $\hat{\theta}_A = \left(\hat{\boldsymbol{\beta}}_A, \hat{\boldsymbol{\tau}}_A\right)$ as

$$\begin{aligned}&\underset{\boldsymbol{\beta},\boldsymbol{\tau}}{\operatorname{argmin}} \left[-2 \log \left(\boldsymbol{\beta}, \boldsymbol{\tau}; y\right)\right] \\ &= \underset{\boldsymbol{\beta},\boldsymbol{\tau}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau} + \sum_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\tau}}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} w_{1j}|\beta_j| + \lambda_2 \sum_{j=1}^{p} w_{2j}|\tau_j|\right]\end{aligned} \tag{3.4}$$

where λ_1 and λ_2 are tuning parameters for the mean and variance respectively.

The weight vectors $\mathbf{w}_1 = (w_{11}, \ldots, w_{1k})$ and $\mathbf{w}_2 = (w_{21}, \ldots, w_{2d})$ are chosen to satisfy $w_{1j} = 1/|\tilde{\beta}_j|$ and $w_{2j} = 1/|\tilde{\tau}_j|$ for any $\sqrt{n}-$consistent estimators $\tilde{\beta}_j$ and $\tilde{\tau}_j$. With this choice, the terms $\lambda_1 \sum_{j=1}^{p} w_{1j}|\beta_j|$ and $\lambda_2 \sum_{j=1}^{p} w_{2j}|\tau_j|$ are adaptive LASSO terms and thus ensure penalization of the elements of $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$. The amount of penalization differs for each parameter and is data-driven to ensure asymptotically unbiased estimates of the large parameter values unlike unweighted L_1 penalties used in the HHR. The exact procedure for choosing the weights and tuning parameters is described in section 3.4.3. We also discuss in 3.4.3 how to obtain the weights when $p \gg n$ when we cannot obtain $\sqrt{n}-$consistent estimators to use as the initial estimators.

For fixed $\boldsymbol{\tau}$, the minimization problem in (3.4) is equivalent to a weighted least squares regression with adaptive LASSO penalty where the weights are $e^{-\mathbf{x}_i'\boldsymbol{\tau}}$. For fixed $\boldsymbol{\beta}$, the minimization problem is equivalent to a gamma-error GLM with log link and an adaptive LASSO penalty. By iterating between these two minimization problems, we will obtain the final set of APLES estimates for the non-zero mean and variance parameters. We describe this next.

3.4.1 Fitting the Location submodel

Consider the expression

$$\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau} + \sum_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\tau}}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} w_{1j}|\beta_j| + \lambda_2 \sum_{j=1}^{p} w_{2j}|\tau_j| \tag{3.5}$$

which is -2 times the loglikelihood with penalties with $\boldsymbol{\tau}$ fixed and known. Then minimizing with respect to $\boldsymbol{\beta}$ is equivalent to finding

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \nu_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} w_{1j}|\beta_j| \tag{3.6}$$

where $\nu_i = e^{-\mathbf{x}_i'\boldsymbol{\tau}}$ and w_{1j} is defined by $1/|\tilde{\beta}_j|$ where $\tilde{\beta}_j$ is any $\sqrt{n}-$consistent estimator.

This is a penalized regression problem. Specifically, it is a weighted linear regression with an adaptive LASSO penalty. Using this fact, the solution to the problem (3.6) has the oracle property for the mean parameters when $\boldsymbol{\tau} = \boldsymbol{\tau}^*$: It asymptotically selects the correct mean model, and estimates it with the optimal rate. This follows from theorem 4.4 of Wagener and Dette (2012).

3.4.2 Fitting the Dispersion submodel

For fixed $\boldsymbol{\beta}$ and λ_2, we may obtain the values of $\boldsymbol{\tau}$ that solve the objective function (3.4) as follows. Consider expression (3.4) with $\boldsymbol{\beta}$ fixed and known. Then minimizing with respect to $\boldsymbol{\tau}$ is equivalent to finding

$$\underset{\boldsymbol{\tau}}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau} + \sum_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\tau}} d_i + \lambda_2 \sum_{j=1}^{p} w_{2j}|\tau_j| \tag{3.7}$$

where $d_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ and w_{2j} is defined by $1/|\tilde{\tau}_j|$ where $\tilde{\tau}_j$ is any $\sqrt{n}-$consistent estimator of τ_j.

When $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, then the random variable d_i is such that d_i/σ_i^2 is distributed as χ^2 with 1 degree of freedom. The optimization problem (3.7) reduces to a penalized likelihood problem where the likelihood is from a GLM. Specifically, it is from a Gamma error regression model with log link. Even though this is not the canonical link for the Gamma distribution, the objective in (3.7) is convex in $\boldsymbol{\tau}$. This is because the matrix of second derivatives, $\sum_{i=1}^{n}(d_i/e^{-\mathbf{x}_i'\boldsymbol{\tau}})\mathbf{x}_i\mathbf{x}_i'$, is positive semidefinite.

Now consider λ_2 as a function of n. Theorem 4 of Zou (2006) states that a penalized GLM fit (such as (3.7)) which satisfies the usual regularity conditions

on the boundedness of the third derivatives of the log likelihood has the oracle property if $\lambda_2(n)/\sqrt{n} \to 0$ and $\lambda_2(n) \to \infty$. Thus the solution to problem (3.7) has the oracle property for the variance parameters ($\boldsymbol{\tau}^*$) if these conditions are satisfied.

3.4.3 Choice of Adaptive Weights and Choice of Tuning Parameters

The adaptive LASSO penalty is able to achieve the oracle property because it does not uniformly penalize all coefficients. Rather, it utilizes the information about the true magnitudes of the coefficients that can be obtained by a $\sqrt{n}-$consistent estimator. When the sample size is larger than the total number of parameters, (i.e. $n \geq 2p$), then the unpenalized maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ may be used as the preliminary estimators.

In the $n \leq 2p$ case we recommend the use of preliminary estimators of $\boldsymbol{\beta}, \boldsymbol{\tau}$ obtained by using the HHR. Using the algorithm described in section 3, the entire HHR solutions path can be efficiently computed. We recommend using the HHR tuned with the AIC rather than the BIC since Yang (2005) shows that the AIC can be minimax-rate optimal whereas the BIC only attains consistent model selection if the true model is among the candidates.

A justification for the use of HHR as the preliminary estimate in the $n \leq 2p$ case is provided by Zou and Li (2008) and Fan et al. (2014). Zou and Li (2008) introduce a local linear approximation (LLA) algorithm for computing solutions to the non-convex penalized likelihood problem (of which the penalized likelihood with SCAD penalty is an example). Zou and Li (2008) showed that the adaptive LASSO is closely related to the SCAD through the LLA algorithm. The LLA algorithm solves a sequence of the adaptive LASSO problems, where the adaptive weights are chosen using the LASSO as the initial solution. Moreover, Fan et al. (2014) proved that the LLA algorithm is able to find the oracle solution for several widely studied problems.

Daye et al. (2012) note that resampling techniques such as cross-validation may be invalid under heteroscedasticity. Following Daye et al. (2012) and Kolar and Sharpnack (2012), we select the optimal settings of λ_1, λ_2 using the Akaike

Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In addition to these, for the applications that we consider in Chapters 4 and 5, we will also use AIC with a finite sample size correction (AICc). These are defined as:

$$
\begin{aligned}
AIC &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) + 2r \\
AICc &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) + 2r + \frac{2r(r+1)}{n-r-1} \\
BIC &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) + r\log(n)
\end{aligned}
$$

where r is the total number of non-zero estimated coefficients. The optimal pair of (λ_1, λ_2) is chosen by solving over a grid of values. This can be computationally expensive for large n. In section 3.5, we recommend using the solutions at neighboring (λ_1, λ_2) pairs as the starting point to solve for a given (λ_1, λ_2) pair. Another option is to use the predictor-corrector algorithm (Park and Hastie, 2007) for performing the grid search for the optimal pair.

3.5 Computational Algorithm for APLES

3.5.1 Framework

To motivate the proposed algorithm, consider maximum likelihood estimation of $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ (i.e. without penalization). This approach has a long-standing history for modeling heteroscedastic data (Harvey 1976; Aitkin 1987; Verbyla 1993).

Let $\boldsymbol{\Sigma}$ denote the diagonal covariance matrix of $\mathbf{y}$ with ith diagonal element σ_i^2, $\mathbf{1}$ denote a vector of ones and the vector $\mathbf{d} = (d_1, \ldots, d_n)'$. The first derivative of $l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*; \mathbf{y})$ (usually called the score vector in the literature) is given by:

$$
\mathcal{S}(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*) = \begin{pmatrix} \sum_{i=1}^n \mathbf{x}_i e^{-\mathbf{x}_i'\boldsymbol{\tau}^*}(y_i - \mathbf{x}_i'\boldsymbol{\beta}^*) \\ \sum_{i=1}^n \frac{1}{2}\mathbf{x}_i \left(e^{-\mathbf{x}_i'\boldsymbol{\tau}^*} d_i - 1\right) \end{pmatrix}_{(p+p)\times 1} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) \\ \frac{1}{2}\mathbf{X}'(\boldsymbol{\Sigma}^{-1}\mathbf{d} - \mathbf{1}) \end{pmatrix}_{(p+p)\times 1} \tag{3.8}
$$

and by taking expectations of the matrix of negative second derivatives (observed

information), we obtain the Fisher information matrix as

$$\mathcal{I}(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*) = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{X}'\mathbf{X} \end{pmatrix}_{(p+p)\times(p+p)} \tag{3.9}$$

Due to the nature of the information matrix given in equation (3.9), we have parameter orthogonality between $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$. In general, two parameters θ_i and θ_j are said to be orthogonal if the ijth element of the Fisher information matrix of $\begin{pmatrix} \theta_i \\ \theta_j \end{pmatrix}$ is zero. This definition extends naturally to sets of parameters.

One way to obtain the unconstrained MLEs is to use Fisher scoring. Let $\hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\tau}}^{(t)}$ be our current estimates. To obtain updated estimates, $\hat{\boldsymbol{\beta}}^{(t+1)}$ and $\hat{\boldsymbol{\tau}}^{(t+1)}$, let us define $\hat{\boldsymbol{\Sigma}}^{(t)}$ to be a diagonal matrix with ith diagonal element $e^{\mathbf{x}_i'\hat{\boldsymbol{\tau}}^{(t)}}$. Then the Fisher scoring updates are given by:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{(t+1)} \\ \hat{\boldsymbol{\tau}}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(t)} \\ \hat{\boldsymbol{\tau}}^{(t)} \end{pmatrix} + \mathcal{I}^{-1}\mathcal{S}. \tag{3.10}$$

Now, using (3.8) and (3.9) the right hand side of (3.10) becomes:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{(t)} \\ \hat{\boldsymbol{\tau}}^{(t)} \end{pmatrix} + \begin{pmatrix} \mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{X}'\mathbf{X} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)}) \\ \frac{1}{2}\mathbf{X}'((\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\hat{\mathbf{d}} - \mathbf{1}) \end{pmatrix}$$

This separates into the two equations:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\mathbf{y} \tag{3.11}$$

$$\hat{\boldsymbol{\tau}}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'((\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\hat{\mathbf{d}}^{(t+1)} - \mathbf{1} + \mathbf{X}\hat{\boldsymbol{\tau}}^{(t)}) \tag{3.12}$$

Thus we iterate between two sets of linear regression equations until they converge to obtain the MLEs. We use $\hat{\mathbf{d}}^{(t+1)} = (\mathbf{y} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(t+1)})^2$ in (3.12). This is

because after updating with equation (3.11), the most current estimate of $\boldsymbol{\beta}^*$ is $\hat{\boldsymbol{\beta}}^{(t+1)}$ and we must use this current estimate in the next step.

Another point of view is this: Let $l_Q(\boldsymbol{\tau})$ denote the local quadratic approximation of the log-likelihood (considered as a function of $\boldsymbol{\tau}$ only) and $l_Q(\boldsymbol{\beta})$ denote the local quadratic approximation of the log-likelihood (considered as a function of $\boldsymbol{\beta}$ only). The Fisher scoring step in equation (3.11) minimizes $l_Q(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{1}{\hat{\sigma}_i^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. The equation in step (3.12) minimizes $l_Q(\boldsymbol{\tau}) = \sum_{i=1}^n (s_i - \mathbf{x}_i'\boldsymbol{\tau})^2$, where $s_i = \mathbf{x}_i'\hat{\boldsymbol{\tau}} + \frac{d_i}{\hat{\sigma}_i^2} - 1$.

Using this point of view, to obtain the current APLES estimates $\boldsymbol{\beta}_A$ of $\boldsymbol{\beta}^*$ we minimize the penalized local quadratic approximation

$$\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} w_{1j}|\beta_j| \tag{3.13}$$

and to obtain current APLES estimates $\boldsymbol{\tau}_A$ of $\boldsymbol{\tau}^*$ we minimize the penalized local quadratic approximation

$$\sum_{i=1}^{n} (s_i - \mathbf{x}_i'\boldsymbol{\tau})^2 + \lambda_2 \sum_{j=1}^{p} w_{2j}|\tau_j|. \tag{3.14}$$

Thus for our proposed algorithm, we embed two sub-algorithms for penalized GLM regression within each loop. We use the CCD algorithm for GLMs proposed by Friedman et al. (2010). The properties of the CCD algorithm in terms of speed and efficiency are well described (e.g. Wu and Lange (2008)). We implement it to solve (3.13) and (3.14) as follows.

We incorporate the adaptive weights into the penalty terms by setting $\lambda_1^* = \lambda_1 w_{1j}$ and $\lambda_2^* = \lambda_2 w_{2j}$. Given the response ($y_i$ for the objective function (3.13) and s_i for the objective function (3.14)), within each sub-algorithm we update a coefficient by regression of the partial residual on the predictor for that coefficient and then perform soft thresholding (LASSO) on the coefficient using the soft

thresholding operator, $S(z, \lambda)$. $S(z, \lambda)$ is defined as

$$S(z,\lambda) = \text{sign}(z)(|z| - \lambda)_+ = \begin{cases} z - \lambda & \text{if } z > \lambda \text{ and } \lambda < |z| \\ z + \lambda & \text{if } z < \lambda \text{ and } \lambda < |z| \\ 0 & \text{if } |z| \leq \lambda \end{cases}$$

By cycling through all predictors several times until convergence we obtain the solution for that value of λ_1, λ_2.

3.5.2 Cyclic Coordinate Descent Algorithm (CCD) for APLES

In this subsection, we present an algorithm to compute APLES estimates. The algorithm is a CCD algorithm similar to the one described for GLMs by Friedman et al. (2010). Suppose that one is interested in obtaining APLES estimates for fixed tuning parameters λ_1 and λ_2. We also assume that we have available preliminary estimators chosen according to section 3.4.3. In step 1 of the algorithm, we initialize the algorithm using the preliminary estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\tau}}$. In other words we set $\hat{\boldsymbol{\beta}}^{(0)} = \tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}^{(0)} = \tilde{\boldsymbol{\tau}}$. We also determine constants, ϵ_1^* and ϵ_2^* to use in the convergence criteria.

In step 2, we state the convergence criteria. In steps 3 and steps 4, we run the sub-algorithms to determine current estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. We run these steps until the convergence criteria are met.

1. Choose λ_1^* and λ_2^*. Set $\hat{\boldsymbol{\beta}}^{(0)} = \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}^{(0)} = \tilde{\boldsymbol{\tau}}$. Choose ϵ_1^* and ϵ_2^* to use in convergence criteria.

2. For $t = 0, 1, 2, \ldots$ until convergence criteria $\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\| < \epsilon_1^*$ and $||\hat{\boldsymbol{\tau}}^{(t+1)} - \hat{\boldsymbol{\tau}}^{(t)}|| < \epsilon_2^*$ are met:

3. Implement sub-algorithm 1

   (a) Form $(W_1^{(t)}, \ldots, W_n^{(t)}) = (1/e^{\mathbf{x}_1' \hat{\boldsymbol{\tau}}^{(t)}}, \ldots, 1/e^{\mathbf{x}_n' \hat{\boldsymbol{\tau}}^{(t)}})$

   (b) For j in $1 : p$, obtain $\hat{\beta}_{j,temp}$ as the weighted least squares coefficient on

the partial residual:

$$\hat{\beta}_{j,temp} = \frac{W_i^{(t)} x_{ij}(y_i - \sum_{l \neq j} x_{il}\hat{\beta}_l^{(t)})}{\sum_{j=1}^{p} W_i^{(t)} x_{ij}^2}$$

(c) Apply soft thresholding to obtain $\hat{\beta}_j^{(t+1)} = S\left(\hat{\beta}_{j,temp}, \lambda_1^*\right)$

(d) Obtain the squared residuals $\hat{d}_i$ and form the working response for the next stage: $s_i = \mathbf{x}_i'\hat{\boldsymbol{\tau}}^{(t)} + \frac{d_i}{\hat{\sigma}_i^2} - 1$. Here $\hat{\sigma}_i^2 = e^{\mathbf{x}_i'\hat{\boldsymbol{\tau}}^{(t)}}$.

4. Implement sub-algorithm 2

(a) For j in $1:p$, obtain $\hat{\tau}_{j,temp}$ as the least squares coefficient on the partial residual:

$$\hat{\tau}_{j,temp} = \frac{u_{ij}(s_i - \sum_{j \neq l} u_{il}\hat{\tau}^{(t)})}{\sum_{j=1}^{p} u_{ij}^2}$$

(b) Apply soft thresholding to obtain $\hat{\tau}_j^{(t+1)} = S\left(\hat{\tau}_{j,temp}, \lambda_2^*\right)$

3.5.3 The APLES Tuning Parameter Path

We can compute the two-dimensional tuning path by computing solutions for $\boldsymbol{\beta}, \boldsymbol{\tau}$ over a rectangular grid of values for $(0, \lambda_1^{max}) \times (0, \lambda_2^{max})$, where $\lambda_1^{max}, \lambda_2^{max}$ are any values of λ_1, λ_2 which are so large that all coefficient estimates other than $\hat{\beta}_0$ and $\hat{\tau}_0$ (the estimates of the intercept terms for the mean and variance, respectively) are set to 0. For example, to obtain $\lambda_1^{max}, \lambda_2^{max}$, set $\lambda_1 = 10, \lambda_2 = 10$ and solve for $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\tau}}_A$. If all coefficients other than $\hat{\beta}_0$ and $\hat{\tau}_0$ equal 0, then set $\lambda_1^{max} = 10, \lambda_2^{max} = 10$. Otherwise, set $\lambda_1 = 20, \lambda_2 = 20$ and solve for $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\tau}}_A$. If all coefficients other than $\hat{\beta}_1$ and $\hat{\tau}_1$ equal 0, then we may use $\lambda_1^{max} = 20, \lambda_2^{max} = 20$. By repeated doubling, it is always possible to arrive at a pair which can be used for $\lambda_1^{max}, \lambda_2^{max}$. Now, suppose we have found $(\lambda_1^{max}, \lambda_2^{max}) = (4, 5)$ and we seek solutions using the grid $(0, 0.4, 0.8, \ldots, 3.6, 4.0) \times (0, 0.5, 1.0, \ldots, 4.5, 5.0)$.

A procedure to obtain the two-dimensional solution path is:

1. Let $\lambda_1 = \lambda_1^{max} (= 4)$.

   (a) Begin with $\lambda_2^{max} (= 5)$ and compute the APLES estimates of $(\boldsymbol{\beta}, \boldsymbol{\tau})$ using the algorithm described in 3.5.2.

   (b) Now decrement λ_2 (i.e. set equal to 4.5) and compute the APLES estimates of $(\boldsymbol{\beta}, \boldsymbol{\tau})$.

   (c) Continue until $\lambda_2 = 0$.

2. Decrement λ_1 (i.e. set equal to 3.6) and repeat steps 1(a) to 1(c).

3. Repeat step 2 until $\lambda_1 = 0$.

If p is large, this can be computationally expensive. For this situation, we recommend that when moving along in the grid, one should use the solutions for nearby points as the starting point for the algorithm. For example, when one has obtained APLES estimates at $(\lambda_1, \lambda_2) = (3.2, 3.5)$ and one seeks solutions for $(\lambda_1, \lambda_2) = (2.8, 3.5)$, the solutions at $(\lambda_1, \lambda_2) = (3.2, 3.5)$ should be used as the starting point.

An alternative approach is to use the CCD algorithm described above (section 3.5.2), which gives point solutions for each λ_1, λ_2 pair in conjunction with the predictor-corrector algorithm of Park and Hastie (2007). Proceed as follows:

1. Let $\lambda_1 = \lambda_1^{max}$. Use the predictor-corrector algorithm to solve the entire path from λ_2^{max} all the way down to 0.

2. Decrement λ_1 and use the predictor-corrector algorithm to solve the entire path from λ_2^{max} all the way down to 0. .

3. Repeat step 2 until $\lambda_1 = 0$.

For a fixed λ_1, λ_2 pair, the solution obtained using the predictor-corrector algorithm is not as accurate as the first approach (in case one does not solve the corrector step to convergence). However, this approach is computationally more efficient and much faster. See Park and Hastie (2007) for more details.

3.5.4 Standard Errors for APLES Estimates

We describe in this section a procedure for computing the standard errors of the non-zero APLES estimates. This procedure is an extension of the method described by Zou (2006) who showed that the local quadratic approximation (LQA) can provide a sandwich formula for the covariance of the non-zero components of the adaptive LASSO estimates.

We recall our notation in section (3.2). Suppose the first k_1 components of $\boldsymbol{\beta}^*$ and the first l_1 components of $\boldsymbol{\tau}^*$ are non-zero. Let $\mathbf{X}_{k_1}$ denote the first k_1 columns of $\mathbf{X}$ and $\mathbf{X}_{l_1}$ denote the first l_1 columns of $\mathbf{X}$. Let $\Sigma(\boldsymbol{\beta}) = \text{Diag}\left(\frac{w_{11}}{|\beta_1^*|}, \ldots, \frac{w_{1k_1}}{|\beta_{k_1}^*|}\right)$ and $\Sigma(\boldsymbol{\tau}) = \text{Diag}\left(\frac{w_{21}}{|\tau_1^*|}, \ldots, \frac{w_{2l_1}}{|\tau_{l_1}^*|}\right)$.

An LQA algorithm for computing the APLES estimates proceeds by iteratively performing the ridge regressions:

$$\left(\hat{\beta}_1, \ldots, \hat{\beta}_{k_1}\right) = \left(\mathbf{X}_{k_1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_{k_1} + \lambda_1\Sigma(\boldsymbol{\beta})\right)^{-1}\mathbf{X}_{k_1}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{3.15}$$

$$(\hat{\tau}_1, \ldots, \hat{\tau}_{l_1}) = \left(\frac{1}{2}\mathbf{X}_{l_1}'\mathbf{X}_{l_1} + \lambda_2\Sigma(\boldsymbol{\tau})\right)^{-1}\frac{1}{2}\mathbf{X}_{l_1}'\left(\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{d} - \mathbf{1}\right) \tag{3.16}$$

This leads to the estimated covariance matrix for the nonzero components of the APLES estimates as:

$$\widehat{cov}\left(\hat{\boldsymbol{\beta}}_A\right) = \left(\mathbf{X}_{k_1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_{k_1} + \lambda_1\Sigma(\hat{\boldsymbol{\beta}})\right)^{-1}\mathbf{X}_{k_1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_{k_1}\left(\mathbf{X}_{k_1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_{k_1} + \lambda_1\Sigma(\hat{\boldsymbol{\beta}})\right)^{-1} \tag{3.17}$$

$$\widehat{cov}\left(\hat{\boldsymbol{\tau}}_A\right) = \left(\frac{1}{2}\mathbf{X}_{l_1}'\mathbf{X}_{l_1} + \lambda_2\Sigma(\hat{\boldsymbol{\tau}})\right)^{-1}\frac{1}{2}\mathbf{X}_{l_1}'\mathbf{X}_{l_1}\left(\frac{1}{2}\mathbf{X}_{l_1}'\mathbf{X}_{l_1} + \lambda_2\Sigma(\hat{\boldsymbol{\tau}})\right)^{-1} \tag{3.18}$$

3.6 Simulation Results

We study the finite-sample properties of APLES in relation to the HHR and HIPPO. For each of these methods, the tuning parameters may be selected using

AIC, BIC or AICc. Thus, our comparison is among nine different estimators. The four scenarios that we consider are:

1. When the mean and variance submodels are correctly specified: In this scenario, each method is estimating the coefficients $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ using a data set generated from the true models (3.1) and (3.2). We use $p = 31$ and the number of non-zero elements for both $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ is 7. Thus, the true mean and variance submodels are actually sparse. This way we are able to compare performance of the methods based both on variable selection as well as predictive accuracy.

2. When outliers are present: In this scenario, we use the same $\boldsymbol{\beta}^*$ as in scenario 1. There are however no non-zero elements of $\boldsymbol{\tau}^*$ apart from the intercept. The data generating process however is not from the true joint models (3.1) and (3.2). We use (3.1) to generate the mean structure. For 90% of the observations, we generate errors with a certain fixed variance and for the remaining we use a larger variance. In this scenario, we wish to find the extent to which the presence of outliers may cause the methods to spuriously identify dispersion effects in a particular predictor when there actually is none.

3. When the error distribution is non-normal: Similar to scenario 2, we use this scenario to compare the performance of the methods outside of true model conditions. We consider when the error distribution is non-normal. We wish to find the extent to which the presence of non-normal errors may cause the methods to spuriously identify dispersion effects in a particular predictor when there actually is none.

4. When the variance submodel is misspecified: In this scenario, we use the same $\boldsymbol{\beta}^*$ and $\boldsymbol{\tau}^*$ as in scenario 1. However, the variance model is not (3.2) but rather a linear variance model. We wish to compare the methods when dispersion effects do exist but do not arise according to the structure specified by (3.2).

With the exception of the fourth scenario, these are the same scenarios that were considered by Daye et al. (2012) in their simulation study. All of the estimates

are computed using the algorithm proposed in section 3.5. For each estimator, we measure closeness to the true parameter values using the mean square error (MSE) for $\hat{\beta}$ and $\hat{\tau}$ separately:

$$MSE(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$$
$$MSE(\hat{\boldsymbol{\tau}}) = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)'(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)$$

In addition, we also report the prediction error (PE) given by:

$$E\left[(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)'(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)\right]$$

We measure the variable selection properties using the specificity and the sensitivity defined as
sensitivity($\hat{\boldsymbol{\beta}}$) = (no. of true non-zero mean parameters identified as non-zero/k_1)
specificity($\hat{\boldsymbol{\beta}}$) = (no. of true zero mean parameters identified as zero/k_2)
sensitivity($\hat{\boldsymbol{\tau}}$) = (no of true non-zero variance parameters identified as non-zero/l_1)
specificity($\hat{\boldsymbol{\tau}}$) = (no of true zero variance parameters identified as zero/l_2)
For each of the following situations, the data are generated as follows:

- *Scenario 1*: The predictors are given by $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1)$, where $\mathbf{X}_1 \sim N(0, \boldsymbol{\Sigma})$ and $\Sigma_{ij} = 0.5^{i-j}$. We set:
  $\boldsymbol{\beta}^* = (2, 3, 3, 3, 0, 0, 0, 2, 2, 2, \underbrace{0, \ldots, 0}_{21 \text{ times}})$
  $\boldsymbol{\tau}^* = (1, 1, 1, 1, 0.5, 0.5, 0.5, \underbrace{0, \ldots, 0}_{24 \text{ times}})$
  Thus p equals 31. Using these values, we simulate from the joint models (3.1) and (3.2) 500 times. We vary n over $\{60, 80, 120, 160, 200, 240\}$ to illustrate the effect of sample size. We report in Table A.1 the means (and standard errors) for the PE, MSEs, sensitivity and specificity using the three approaches: HHR, HIPPO and APLES.

- *Scenario 2*: The second experiment looks at how the methods perform in the presence of outliers. We use the same $\boldsymbol{\beta}^*$ as scenario 1. However, $\boldsymbol{\tau}^*$ has all elements set to 0. The error structure is generated as follows. We use $n = 100$ observations with $n_{outlier} = 10$ of these being outliers. The outliers are generated from a truncated normal distribution similar to Daye

et al. (2012) with α measuring the extremeness of the outliers present. For the outliers, we have $\sigma\epsilon_i^{\text{outlier}} \sim N(0, \sigma^2)$ and $|\sigma\epsilon_i^{\text{outlier}}| > \alpha\sigma$. For the non-outlying observations, we have $\sigma\epsilon_i^{\text{outlier}} \sim N(0, \sigma^2)$ and $|\sigma\epsilon_i^{\text{outlier}}| \leq \alpha\sigma$. In the results presented, we have used $\sigma = 3$. Table A.2 presents the results of this experiment for $\alpha = 1, 3$ and 5.

- *Scenario 3*: We use the same $\boldsymbol{\beta}^*$ as scenario 1. However, $\boldsymbol{\tau}^*$ has all elements set to 0. The error structure is generated as follows. We use $n = 100$ observations drawn from the t distribution with ν degrees of freedom. We vary ν over $\{3, 5, 10\}$. Table A.3 presents the results of this simulation.

- *Scenario 4*: Finally, we consider a similar situation as scenario 1. We use $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1)$, where $\mathbf{X}_1 \sim N(0, \boldsymbol{\Sigma})$ and $\Sigma_{ij} = 0.5^{i-j}$. However, the variance submodel is misspecified in that, we generate σ_i^2 according to $\sigma_i^2 = \max(0, \mathbf{x}_i'\boldsymbol{\tau})$, where
  $\tau = (1, 1, 1, 1, 0.5, 0.5, 0.5, \underbrace{0, \ldots, 0}_{24 \text{ times}})$
  Instead of a log-linear variance model, the variance at each setting is generated according to a linear variance model. Thus this experiment compares performance under misspecification of the variance submodel. Table A.4 presents the results of this simulation.

3.6.1 Discussion of Simulation Results

The smallest two PE, MSE and highest sensitivity and specificity are shown in boldface in Tables A.1 - A.4. When the mean and variance submodels are correctly specified, Table A.2 shows that APLES performs the best in terms of prediction error and MSE for $\boldsymbol{\beta}$ when the sample size is small. For larger sample sizes HIPPO performs the best. This can also be seen from figures 3.1 and 3.2. For variable selection measured using sensitivity and specificity, APLES is again very competitive. It performs better than HHR and HIPPO for small sample sizes ($n = 60$) and is competitive with HIPPO for medium ($n = 120$) and large sample sizes ($n = 200$).

When outliers are present, APLES generally outperforms the other methods in terms of PE (figures 3.3 and 3.4) and MSE. It however has reduced sensitivity for

$\boldsymbol{\tau}$ relative to HIPPO.

HIPPO performs the best when we have non-normal data as seen in Table A.3. It has the lowest PE and MSE for $\boldsymbol{\beta}$. However, APLES is usually the best in terms of MSE and specificity for $\boldsymbol{\tau}$.

HIPPO again performs the best in terms of PE and MSE for $\boldsymbol{\beta}$ when the variance submodel is misspecified.

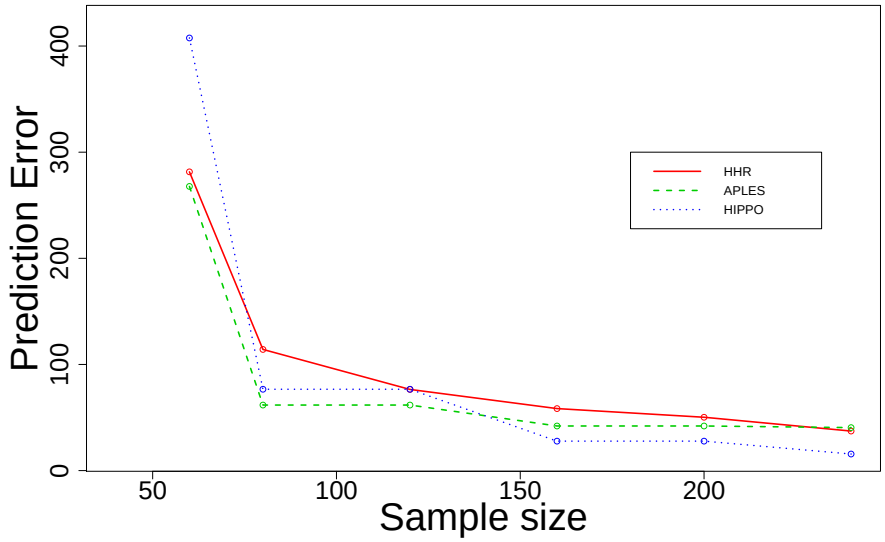

Figure 3.1. Plot of Prediction Error versus Sample Size Using AIC

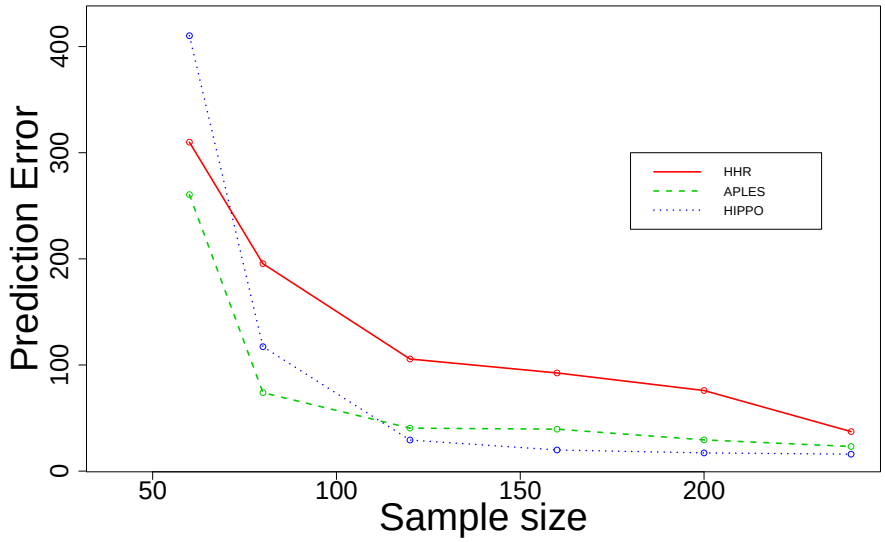

Figure 3.2. Plot of Prediction Error versus Sample Size Using BIC

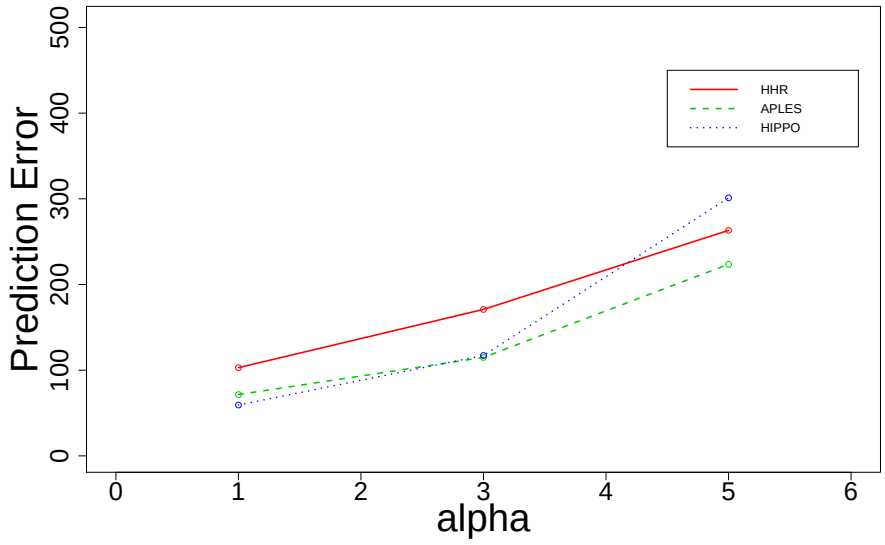

Figure 3.3. Plot of Prediction Error versus α Using AIC

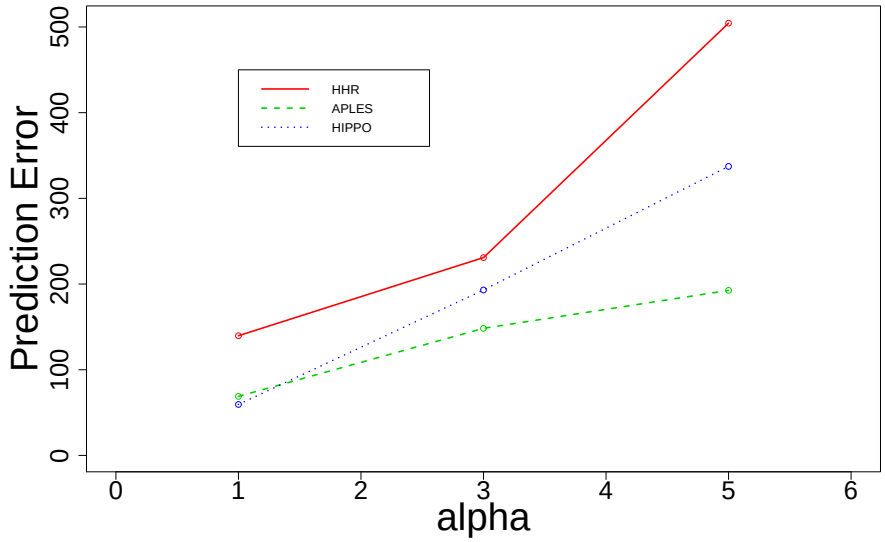

Figure 3.4. Plot of Prediction Error versus α Using BIC

3.7 Extension to the Linear Variance Model

Most of the work on doubly penalized likelihood estimation for heteroscedastic models has focused on the log-linear or multiplicative variance representation. However, the method of this chapter is general enough to encompass other variance functions. In this section we consider the linear variance function:

$$\sigma_i^2 = h(\mathbf{x}_i, \boldsymbol{\tau}) = \mathbf{x}_i'\boldsymbol{\tau} \tag{3.19}$$

When used without a constraint, it has the disadvantage that it may lead to negative values of the variance. However, this representation has found extensive usage in econometrics and has applications to random coefficient estimation (e.g. Breusch and Pagan (1979)).

It can be shown that, under the linear variance model,

$$l'(\boldsymbol{\beta}, \boldsymbol{\tau}) = \begin{pmatrix} \sum_{i=1}^n \mathbf{x}_i \sigma_i^{-2}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \\ \sum_{i=1}^n \frac{1}{2}\mathbf{x}_i \left(d_i \sigma_i^{-4} - \sigma_i^{-2}\right) \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ \frac{1}{2}\mathbf{X}'(\boldsymbol{\Sigma}^{-2}\mathbf{d} - \boldsymbol{\Sigma}^{-1}\mathbf{1}) \end{pmatrix} \tag{3.20}$$

and by taking expectations of the matrix of negative second derivatives, the Fisher

information matrix is obtained as

$$\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{X}'\boldsymbol{\Sigma}^{-2}\mathbf{X} \end{pmatrix} \tag{3.21}$$

To obtain the unconstrained MLEs, the Fisher scoring updates are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y} \tag{3.22}$$

$$\hat{\boldsymbol{\tau}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\tau}} + \mathbf{d} - \hat{\boldsymbol{\Sigma}}\mathbf{1}) \tag{3.23}$$

However, the ith element of $\mathbf{X}\hat{\boldsymbol{\tau}} + \mathbf{d} - \hat{\boldsymbol{\Sigma}}\mathbf{1}$ is $\mathbf{x}_i'\boldsymbol{\tau} - d_i - \sigma_i^2 = d_i$, so that the working response for fitting the variance submodel is exactly equal the squared residual from fitting the mean model. By making this change in the algorithm presented in section 3.2, it is straightforward to obtain APLES estimates.

3.8 Discussion

We began this chapter with a presentation of recent work that has been done on simultaneous variable selection and estimation in the heteroscedastic linear model. We have proposed in section 3.4 a new methodology for this framework, the Adaptive Penalized Likelihood Effects Selection (APLES) estimation procedure. We have shown that the APLES estimator can simultaneously select and estimate variables for the mean and the variance in a heteroscedastic model.

We have provided a new perspective on fitting of the penalized heteroscedastic linear model (3.1 and 3.2) with log-linear variance structure. We achieve this by using a link between the well-known (unpenalized) MLE fitting procedure of Fisher Scoring and the more recent Coordinate Descent approach to GLMs of Friedman et al. (2010). This perspective enabled us to propose an efficient and fast algorithm for obtaining APLES estimators which we presented in section 3.5. The proposed algorithm can be used for obtaining HHR and HIPPO estimates as well. When a different parametric variance function from the ones discussed here are specified, the algorithm described here can still be applied with the appropriate modifications.

In extensive simulation studies (section 3.6) we found that APLES compares favorably with existing techniques and offers a valuable resource in analyzing small-

as well as large-sample data with heteroscedasticity. It performs well under some non-standard conditions such as data with extreme outliers. It does not perform as well under other non-standard conditions such as non-normality and misspecification of the variance structure and therefore we recommend the use of HIPPO in these situations.

Chapter 4

Analysis of Robust Parameter Design Experiments using APLES

4.1 Introduction

In the dual response robust parameter design (RPD), estimated response surfaces for the mean and the variance are obtained and based on the goal of the experiment an appropriate objective function determined. This is then optimized with respect to the control factors with the goal of finding the best settings at which the process should be carried out. In this chapter, we apply the Adaptive Penalized Likelihood Effects Selection (APLES) of Chapter 3 to simultaneously estimate the mean and variance response surfaces for use in RPD.

We suppose that there is a response variable, y, of interest obtained from the running of a process. There is an unknown function, f, relating the response to certain noise and control variable vectors. Let $\mathbf{x}$ represent the control factors and $\mathbf{z}$ represent the noise factors affecting the process. Every observable value of y is made with a zero mean, additive error ϵ. In the process, we have the relation:

$$y = f(\mathbf{x}, \mathbf{z}) + \epsilon$$

There are two sources of randomness on the right hand side: the noise factors as well as ϵ. Suppose an RPD experiment is performed. This may either be a classical Taguchi design or it could be a combined array experiment (described in section

2.2). In this experiment, the responses are obtained at fixed levels of the control as well as the noise variables.

If we knew f exactly, our next step would be to formulate a robust design problem. The criteria for formulating the robust design problem depend on the application, the goal of the process expert or engineer as well as various practical constraints that may affect the running of the process. We will focus on some of the more popular formulations in the RPD literature. Let the true mean function be $\mu_y(\mathbf{x}) = E_{\mathbf{z},\epsilon}\{f(\mathbf{x},\mathbf{z})\}$ and the true variance function be $\sigma_y^2(\mathbf{x}) = Var_{\mathbf{z},\epsilon}\{f(\mathbf{x},\mathbf{z})\}$ where the E and Var operators are with respect to the joint distribution of the noise variables as well as ϵ. Thus, μ_y and σ_y^2 are functions of the control variables only. In the following, let $\mathbf{x}_{\text{opt}}$ be the robust optimum, i.e. it is the solution to an optimization problem chosen to capture the goals of the experimenter: meet a target mean with as little variation in process response as possible. The following are three commonly used robust design criteria.

1. Robust Design Criterion 1 (assuming true mean and variance relationships known)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}} \ \left[\sigma_y^2(\mathbf{x})\right] \ \text{subject to } \mu_y(\mathbf{x}) = T \tag{4.1}$$

   where T is a target value for the mean response.

2. Robust Design Criterion 2 (assuming true mean and variance relationships known)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}} \ \left[(\mu_y - T)^2 + \sigma_y^2\right] \tag{4.2}$$

   where T is a target value for the mean response.

3. Robust Design Criterion 3 (assuming true mean and variance relationships known)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}} \ \left[\kappa_1(\mu_y) + \kappa_2\sigma_y\right] \tag{4.3}$$

   where κ_1 and κ_2 are parameters chosen by the engineer or process expert.

Vining and Myers (1990) proposed the first criterion as a solution to Taguchi's target-the-best (TTB) RPD problem. We will subsequently call it the TTB criterion. The second criterion is described in Lin and Tu (1995) as an improvement to criterion 1. The objective function is a mean square error (MSE) from the target, T. It permits deviations from the target (i.e. biased optima) as long as the MSE is kept at a minimum. The third formulation is attributed to Charnes and Cooper (1963) although several variants abound in the literature, with varying conditions on the choice of κ_1 and κ_2. This criterion is one possible solution when we are interested in simultaneously minimizing the mean and variance response, sometimes called minimum-the-best (MTB) RPD. We note that these three are merely a sample of the many possible criteria for solving the robust design problem.

Thus, the nature of the optimization problem differs from problem to problem. However, we may think of the objective function of each such optimization problem as a "quality" objective function, being the function which when optimized would provide settings of the control variable that would provide a quality product or process. This objective function usually combines the estimated mean and variance response surfaces similar to the robust design formulations shown above. However, it is usual to include extra terms to account for the parameter uncertainty in the estimation process (section 4.3.3.2).

The main difference between the model that we use in this chapter (equation (4.4)) and the model used in the usual approaches to RPD are that $\boldsymbol{\epsilon}(\mathbf{x})$ is a parametric function of $\mathbf{x}$. This results in a different variance function. The variance function is no longer a polynomial in $\mathbf{x}$. Most methods would estimate the mean function using least squares and the variance function using a plug-in or some modification of plug-in estimators resulting from a least squares fit. In this sense it is not fair to compare our approach to least squares when the data is generated according to our model. If the researcher knows in advance that the data is generated according to our model, then he would use MLE to estimate the mean and variance models. It is for this reason that we compare our approach to the full MLE.

Now, if the number of parameters then a penalized likelihood approach is desirable because the ML estimates do not exist. Indeed if the number of parameters is less than the sample size but is close to it, the ML estimates are extremely unsta-

ble. Intuitively, this makes sense because after the initial location fit the residuals are all close to zero. Thus, we compare our approach to the typical sample sizes that one finds in reported RPD datasets in the literature. n is larger than p, and the researcher knows that there is non-constant residual variance. In this situation, we seek to investigate using our simulation study how much better our procedure performs and in which settings.

Given the results of the RPD experiment, we build a predictive model $\hat{y} = \hat{f}(\mathbf{x}, \mathbf{z})$. Several approaches for building the predictive or surrogate model are available. This chapter will focus on comparing the following approaches to building the predictive models:

1. Response Surface Methodology (RSM) which uses OLS to fit a second-order regression model (usually in conjunction with a selection criterion such as C_p).

2. Maximum likelihood estimation: When there are enough data points it is possible to use maximum likelihood to estimate all of the parameters and then proceed with the resulting mean and variance functions to obtain robust solutions.

3. APLES: Developed in Chapter 3 of this dissertation may be used to obtain the predicted mean and variance response surfaces.

For each of these approaches to finding the best model we will look at the closeness of the optimum values of the formulated robust problem using a particular surrogate model to the optimum values of the corresponding robust problem using the true response.

The flowchart in figure 4.1 adapted from Simpson et al. (1998) illustrates the process of RPD. Our contributions in this chapter correspond to the steps that are circled.

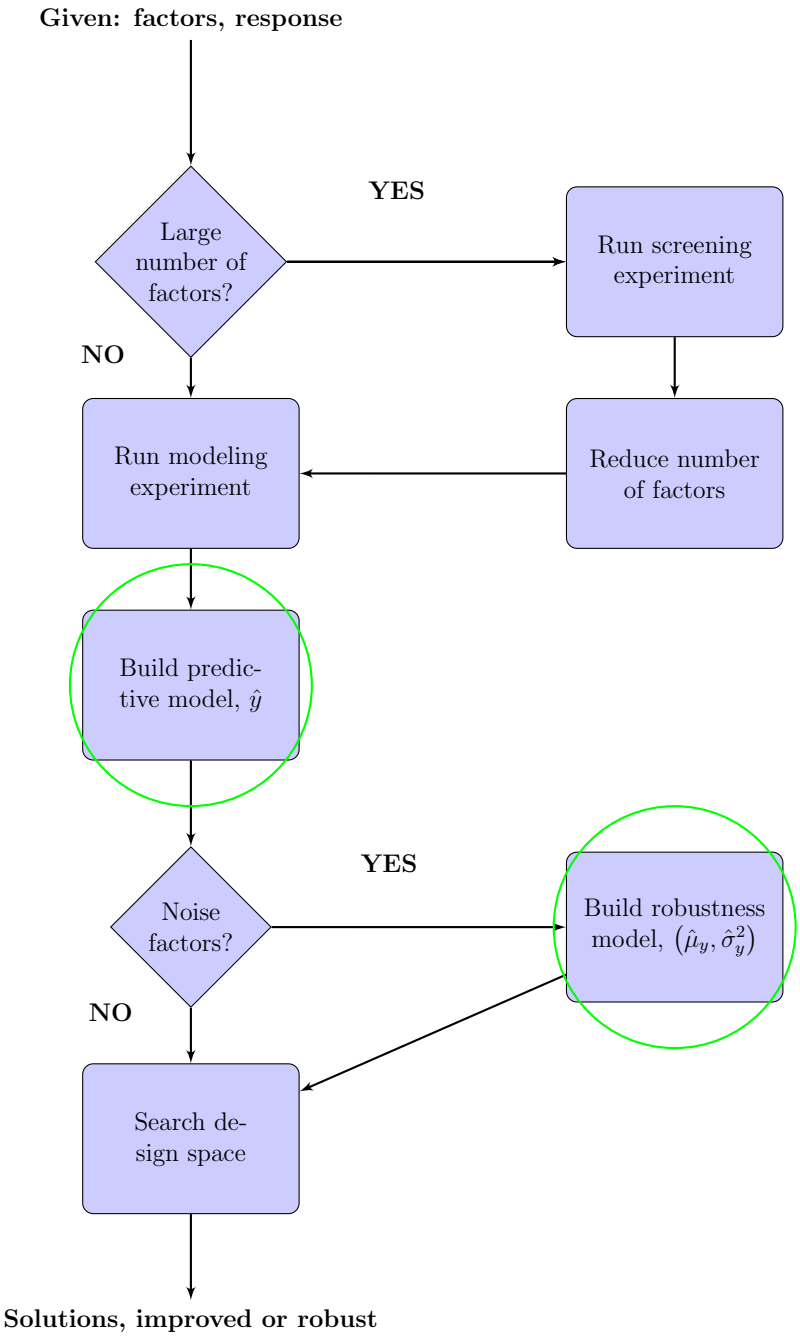

Figure 4.1. General Response Surface Modeling Approach to RPD

4.2 The heteroscedastic RPD Model

Consider a combined array RPD experiment with r_c control factors, $\mathbf{x}' = (x_1, x_2, ..., x_{r_c})$ and r_n noise factors, $\mathbf{z}' = (z_1, z_2, ..., z_{r_n})$. Thus for the ith factor settings in the experiment, we have:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ir_c} \end{pmatrix}$$

and

$$\mathbf{z}_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ir_n} \end{pmatrix}.$$

We assume the single response model (Engel and Huele, 1996a):

$$y_i(\mathbf{x}_i, \mathbf{z}_i) = \alpha_0 + \mathbf{x}_i'\boldsymbol{\alpha} + \mathbf{x}_i'\mathbf{A}\mathbf{x}_i + \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\Delta}\mathbf{z}_i + \epsilon_i(\mathbf{x}_i) \tag{4.4}$$

for $i = 1, \ldots, n$, where $y_i(\mathbf{x}_i, \mathbf{z}_i)$ is the observed response for a fixed setting of the control variables $\mathbf{x}_i$ and a fixed setting of the noise variables $\mathbf{z}_i$ and $\{\epsilon_i(\mathbf{x}_i), i = 1, \ldots, n\}$ is a set of independent normal random variables with mean 0 and variance σ_i^2. $\alpha_0, \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\gamma}$, and $\boldsymbol{\Delta}$ are the model parameters. Specifically, α_0 is the intercept, $\boldsymbol{\alpha}$ is the $r_c \times 1$ vector of control factor coefficients, $\mathbf{A}$ is the $r_c \times r_c$ matrix of second order coefficients for the control factor quadratic terms and interactions, $\boldsymbol{\gamma}$ is the $r_n \times 1$ vector of noise factor coefficients and $\boldsymbol{\Delta}$ is the $r_c \times r_n$ matrix of control by noise interaction coefficients.

$\mathbf{A}$ and $\boldsymbol{\Delta}$ are shown below:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \dots & \frac{1}{2}a_{1r_c} \\ \frac{1}{2}a_{12} & a_{22} & \dots & \frac{1}{2}a_{2r_c} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{1r_c} & \frac{1}{2}a_{2r_c} & \dots & a_{r_cr_c} \end{pmatrix}_{r_c \times r_c}$$

$$\mathbf{\Delta} = \begin{pmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1r_n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2r_n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{r_c,1} & \delta_{r_c,2} & \dots & \delta_{r_c,r_n} \end{pmatrix}_{r_c \times r_n}$$

The difference between model (4.4) and the original dual response approach by Vining and Myers (1990) is that the residual variance in model (4.4) is not constant, but allowed to depend on the control factors and the interactions between them. Thus, for fixed settings of the noise levels we have the usual heteroscedastic linear model $y_i = \mathbf{g}'(\mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\beta}^* + \epsilon_i$ with

$$\boldsymbol{\beta}^* = (\, \alpha_0, \alpha_1, \dots, \alpha_{r_c}, a_{11}, \dots, a_{r_cr_c}, \gamma_1, \dots, \gamma_{r_n}, \delta_{11}, \dots, \delta_{r_cr_n} \,) \tag{4.5}$$

$$\mathbf{g}'(\mathbf{x}, \mathbf{z}) = \left(1, \underbrace{x_1, x_2 \dots x_{r_c}, x_1x_2, x_1x_3 \dots x_{r_c-1}x_{r_c}, x_1^2, x_2^2 \dots x_{r_c}^2}_{\mathbf{u}(\mathbf{x})}, \underbrace{z_1, \dots, z_{r_n}}_{\mathbf{z}}, \underbrace{x_1z_1, \dots, x_{r_c}z_{r_n}}_{(\mathbf{x}*\mathbf{z})} \right) \tag{4.6}$$

and $\epsilon_i \sim N(0, \sigma_i^2(\mathbf{x}_i))$.
$\boldsymbol{\beta}^*$ contains $1 + \frac{r_c(r_c+3)}{2} + r_cr_n + r_n$ elements. This can be seen for example by writing the model in the form:

$$y_i\,(\mathbf{x}_i, \mathbf{z}_i) = \alpha_0 + \sum_{j=1}^{r_c} \alpha_j x_{ji} + \sum_{j=1}^{r_c} \sum_{j \le j'}^{r_c} a_{jj'} x_{ji} x_{j'i} + \sum_{k=1}^{r_n} \gamma_k z_{ki} + \sum_{j=1}^{r_c} \sum_{k=1}^{r_n} \delta_{jk} x_{ji} z_{ki} + \epsilon_i \tag{4.7}$$

There are $(r_c + 1)$ distinct α terms including the intercept α_0, $(\frac{r_c(r_c+1)}{2})$ distinct

terms in the matrix $\mathbf{A}$, r_n distinct γ terms and $(r_c r_n)$ distinct δ terms.

Let $\mathbf{X}$ be the design matrix whose ith row is $\mathbf{g}_i' = \mathbf{g}'(\mathbf{x}_i, \mathbf{z}_i)$. Then $\mathbf{X}$ is an $n \times k$ matrix, where k is the number of distinct elements of $\mathbf{g}_i$. $\mathbf{X}$ divides naturally into three sets of columns, the control matrix $\mathbf{U}(\mathbf{x})$, the noise matrix, $\mathbf{Z}$, and the control-by-noise interaction matrix, corresponding to the partitions indicated in equation (4.6). $\mathbf{u}'(\mathbf{x}_i)$, the ith row of $\mathbf{U}(\mathbf{x})$, is usually quadratic in the control factors $\mathbf{x}$ as shown above. However, in some problems, if the experimenter believes higher order terms are relevant and there are enough observations to estimate them then these could be included.

We assume the heteroscedastic log-linear model:

$$\sigma_i^2 = \exp\left(\mathbf{u}'(\mathbf{x}_i)\boldsymbol{\tau}^*\right) = \exp(\mathbf{u}_i'\boldsymbol{\tau}^*) \tag{4.8}$$

where $\boldsymbol{\tau}$ is the vector of dispersion parameters associated with the control matrix.

A distinguishing feature of the Taguchi RPD paradigm is that in the experiment, the levels of $\mathbf{z}$ are manipulated by the experimenter and therefore are not random. In the process, however, it is assumed that $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{z})$ with the (diagonal) covariance matrix $\boldsymbol{\Sigma}_\mathbf{z} = \text{diag}(\, \sigma_{Z_1}^2, \ldots, \sigma_{Z_{rn}}^2\,)$, known. However, this can be extended to the case of a more general covariance matrix (Dellino et al., 2010).

The covariance matrix $\boldsymbol{\Sigma}_\mathbf{z}$ has always been assumed to be known in the literature, and we follow this in our current work. However, we note that this is a very key assumption which can have an effect on the subsequent analysis when it is not satisfied. In actual fact, the elements of this matrix are estimated by process experts based on previous experimentation or some prior information.

Conditional on the the noise variables, the mean and variance are given by:

$$E_\epsilon(y(\mathbf{x}, \mathbf{z})|\mathbf{z}) = \alpha_0 + \mathbf{x}'\boldsymbol{\alpha} + \mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{z}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\Delta}\mathbf{z} \tag{4.9}$$

$$Var_\epsilon(y(\mathbf{x}, \mathbf{z})|\mathbf{z}) = \sigma_i^2 \tag{4.10}$$

The (unconditional) process mean and variance are given by:

$$\mu_y = E_{\mathbf{z},\epsilon}(y(\mathbf{x}, \mathbf{z})) = \alpha_0 + \mathbf{x}'\boldsymbol{\alpha} + \mathbf{x}'\mathbf{A}\mathbf{x} \tag{4.11}$$

$$\sigma_y^2 = Var_{\mathbf{z},\epsilon}(y(\mathbf{x}, \mathbf{z})) = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}_i)'\boldsymbol{\Sigma}_\mathbf{z}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}_i) + \sigma_i^2 \tag{4.12}$$

Expression (4.12) is obtained by applying the law of total variance or variance decomposition formula:

$$Var_{\mathbf{z},\epsilon}(y(\mathbf{x},\mathbf{z})) = Var_{\mathbf{z}}\left(E_{\epsilon}(y(\mathbf{x},\mathbf{z})|\mathbf{z})\right) + E_{\mathbf{z}}\left(Var_{\epsilon}(y(\mathbf{x},\mathbf{z})|\mathbf{z})\right) \tag{4.13}$$

Thus, even though we began with a single model (4.4), we have obtained two response surfaces (for the mean and the variance) and we can apply the dual response approach to obtain optimum settings of the control factors. There are two ways of controlling the process variance: either through those control factors that have a non-zero interaction with a noise factor or through those factors which affect the residual variance, i.e. factors with dispersion effects.

Although the heteroscedastic RPD model given in equation (4.4) was suggested in Engel and Huele (1996a), recent published work has focused more on the homoscedastic formulation (e.g. Matsuura et al. (2011) and Bingham and Nair (2012)).

4.2.1 Single Control Variable and Single Noise Variable

For clarification of the key components of the model, we consider the case where we have a single control variable x_1 and a single noise variable z_1. i.e. $r_c = 1, r_n = 1$.

The model in equation (4.4) is then given by:

$$y_i(x_{i1}, z_{i1}) = \alpha_0 + \alpha_1 x_{i1} + a_{11}x_{i1}^2 + \gamma_1 z_{i1} + \delta_{11}x_1 z_{i1} + \epsilon(x_{i1}) \tag{4.14}$$

Suppose this is a model for an industrial experiment where the engineers have pre-identified z_1 as a noise variable. Suppose they know that it has mean 0 and variance $\sigma_{z_1}^2$. We assume ϵ is $N(0, \sigma^2(x_1))$.
The four aspects of this model that are of interest for an RPD analysis are:

1. The conditional means model:

$$\begin{aligned} E_{\epsilon}(y|z_{i1}) &= E_{\epsilon}(\alpha_0 + \alpha_1 x_{i1} + a_{11}x_{i1}^2 + \gamma_1 z_{i1} + \delta_{11}x_1 z_{i1} + \epsilon(x_{i1})) \\ &= \alpha_0 + \alpha_1 x_{i1} + a_{11}x_{i1}^2 + \gamma_1 z_{i1} + \delta_{11}x_{i1} z_{i1} \end{aligned}$$

2. The conditional variance model:

$$\begin{aligned} Var_\epsilon(y|z_{i1}) &= Var_\epsilon(\alpha_0 + \alpha_1 x_{i1} + a_{11}x_{i1}^2 + \gamma_1 z_{i1} + \delta_{11}x_{i1}z_{i1} + \epsilon(x_{i1})) \\ &= \sigma^2(x_{i1}) \end{aligned}$$

3. The unconditional means model:

$$\begin{aligned} E(y(x_{i1}, z_{i1})) &= E_{z_{i1}}\left[E_\epsilon(y|z_{i1})\right] \\ &= \alpha_0 + \alpha_1 x_{i1} + a_{11}x_{i1}^2 \end{aligned}$$

4. The unconditional variance model:

$$\begin{aligned} Var(y(x_{i1}, z_{i1})) &= Var_{z_{i1}}\left[E_\epsilon(y|z_{i1})\right] + E_{z_{i1}}\left[Var_\epsilon(y|z_{i1})\right] \\ &= (\gamma_1 + \delta_{11}x_{i1})^2 var(z_{i1}) + \sigma^2(x_{i1}) = (\gamma_1 + \delta_{11}x_{i1})^2\sigma_{z_{i1}}^2 + \sigma^2(x_{i1}) \end{aligned}$$

The unconditional variance is the sum of two components: the variance of the conditional mean and the mean of the conditional variance. The term *heteroscedastic* used in describing equations (4.4) and (4.14) refers to the heteroscedasticity in this last term. This is because RPD is inherently heteroscedastic. Even if $\sigma^2(x_1)$ is a constant, say 1, the variance of the process can be tuned through the first term $(\gamma_1 + \delta_{11}x_1)^2\sigma_{z_1}^2$. This term only arises because of the interactions between the noise factors and some of the control factors. This is sometimes called the transmitted variance Wu and Hamada (2000). However, it is always possible for the heteroscedasticity to be a function of the control variable x_1 directly, without being transmitted through an interaction with a noise variable. Thus, we may have, for example $\sigma^2(x_1) = \psi x_1^2$, $\sigma^2(x_1) = \exp(x_1/\psi)$ or $\sigma^2(x_1) = |\psi x_1|$, for some parameter ψ.

Suppose we obtain data for the model in equation (4.14) and assume $\sigma^2(x_1) = \sigma^2$ is a constant. Then the fitted model will be:

$$\hat{y}(x_1, z_1) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{a}_{11}x_1^2 + \hat{\gamma}_1 z_1 + \hat{\delta}_{11}x_1 z_1$$

From the fitted model, a model for the process mean is given by:

$$\widehat{E(y|z_1)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{a}_{11} x_1^2$$

The unconditional variance of the responses is:

$$\widehat{Var(y(x_1, z_1))} = (\hat{\gamma}_1 + \hat{\delta}_{11} x_1)^2 \sigma_{z_1}^2 + \hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the residual mean square. If $\sigma_{z_1}^2 = 1$, we obtain:

$$\widehat{Var(y(x_1, z_1))} = (\hat{\gamma}_1 + \hat{\delta}_{11} x_1)^2 + \hat{\sigma}^2 = \hat{\gamma}_1^2 + \hat{\delta}_{11}^2 x_1^2 + 2\hat{\gamma}\hat{\delta}_{11} x_1 + \hat{\sigma}^2$$

Finally,

$$\widehat{Var(y(x_1, z_1))} = \hat{\gamma}_1^2 + \hat{\delta}_{11}^2 x_1^2 + 2\hat{\gamma}\hat{\delta}_{11} x_1 + \hat{\sigma}^2$$

These suggest a way to perform an RPD analysis. In this simple situation, we can find the value $x = x^*$ that minimizes the estimated variance or the mean square error from a certain target. In the general case, once the unconditional mean and unconditional variance are estimated, we can use optimization techniques to minimize the variance subject to a target value for the mean (Vining and Myers 1990; Del Castillo and Montgomery 1993; Lin and Tu 1995)..

Sometimes, the nature of the relationship between the control and noise variable is such that we can obtain a robust setting for the control variable without even solving an optimization problem. The following interaction plots (figure 4.2) adapted from Wu and Hamada (2000) help to illustrate this. The noise variable z and the control variable x each have two levels. In the topmost figure, suppose the objective is to maximize the response. Then it is clear that one should set the control to its high level. Even though there is greater variability at the high level of the control variable, one obtains a higher value for the response at $x = +$ over the entire range of the noise variable. In the middle figure, suppose the nominal target value for the response is 3 with specification limits of ± 2. Then one would set the control variable at its low level. For the bottom figure, again assuming a nominal target value of 3, it appears that neither the high or low levels of the

control variable is preferable.

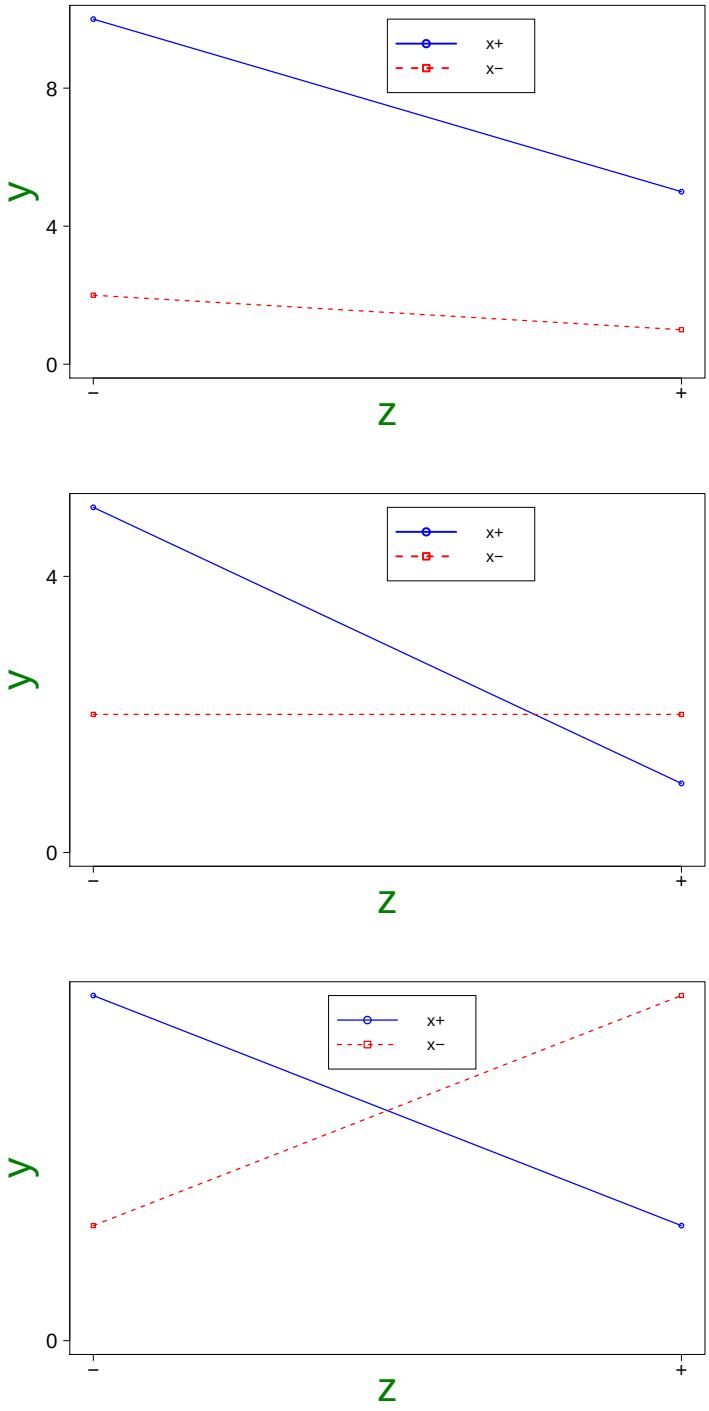

Figure 4.2. Interaction Plots for Single Control and Single Noise Variables.

4.3 APLES Applied to RPD

4.3.1 Motivation

Validation of response surface models is based on testing statistical hypotheses derived from error estimates of the variability in the data, plotting and checking the residuals, and computing criteria such as R^2 or C_p (Wu and Hamada, 2000). Penalized likelihood estimators are a competitor to the least squares estimator (which is the maximum likelihood estimator under normal, constant variance error structure) traditionally employed for fitting response surface models. The advantage of penalized likelihood estimators is the potential for reduced prediction error relative to the OLS estimators.

We describe the analysis of the heteroscedastic dual response RPD in the following steps:

0. Decide on an objective function based on the goal of the RPD experiment.

1. Identify the conditional mean model by determining which predictors or terms should be included in the conditional mean model.

2. Estimate the conditional mean response.

3. Identify the conditional variance model by determining which control factors and interactions between control factors should be included in the residual variance model.

4. Estimate the conditional variance response.

5. Based on steps one through four, form the unconditional mean and variance surfaces.

6. Optimize the objective function and obtain robust settings, $\mathbf{x}_{opt}$, for the control factors.

Thus, there are two variable selection problems identified in steps 1 and 3. We therefore utilize APLES developed in Chapter 3 which achieves the steps 1 through 4 in a single step: Identify and estimate the conditional mean and conditional variance response surfaces.

4.3.2 APLES estimators for the conditional process mean and process variance

Let $L(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$ be the likelihood for model (4.4) and $l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$ be the log-likelihood. Then $l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$ is given by:

$$l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*) = -\frac{n}{2} log 2\pi - \frac{1}{2}\sum_{i=1}^{n} \mathbf{u}_i'\boldsymbol{\tau}^* - \frac{1}{2}\sum_{i=1}^{n} e^{-\mathbf{u}_i'\boldsymbol{\tau}^*}(y_i - \mathbf{g}_i\boldsymbol{\beta}^*)^2 \tag{4.15}$$

and ignoring the constant term, we obtain:

$$-2l(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*) = \sum_{i=1}^{n} \mathbf{u}_i'\boldsymbol{\tau}^* + \sum_{i=1}^{n} e^{-\mathbf{u}_i'\boldsymbol{\tau}^*}(y_i - \mathbf{g}_i'\boldsymbol{\beta}^*)^2 \tag{4.16}$$

Let d denote the dimension of $\boldsymbol{\tau}^*$.

In Chapter 3, we proposed the Adaptive Penalized Likelihood Effects Selection (APLES) estimator which minimizes $-2l(\boldsymbol{\beta}, \boldsymbol{\tau})$ subject to bounds on the weighted L_1 norms of the location parameter vector $\boldsymbol{\beta}$ and the dispersion parameter vector $\boldsymbol{\tau}$. The APLES estimator, denoted,

$$\hat{\theta}_A = \begin{pmatrix} \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\tau}}_A \end{pmatrix} \tag{4.17}$$

is defined by

$$\underset{\boldsymbol{\beta},\boldsymbol{\tau}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \mathbf{u}_i'\boldsymbol{\tau} + \sum_{i=1}^{n} e^{-\mathbf{u}_i'\tau}(y_i - \mathbf{g}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{k} w_{j1}|\beta_j| + \lambda_2 \sum_{j=1}^{d} w_{j2}|\tau_j| \right) \tag{4.18}$$

where λ_1 and λ_2 are regularization parameters for the mean and variance respectively.

The APLES estimator possesses the following desirable properties: It has variable selection consistency. It has the oracle property in $\boldsymbol{\tau}, \boldsymbol{\beta}$. This means it performs as well as if the true model were known in advance. This is because the APLES estimator solves two adaptive LASSO (Zou, 2006) problems: A weighted least squares regression with adaptive LASSO penalty followed by a penalized likelihood problem with adaptive LASSO penalty. An efficient fitting algorithm for

solving equation (4.18) as well as further properties of APLES can be found in Chapter 3. In the rest of this chapter, we focus on applying the APLES procedure for achieving RPD.

4.3.3 Some APLES-based objective functions RPD

4.3.3.1 Certainty Equivalence

Given the two response surfaces, the goal of the experimenter determines the type of RPD problem that is formulated and solved. For example, in many processes the goal is to minimize the response variance while keeping the mean response as close as possible to a pre-defined target, T. This is known as target-the-best RPD. To achieve target-the-best RPD, Vining and Myers (1990) suggested solving the problem (4.1). As an alternative, Lin and Tu (1995) formulated the MSE criterion (4.2). The MSE criterion has the advantage that it may permit solutions, which although not meeting the target exactly are close enough to be deemed acceptable and also lead to far lesser variances than criterion (4.1).

In actual fact, the process mean and variance are unknown and have to be replaced with estimates, $\hat{\mu}_y$, and $\hat{\sigma}_y^2$. This situation is known as Certainty Equivalence (CE) and optimal solutions are denoted $\mathbf{x}_{CE}$ (Apley and Kim, 2011). Thus corresponding to the three robust design criteria in section (4.1) we will use the following in our simulations:

1. Robust Design Criterion 1, Target-the-best (assuming Certainty Equivalence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ \hat{\sigma}_y^2(\mathbf{x}) \text{ subject to } \hat{\mu}_y(\mathbf{x}) = T \tag{4.19}$$

2. Robust Design Criterion 2, Mean Square Error (assuming Certainty Equivalence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ (\hat{\mu}_y - T)^2 + \hat{\sigma}_y^2 \tag{4.20}$$

3. Robust Design Criterion 3, Minimum-the-best (assuming Certainty Equiva-

lence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ \kappa_1(\hat{\mu}_y) + \kappa_2\hat{\sigma}_y \tag{4.21}$$

In the tables summarizing our simulation results in section (4.5) below, we will refer to these as Target-The-Best (TTB-CE) under Certainty Equivalence, Mean Square Error under Certainty Equivalence (MSE-CE) and Minimum-The-Best under Certainty Equivalence (MTB-CE) respectively. It has been noted in the literature that these solutions are generally too optimistic as they do not account for the parameter estimation uncertainty (Miro-Quesada and Del Castillo 2004; Apley and Kim 2011). In fact they may lead to settings which do not meet the target and have high variance for the response.

4.3.3.2 Parameter Estimation Uncertainty

When we do not have Certainty Equivalence, we must use an objective function which accounts for the uncertainty in parameter estimation. Several approaches exist in the literature. Apley and Kim (2011) minimize the Bayesian MSE and arrive at an objective function which they call the cautious robust design (CRD) objective function. Vanli and Del Castillo (2009) also propose two Bayesian approaches in the context of on-line RPD. We take a frequentist approach and we describe next a proposal by Miro-Quesada and Del Castillo (2004) which we will modify to fit our heteroscedastic framework.

Miro-Quesada and Del Castillo (2004) propose minimizing the variance of the predicted response where the variance is taken with respect to the parameter estimates as well as the noise factors subject to the mean being on target. They obtain an unbiased estimate of the variance of the predicted response. Thus, the minimization problem that they proposed is (in the notation of this dissertation): Minimize with respect to $\mathbf{x}$

$$\widehat{Var}_{\mathbf{z},\hat{\boldsymbol{\beta}}}\left[\hat{y}(\mathbf{x},\mathbf{z})\right] = (\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}) + \hat{\sigma}_\epsilon^2\left[\mathbf{u}'(\mathbf{x})(\mathbf{U}'(\mathbf{x})\mathbf{U}(\mathbf{x}))^{-1}\mathbf{u}(\mathbf{x})\right] \tag{4.22}$$

Subject to :

$$\widehat{E}_{\mathbf{z},\epsilon}\left[y(\mathbf{x},\mathbf{z})\right] \equiv \hat{\alpha}_0 + \mathbf{x}'\hat{\boldsymbol{\alpha}} + \mathbf{x}'\hat{\mathbf{A}}\mathbf{x} = T \tag{4.23}$$

where $\mathbf{u}'(\mathbf{x})$ is a row of $\mathbf{U}(\mathbf{x})$ defined in section 4.2. This can be extended to an MSE criterion as: Minimize with respect to $\mathbf{x}$

$$\underbrace{(\hat{\alpha}_0 + \mathbf{x}'\hat{\boldsymbol{\alpha}} + \mathbf{x}'\hat{\mathbf{A}}\mathbf{x} - T)^2}_{\text{squared deviation from target}} + \underbrace{(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})}_{\text{Plug-in variance estimate}} + \underbrace{\hat{\sigma}^2_{\epsilon}\left[\mathbf{u}'(\mathbf{x})(\mathbf{U}'(\mathbf{x})\mathbf{U}(\mathbf{x}))^{-1}\mathbf{u}(\mathbf{x})\right]}_{\text{accounts for estimation uncertainty}} \tag{4.24}$$

It is worth noting that the expression $\hat{\sigma}^2_{\epsilon}\left[\mathbf{u}'(\mathbf{x})(\mathbf{U}'(\mathbf{x})\mathbf{U}(\mathbf{x}))^{-1}\mathbf{u}(\mathbf{x})\right]$ in the objective function prevents optimal solutions that are too far out in $\mathbf{x}$ space. Thus in this sense this objective function provides some constraint on the solution.

Under heteroscedasticity, the expression $\hat{\sigma}^2_{\epsilon}\left[\mathbf{u}'(\mathbf{x})(\mathbf{U}'(\mathbf{x})\mathbf{U}(\mathbf{x}))^{-1}\mathbf{u}(\mathbf{x})\right]$ may be replaced by $\mathbf{u}'(\mathbf{x})\left[\mathbf{U}'(\mathbf{x})\hat{\boldsymbol{\Sigma}}^{-1}_{\epsilon}\mathbf{U}(\mathbf{x})\right]^{-1}\mathbf{u}(\mathbf{x})$ where $\hat{\boldsymbol{\Sigma}}^{-1}_{\epsilon}$ is a diagonal matrix with ith diagonal element $e^{-\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A}$. In addition, the plug-in variance estimate may be replaced by $(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}) + e^{\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A}$.

Therefore, under heteroscedasticity and using APLES, we have the following MSE objective function for RPD which accounts for estimation uncertainty:

$$\begin{aligned}&(\hat{\alpha}_{0A} + \mathbf{x}'\hat{\boldsymbol{\alpha}}_A + \mathbf{x}'\hat{\mathbf{A}}_A\mathbf{x} - T)^2 + (\hat{\boldsymbol{\gamma}}_A + \hat{\boldsymbol{\Delta}}'_A\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}}_A + \hat{\boldsymbol{\Delta}}'_A\mathbf{x}) + e^{\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A}\\ &+ \mathbf{u}'(\mathbf{x})\left[\mathbf{U}'(\mathbf{x})\hat{\boldsymbol{\Sigma}}^{-1}_{\epsilon}\mathbf{U}(\mathbf{x})\right]^{-1}\mathbf{u}(\mathbf{x})\end{aligned} \tag{4.25}$$

where $\hat{\alpha}_{0A}, \hat{\boldsymbol{\alpha}}_A, \hat{\mathbf{A}}_A, \hat{\boldsymbol{\gamma}}_A$, and $\hat{\boldsymbol{\Delta}}_A$ are the APLES estimates of $\alpha_0, \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\gamma}$, and $\boldsymbol{\Delta}$. $\hat{\boldsymbol{\Sigma}}^{-1}_{\epsilon}$ is a diagonal matrix with ith diagonal element $e^{-\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A}$ and $\hat{\boldsymbol{\tau}}_A$ is the APLES estimate of $\boldsymbol{\tau}$.

In the general case (i.e. not necessarily using APLES for the estimates), we have the following robust design formulations:

1. Robust Design Criterion 1, Target-the-best (not assuming Certainty Equiv-

alence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ (\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}) + e^{\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A} + \mathbf{u}'(\mathbf{x})\left[\mathbf{U}'(\mathbf{x})\hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1}\mathbf{U}(\mathbf{x})\right]^{-1}\mathbf{u}(\mathbf{x})$$
$$\text{subject to } \hat{\alpha}_0 + \mathbf{x}'\hat{\boldsymbol{\alpha}} + \mathbf{x}'\hat{\mathbf{A}}\mathbf{x} = T \tag{4.26}$$

2. Robust Design Criterion 2, Mean Square Error (not assuming Certainty Equivalence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ (\hat{\alpha}_0 + \mathbf{x}'\hat{\boldsymbol{\alpha}} + \mathbf{x}'\hat{\mathbf{A}}\mathbf{x} - T)^2 + (\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}) + e^{\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A}$$
$$+ \mathbf{u}'(\mathbf{x})\left[\mathbf{U}'(\mathbf{x})\hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1}\mathbf{U}(\mathbf{x})\right]^{-1}\mathbf{u}(\mathbf{x}) \tag{4.27}$$

3. Robust Design Criterion 3, Minimum-the-best (not assuming Certainty Equivalence)

$$\mathbf{x}_{\text{opt}} = \underset{\mathbf{x}}{\text{argmin}}\ \kappa_1(\hat{\alpha}_0 + \mathbf{x}'\hat{\boldsymbol{\alpha}} + \mathbf{x}'\hat{\mathbf{A}}\mathbf{x})$$
$$+ \kappa_2\sqrt{(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\Delta}}'\mathbf{x}) + e^{\mathbf{u}_i'\hat{\boldsymbol{\tau}}_A} + \mathbf{u}'(\mathbf{x})\left[\mathbf{U}'(\mathbf{x})\hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1}\mathbf{U}(\mathbf{x})\right]^{-1}\mathbf{u}(\mathbf{x})} \tag{4.28}$$

We will refer to these as Target-The-Best (TTB-U) under Uncertainty, Mean Square Error under Uncertainty (MSE-U) and Minimum-The-Best under Uncertainty (MTB-U), respectively.

4.4 Examples

We now apply and illustrate APLES using several well known examples in the literature to compare with other methods.

4.4.1 Example 1: Injection Molding Experiment

The first example we consider is the injection molding experiment presented and analyzed in several works on RPD such as Engel (1992), Steinberg and Bursztyn (1994), Engel and Huele (1996a), and most recently in Bingham and Nair (2012). Engel and Huele (1996a) and Lee and Nelder (1998) formulate the problem as a

heteroscedastic RPD problem and so we will subsequently compare our results with theirs. The response is the percentage of shrinkage and the goal of the experiment is to determine the settings of the control factors which bring the response to a certain target value while being robust against environmental variation, such as ambient temperature. The experimental design is a saturated product design in which the control factors and noise factors are varied according to a $2^7 \times 2^3$ design. To use Taguchi's terminology, this is an $L_8 \times L_4$ design. The data is shown in Table 4.1.

								M			
								-1	-1	1	1
								N			
								-1	1	-1	1
								O			
Run	A	B	C	D	E	F	G	-1	1	1	-1
1	-1	-1	-1	-1	-1	-1	-1	2.2	2.1	2.3	2.3
2	-1	-1	-1	1	1	1	1	2.5	0.3	2.5	0.3
3	-1	1	1	-1	-1	1	1	0.5	3.1	0.4	2.8
4	-1	1	1	1	1	-1	-1	2.0	1.9	1.8	2.0
5	1	-1	1	-1	1	-1	1	3.0	3.1	3.0	3.0
6	1	-1	1	1	-1	1	-1	2.1	4.2	1.0	3.1
7	1	1	-1	-1	1	1	-1	4.0	1.9	4.6	2.2
8	1	1	-1	1	-1	-1	1	2.0	1.9	1.9	1.8

Table 4.1. Data for Injection Molding Experiment (Experiment 1)

The control factors are: A = cycle time, B = mold temperature, C = cavity thickness, D = holding pressure, E = injection speed, F = holding time, G = gate size , and the noise factors are: M = percentage regrind, N = moisture content and O = ambient temperature. Steinberg and Bursztyn (1994) note that the design can be regarded as a 2^{10-5}_{III} design, with defining relation $I = -ABC = -ADE = -BDF = ABDG = -MNO$. This alias structure allows estimation of all main effects and all two-factor interactions between design and noise factors.

Even though a specific target value is not assumed in any of these papers, for the purposes of comparison, we will assume in our subsequent discussion that this is $T = 2.25$. We chose this because it is the mean of the responses at the actual settings the experiment was performed at. To ensure that the results are not sensitive to this choice of target we also performed the analysis with several values in the range from $T = 0.00$ to $T = 3.00$.

4.4.1.1 Results Assuming Certainty Equivalence

The analysis by Engel and Huele (1996a) identified A, D, G, CN and EN as having significant effects on the location and A as having a significant effect on the variance. Based on this, they fitted the mean model and the variance model shown in Table 4.2 and Table 4.3 using maximum likelihood estimation.

Thus, using their model and assuming Certainty Equivalence and the MSE objective function, the RPD problem is to find A, B, C, D, E, F and G to minimize:

$$(2.25 + 0.43A - 0.15D - 0.25G - T)^2 + (0.60C - 0.58E)^2\sigma_N^2 + \exp(-2.35 + 1.41A)$$

The solution is given by A=-1, D=-0.31, G=-0.53, C=E and B, F may take on any values. We have assumed a cuboidal region of interest, $-1 \leq x_i \leq 1$. We have also assumed $\sigma_N^2 = 1$.

The identification of the factor A as having a dispersion effect by Engel and Huele (1996a) is based on fitting a GLM with the squared residuals and using both a normal plot of the estimated effects as well as the Wald test. As several authors have noted, the outcome of such a procedure is highly dependent on the pre-selected location model. Thus to be entirely sure that the mean and variance submodels chosen are the best, one would have to alternatively search through the entire space of location models and find for each location model, the best dispersion model. However, it is this process of alternatively searching through the entire model space and refitting several possible models that the procedure developed in Chapter 3 and used in this chapter avoids. We now proceed to implement our proposed procedure for this dataset.

We consider as possible location terms all 32 terms consisting of intercept, 7 control factor main effects, 3 noise factor main effects, as well as all 21 control by noise factor interactions. We consider 8 terms as possible dispersion terms (the intercept and the 7 control factors). Thus, we have $k = 32$ and $d = 8$. The design matrix does not contain center points and therefore cannot estimate any quadratic effects in the control factors.

Tables 4.2 and 4.3 below shows the best selected models using APLES with the corrected AIC (AICc) criterion.

Term	APLES	s.e.	Engel and Huele (1996a)	s.e.	Lee and Nelder (1998)	s.e.
1	2.250	0.101	2.250	0.104	0.800	0.010
A	0.411	0.169	0.430	0.168	0.149	0.035
C	0.000	0.000	0.000	0.000	0.069	0.031
D	-0.260	0.101	-0.150	0.168	-0.152	0.010
E	0.103	0.101	0.000	0.000	0.012	0.032
G	-0.206	0.169	-0.250	0.168	-0.074	0.035
N	0.000	0.000	0.000	0.000	-0.006	0.008
CN	0.577	0.169	0.610	0.168	0.189	0.034
EN	-0.546	0.169	-0.580	0.168	-0.173	0.034

Table 4.2. Comparison of Mean Models Selected: Injection Molding Experiment

Term	APLES	s.e.	Engel and Huele (1996a)	s.e.	Lee and Nelder (1998)	s.e.
1	-3.784	0.244	-3.170	0.250	-2.849	0.300
A	0.707	0.230	1.530	0.250	-0.608	0.296
C	0.351	0.231	0.000	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	2.324	0.300

Table 4.3. Comparison of Variance Models Selected: Injection Molding Experiment

We see that in addition to factor A, APLES identifies factor C as having a dispersion effect. Also, APLES includes the main effect of factor E in the mean model whereas Engel and Huele (1996a) does not.

Next, we compare the sets of response surfaces, Engel and Huele (1996a) and APLES in terms of which settings of the control factors they identify for RPD. Under Certainty Equivalence, optimization of the MSE objective function using the response surfaces obtained using APLES gives the optimal solution: A=-1, C=-1, D=-0.18, E=-1, G=-1, and B, F may take on any values. The major difference between this solution and that from Engel and Huele (1996a) is for APLES there are specific values of C and E to obtain RPD rather than the requirement that they simply be equal. The objective functions (MSE) are plotted in Figure (4.3) as functions of C and E, with all other factors held at their optimum values and as functions of A and G, with all other factors held at their optimum values in Figure (4.4).

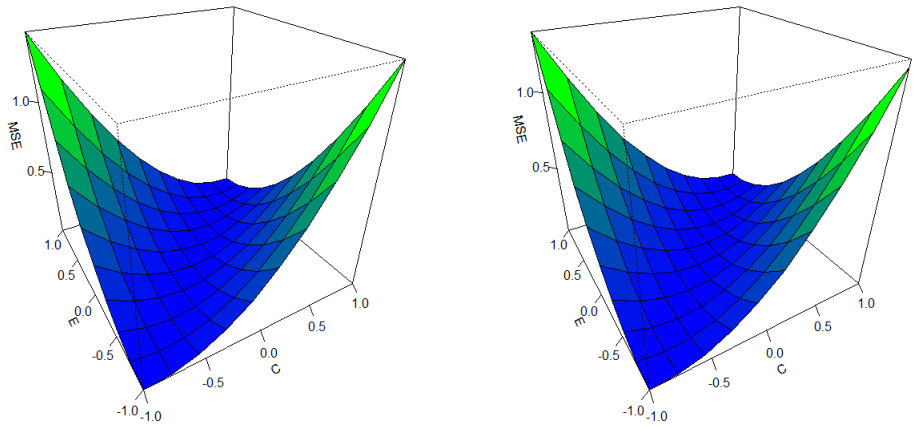

Figure 4.3. MSE Objective Functions Under Certainty Equivalence Versus Factors C and E (Engel-Huele On Left, APLES On Right)

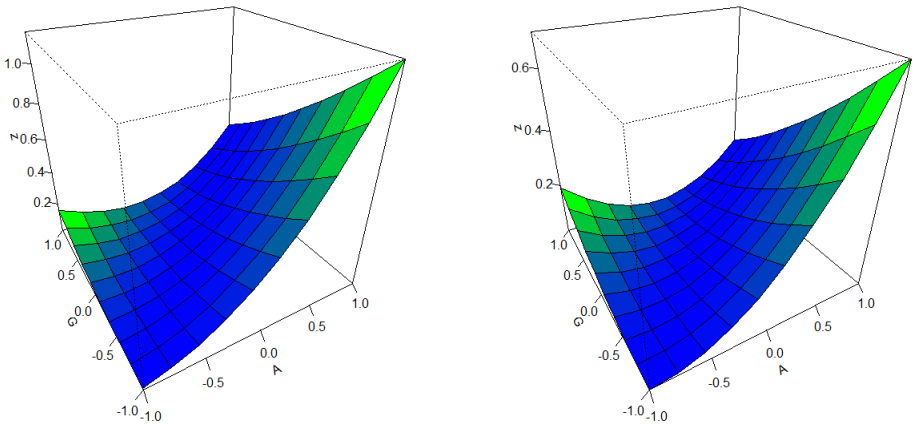

Figure 4.4. MSE Objective Functions Under Certainty Equivalence Versus Factors A and G (Engel-Huele On Left, APLES On Right)

4.4.1.2 Results Considering Parameter Estimation Uncertainty

Now let us consider parameter uncertainty, so that we now use the objective function (4.27). The optimum values, $\mathbf{x}_{\text{opt}}$, for Engel and Huele (1996a) and APLES

are shown in Table 4.4.

	A	B	C	D	E	F	G
Engel and Huele (1996a)	-1.00	0.47	0.46	0.47	0.46	-1.00	-1.00
APLES	-0.42	0.56	-0.22	0.56	-0.22	-1.00	-0.41

Table 4.4. Optimum Settings When Minimizing MSE Objective Function Under Parameter Estimation Uncertainty

The objective functions (MSE) are plotted below as functions of C and E, with all other factors held at their optimum values (figure 4.5) and as functions of A and G, with all other factors held at their optimum values (figure 4.6).

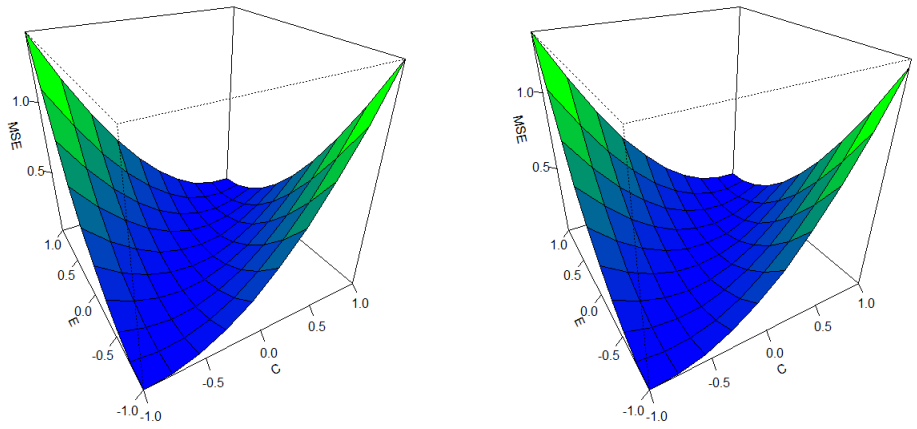

Figure 4.5. MSE Objective Functions Under Parameter Estimation Uncertainty Versus Factors C and E (Engel-Huele On Left, APLES On Right)

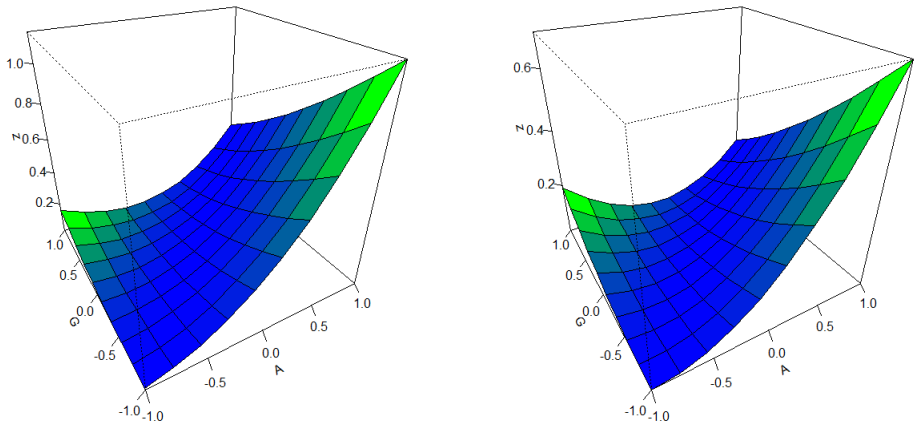

Figure 4.6. MSE Objective Functions Under Parameter Estimation Uncertainty Versus Factors A and G (Engel-Huele On Left, APLES On Right)

4.4.1.3 Discussion

We illustrated with this example the application of the APLES estimation procedure on a well-known dataset. In figure 4.7 we show a plot of the MSE objective function versus the target value T for values of T varying from 0 to 3. The plot shows minimal values in the MSE objective function for values of the target greater than 1.5. This indicates that the optima of the MSE objective function seems to be insensitive to the estimation method used to generate the response surfaces for these values of the target. Intuitively, these targets are easy to meet with little "cost" in terms of MSE. For target values between 0 and 1.5 however, the Lee and Nelder (1998) MSE values are much higher than for APLES and the Engel and Huele (1996a) values.

We also performed an analysis for this dataset when the objective function is the MTB objective function. Table 4.5 provides the optimizing points of the design for each approach and table 4.6 provides the value of the MTB objective functions at the respective optima. Even though the optima differ (in the settings of factors A, C and E), the actual value of the MTB objective functions at the respective optima do not differ greatly, as seen in table 4.5.

Factor	APLES	Engel and Huele (1996a)	Lee and Nelder (1998)
A	0.86	0.46	1.00
B	0.01	0.01	0.01
C	-0.96	0.01	-0.16
D	-1.00	-1.00	-1.00
E	-1.00	0.01	-0.16
F	0.01	0.01	0.01
G	-1.00	-1.00	-1.00

Table 4.5. Comparison of Optimal Factor Settings Using MTB Criterion Under Certainty Equivalence

APLES	Engel and Huele (1996a)	Lee and Nelder (1998)
1.57	1.57	1.51

Table 4.6. Value of MTB Objective at Optima

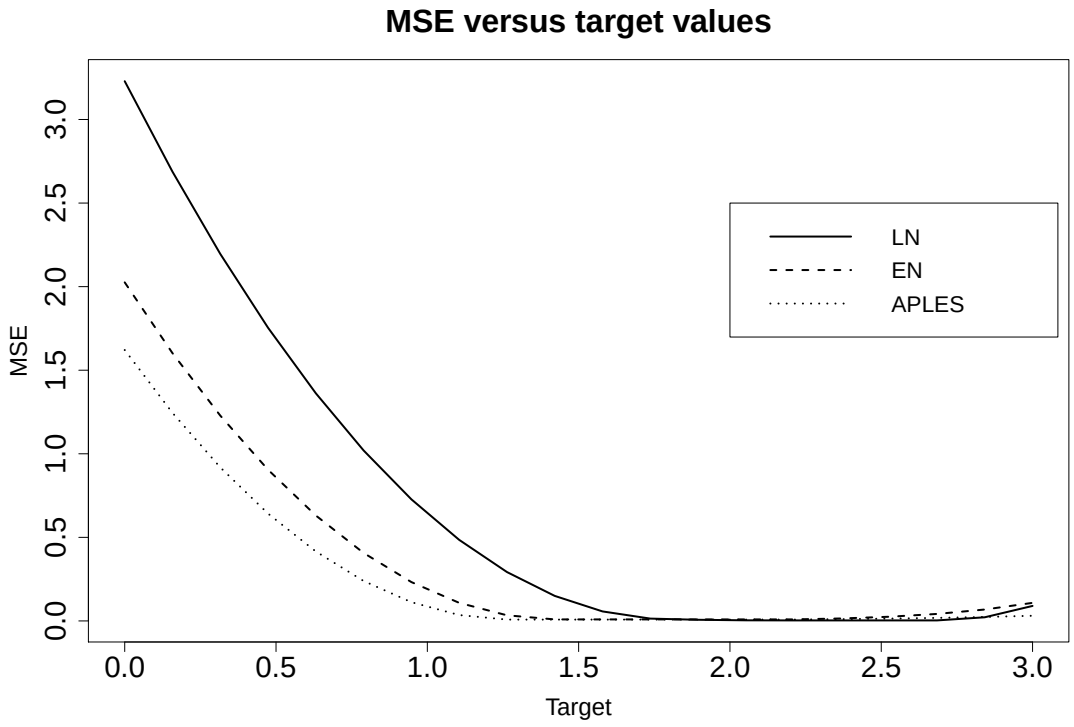

Figure 4.7. MSE Versus Target Values

4.4.2 Example 2: Printing Process Experiment

The printing process experiment, described in Box and Draper (1987), is also a very popular dataset that has been well analyzed in the RPD literature (Vining and Myers 1990; Lin and Tu 1995; Pickle et al. 2008). This experiment involves three control factors x_1, x_2, x_3, representing speed, pressure and distance respectively. The response is a quality characteristic for the performance of a printing process, specifically, a machine's ability to apply colored inks to package labels. The experimental design is a 3^3 design with three replicates at each point and thus 81 observations in all. There are no noise factors present in the experiment. The target value for the response is $T = 500$. Thus the ultimate goal from the point of view of RPD is to choose settings of x_1, x_2, x_3 to bring the mean of the response y as close as possible to 500 while keeping its variance to a minimum. The data is presented in Table 4.7.

	x_1	x_2	x_3	y_1	y_2	y_3	$\bar{y}$	s
1	-1	-1	-1	34.00	10.00	28.00	24.00	12.49
2	0	-1	-1	115.00	116.00	130.00	120.33	8.39
3	1	-1	-1	192.00	186.00	263.00	213.67	42.83
4	-1	0	-1	82.00	88.00	88.00	86.00	3.46
5	0	0	-1	44.00	178.00	188.00	136.67	80.41
6	1	0	-1	322.00	350.00	350.00	340.67	16.17
7	-1	1	-1	141.00	110.00	86.00	112.33	27.57
8	0	1	-1	259.00	251.00	259.00	256.33	4.62
9	1	1	-1	290.00	280.00	245.00	271.67	23.63
10	-1	-1	0	81.00	81.00	81.00	81.00	0.00
11	0	-1	0	90.00	122.00	93.00	101.67	17.67
12	1	-1	0	319.00	376.00	376.00	357.00	32.91
13	-1	0	0	180.00	180.00	154.00	171.33	15.01
14	0	0	0	372.00	372.00	372.00	372.00	0.00
15	1	0	0	541.00	568.00	396.00	501.67	92.50
16	-1	1	0	288.00	192.00	312.00	264.00	63.50
17	0	1	0	432.00	336.00	513.00	427.00	88.61
18	1	1	0	713.00	725.00	754.00	730.67	21.08
19	-1	-1	1	364.00	99.00	199.00	220.67	133.82
20	0	-1	1	232.00	221.00	266.00	239.67	23.46
21	1	-1	1	408.00	415.00	443.00	422.00	18.52
22	-1	0	1	182.00	233.00	182.00	199.00	29.44
23	0	0	1	507.00	515.00	434.00	485.33	44.64
24	1	0	1	846.00	535.00	640.00	673.67	158.21
25	-1	1	1	236.00	126.00	168.00	176.67	55.51
26	0	1	1	660.00	440.00	403.00	501.00	138.94
27	1	1	1	878.00	991.00	1161.00	1010.00	142.45

Table 4.7. Data for Printing Process Experiment (Example 2)

Previous authors (e.g. Vining and Myers (1990) and Lin and Tu (1995)) have used the standard deviation of the three replicates at each setting to estimate a response surface for the standard deviation of the process. This is then squared to obtain the variance response surface. Due to the small number of replicates, this measure can be unreliable and may not accurately reflect the variability for that setting as a function of the control factors. We will present a comparison of the results of Vining and Myers (1990) and Lin and Tu (1995) with the results of an analysis using APLES. We will compare the estimated response surfaces obtained as well as the optimum values $\mathbf{x}_{\text{opt}} = (x_1, x_2, x_3)$ obtained after plugging the estimated response surfaces into the Robust Design Criteria 1 and 2 (i.e. the TTB and MSE criteria respectively).

The columns of the design matrix correspond to the terms:

$$\left(1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2, x_1x_2x_3\right)$$

The mean model can possibly have any of the 10 terms listed above and similarly for the variance model. Thus there are potentially 20 coefficients to be estimated. It is possible to estimate all coefficients because the 27 unique $\mathbf{x}$ settings at which responses are taken is larger than 20. Vining and Myers (1990) fit full quadratic models for the mean and the standard deviation, thus the coefficient of the three-factor interaction is 0 for their model. Lin and Tu (1995) use different model selection procedures to obtain the best subset model. We apply APLES to select the best means model using the AICc. Tables 4.8 and 4.9 show three different models, the first by Vining and Myers (1990), the second by Lin and Tu (1995) and the third is obtained by APLES using AICc to select the tuning parameter. We label the Vining and Myers (1990) results as VM and the Lin and Tu (1995) results as LT.

Term	VM	LT	APLES(AICc)
1	327.60 (25.63)	314.70 (21.52)	317.00 (23.22)
x_1	177.00 (11.86)	177.00 (9.96)	180.70 (10.84)
x_2	109.40 (11.86)	109.40 (9.96)	108.50 (10.84)
x_3	131.50 (11.86)	131.50 (9.96)	132.60 (10.84)
x_1x_2	66.00 (14.53)	66.00 (12.20)	53.60 (13.23)
x_1x_3	75.50 (14.53)	75.50 (12.20)	81.90 (13.23)
x_2x_3	43.60 (14.53)	43.60 (12.20)	65.10 (13.23)
x_1^2	32.00 (20.55)	0.00 (0.00)	33.80 (20.45)
x_2^2	-22.40 (20.55)	0.00 (0.00)	-19.50 (20.45)
x_3^2	-29.10 (20.55)	0.00 (0.00)	-20.90 (20.45)
$x_1x_2x_3$	0.00 (0.00)	82.80 (17.25)	64.80 (15.15)

Table 4.8. Comparison of Mean Model Selected for Example 2

Term	VM	LT	APLES
1	34.90 (22.32)	47.99 (7.21)	8.72 (0.30)
x_1	11.50 (10.33)	11.53 (8.83)	0.26 (0.14)
x_2	15.30 (10.33)	15.32 (8.83)	0.57 (0.14)
x_3	29.20 (10.33)	29.19 (8.83)	0.64 (0.14)
x_1x_2	7.70 (12.65)	0.00 (0.00)	0.00 (0.00)
x_1x_3	5.10 (12.65)	0.00 (0.00)	0.00 (0.00)
x_2x_3	14.10 (12.65)	0.00 (0.00)	0.00 (0.00)
x_1^2	4.20 (17.89)	0.00 (0.00)	-0.61 (0.00)
x_2^2	-1.30 (17.89)	0.00 (0.00)	0.00 (0.00)
x_3^2	16.80 (17.89)	0.00 (0.00)	0.00 (0.00)
$x_1x_2x_3$	0.00 (0.00)	29.57 (13.24)	1.33 (0.24)

Table 4.9. Comparison of Variance Model Selected for Example 2

When assuming Certainty Equivalence, the optimal setting for $\mathbf{x}$ using the MSE criterion and assuming a cuboidal region of interest $-1 \leq x_i \leq 1$ are shown in the Table 4.10.

Factor	VM	LT	APLES
x_1	1.00	1.00	1.00
x_2	-0.05	1.00	-0.76
x_3	-0.17	-0.52	1.00

Table 4.10. Optimum Factor Settings for Example 2 Under Certainty Equivalence and using the MSE criterion

Table 4.11 shows the best values using the TTB criterion and assuming a cuboidal region of interest $-1 \leq x_i \leq 1$.

Factor	VM	LT	APLES
x_1	1.00	1.00	1.00
x_2	-0.01	1.00	-0.74
x_3	-0.16	-0.50	1.00

Table 4.11. Optimum Factor Settings for Example 2 Under Certainty Equivalence and using the TTB criterion

We draw the following conclusions from this comparison. For a fixed selected model (VM, LT or APLES), the optimal factor settings corresponding to MSE and TTB criteria are quite close to each other. The optimal factor settings are however very different for VM, LT or APLES. Since this is an example for which we do not know the true model, we cannot make a conclusive statement on the closeness of these predicted optimal settings to the 'true' one. This motivates us to perform a simulation study in which we know the true model and the 'true' optimal factor settings.

4.5 Simulation Study

In this section we present the results of a simulation experiment to compare the APLES estimator to the ordinary least squares estimator and the maximum likelihood estimator.

Specifically, the three methods that we compare are:

1. Maximum likelihood estimation for both location and dispersion effects (MLE).

2. Least squares estimation for location effects ignoring the fact that $\epsilon(\mathbf{x})$ may be a function of the control factors (LS).

3. APLES estimation with the tuning parameter chosen using AICc.

We also present results for the maximum likelihood estimator based on an oracle (i.e. one based on knowledge of which coefficients are exactly zero). We call this estimator MLO in our subsequent results.

Our reason to compare the performance of APLES in RPD to these two procedures is because they are the two methods that are usually used for the analysis of designed experiments. The least squares estimates have been used traditionally in analysis of data from designed experiments (including the RSM approach to RPD) when we have constant residual variance . Usually, a good model is chosen using stepwise regression together with Cp or AIC. When the constant residual variance assumption is not met, the use of maximum likelihood to estimate the elements of $\boldsymbol{\tau}$ is the usual choice.

4.5.1 Goals of the Simulation Study

The ultimate goal of this simulation is to study the performance of these methods in obtaining robust settings of the control factors. By a robust setting of the control factors, we refer to levels of the control factors which lead to the attainment of a certain target while making the process response insensitive to variations caused by changes in the noise factors. To make this precise, we will use the three objective functions described in (4.3). We will consider both the versions of these objective functions obtained through certainty equivalence as well as under parameter estimation uncertainty. We are interested in how close the optima of the predicted objective functions (predicted optima) are to the optima of the true objective functions.

The predicted objective functions may be non-convex and the chosen optima may not truly reflect the overall optimum for that specific objective function. Thus, there is the possibility for the comparison to depend more on the capabilities of the optimization tools at hand rather than on the intrinsic performance of the methods. For this reason, we also present a more direct comparison of the procedures using

the simulated integrated mean square error for the mean (SIMSEM) and for the variance (SIMSEV). We describe these in section 4.5.4. These criteria were also used by Pickle et al. (2008).

All computations and simulation results that are presented below were obtained using the R statistical software. Optimization was done with the 'Rsolnp' package.

4.5.2 Description of Simulation Procedure

In this section, we describe the procedure used for the simulation study. We generate data according to the model

$$\begin{aligned} y =& \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + a_{12} x_1 x_2 + a_{13} x_1 x_3 + a_{23} x_2 x_3 + a_{11} x_1^2 + a_{22} x_2^2 + a_{33} x_3^2 + \\ & \gamma_1 z_1 + \gamma_2 z_2 + \delta_{11} x_1 z_1 + \delta_{12} x_1 z_2 + \delta_{21} x_2 z_1 + \delta_{22} x_2 z_2 + \delta_{31} x_3 z_1 + \delta_{32} x_3 z_2 + \epsilon \end{aligned} \tag{4.29}$$

where $\epsilon \sim N(0, \sigma^2(x_1, x_2, x_3))$ and

$$\begin{aligned} \sigma^2(x_1, x_2, x_3) = \exp\ (& \tau_0 + \tau_1 x_1 + \tau_2 x_2 + \tau_3 x_3 + \\ & \tau_{12} x_1 x_2 + \tau_{13} x_2 x_3 + \tau_{23} x_2 x_3 + \tau_{11} x_1^2 + \tau_{22} x_2^2 + \tau_{33} x_3^2) \end{aligned}$$

We set $\boldsymbol{\beta} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, a_{12}, a_{13}, a_{23}, a_{11}, a_{22}, a_{33}, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}, \delta_{31}, \delta_{32})$ and $\boldsymbol{\tau} = (\tau_0, \tau_1, \tau_2, \tau_3, \tau_{12}, \tau_{13}, \tau_{23}, \tau_{11}, \tau_{22}, \tau_{33}) = (\tau_0, \tau_1, \tau_2, \tau_3, 0, 0, 0, 0, 0, 0)$.

There are 18 parameters in $\boldsymbol{\beta}$ and 10 in $\boldsymbol{\tau}$ making a total of 28 parameters in all. Each of the methods APLES and MLE has to estimate all of these based on 100 data points generated from model (4.29). The LS method estimates all 18 parameters in $\boldsymbol{\beta}$ and only τ_0 in $\boldsymbol{\tau}$. It sets all other parameters in $\boldsymbol{\tau}$ to 0, no matter their true value. MLO only estimates those parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ which are non-zero.

We choose the covariance matrix of the noise variables as:

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \begin{pmatrix} \sigma_{z_1}^2 & \sigma_{z_1 z_2} \\ \sigma_{z_1 z_2} & \sigma_{z_2}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Now, consider a fixed setting of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. These fixed values of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ determine a true mean response and a true variance response, μ_y and σ_y^2 respectively.

Using μ_y and σ_y^2, we can form the true quality objective functions: MSE, TTB, and MTB. The simulation is carried out as follows:

1. Generate a dataset according to model (4.29).
2. Use each of the four methods (MLO, MLE, LS and APLES) to obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$.
3. For each of the four methods, obtain the estimated mean and variance response surfaces $\hat{\mu}_y$ and $\hat{\sigma}_y^2$.
4. Next, we plug these into the quality objective function. There are six objective functions in all: MSE-CE, TTB-CE, MTB-CE, MSE-U, TTB-U and MTB-U.
5. We then determine the optimum that is obtained within the design region. We choose the design region of interest to be cuboidal: $-1 \leq x_i \leq 1$ i.e., we solve the optimization problem for x_i in this range.
6. We compute the L_1 distance between this predicted optimum and the optimum of the corresponding true quality objective function.
7. We compute the values of the true quality objective functions when evaluated at the predicted optimum.

We repeat steps 1-7 for 500 datasets generated from model (4.29). We report the median of the L_1 distances from step 6 for the 500 datasets. Using this measure, a method is considered better if it's predicted optimum is closer to the true optimum in the L_1 distance. We also report the median of the values obtained from step 7 for the 500 datasets. Using this measure, a method is considered better if the optimum it predicts leads to a smaller value of the true quality objective function than other methods. As a measure of scale to accompany these median measures, we report $\frac{1.486\times \text{ MAD}}{\sqrt{500}}$, a robust estimate of the standard error, where MAD is the median absolute deviation.

Next, we describe the experimental design used. We use three control factors x_1, x_2, x_3 and two noise factors z_1, z_2. We use a central composite design for the control factors. We choose the distance between the center of the design and the

axial points to be 2. We also choose the number of center points, $n_0 = 3$. We use four replicates so that the total sample size is $n = 25 \times 4 = 100$. The design matrix is shown in Table 4.12.

	1	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2	z_1	z_2	x_1z_1	x_1z_2	x_2z_1	x_2z_2	x_3z_1	x_3z_2
1	1	-1	-1	-1	1	1	1	1	1	1	-1	-1	1	1	1	1	1	1
2	1	1	-1	1	-1	1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1
3	1	-1	1	1	-1	-1	1	1	1	1	-1	-1	1	1	-1	-1	-1	-1
4	1	1	1	-1	1	-1	-1	1	1	1	-1	-1	-1	-1	-1	-1	1	1
5	1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1	1	-1	1	1	-1
6	1	1	-1	-1	-1	-1	1	1	1	1	1	-1	1	-1	-1	1	-1	1
7	1	-1	1	-1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	-1	1
8	1	1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	-1	1	-1
9	1	-1	-1	1	1	-1	-1	1	1	1	-1	1	1	-1	1	-1	-1	1
10	1	1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	1	-1
11	1	-1	1	-1	-1	1	-1	1	1	1	-1	1	1	-1	-1	1	1	-1
12	1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	-1	1	-1	1
13	1	-1	-1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
14	1	1	-1	1	-1	1	-1	1	1	1	1	1	1	1	-1	-1	1	1
15	1	-1	1	1	-1	-1	1	1	1	1	1	1	-1	-1	1	1	1	1
16	1	1	1	-1	1	-1	-1	1	1	1	1	1	1	1	1	1	-1	-1
17	1	-2	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
18	1	2	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
19	1	0	-2	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
20	1	0	2	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
21	1	0	0	-2	0	0	0	0	0	4	0	0	0	0	0	0	0	0
22	1	0	0	2	0	0	0	0	0	4	0	0	0	0	0	0	0	0
23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.12. Design Matrix for Simulation Experiment

We considered 15 settings corresponding to a combination of one of the following $\boldsymbol{\tau}$ scenarios:

1. No τ effect

   $\boldsymbol{\tau} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

2. Single small τ effect

   $\boldsymbol{\tau} = (0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0)$

3. Single medium τ effect

   $\boldsymbol{\tau} = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$

4. Single large $\boldsymbol{\tau}$ effect

   $\boldsymbol{\tau} = (0, 1.5, 0, 0, 0, 0, 0, 0, 0, 0)$

5. Several large $\boldsymbol{\tau}$ effects

   $\boldsymbol{\tau} = (0, 1, 1, 1, 0, 0, 0, 0, 0, 0)$

and one of the following $\boldsymbol{\beta}$ scenarios:

1. Few non-zero β_j

   $\boldsymbol{\beta} = (5, 2, 0, 0, 0, 3, 0, 0, 0, 0, 0, -1, 0, 0, 0, -3, 0, 0)$

2. Many non-zero β_j

   $\boldsymbol{\beta} = (5, 2, 2, 0, 0, 3, -3, 2, 2, 0, 0, -1, 0, 2, 0, -3, 2, 0)$

3. All β_j non-zero

   $\boldsymbol{\beta} = (5, 10, 10, 10, 5, 5, 5, 6, 6, 5, 2, 2, 3, 3, 2, 2, 2, 2)$

Table 4.13 shows the fifteen settings that were used in the simulation study.

Factor	Few non-zero β_j	Many non-zero β_j	All β_j non-zero
No τ effect	1	6	11
Single small τ effect	2	7	12
Single medium τ effect	3	8	13
Single large τ effect	4	9	14
Several large τ effects	5	10	15

Table 4.13. Settings Used in Simulation Study

4.5.3 Description of the Settings Used in Simulation Study

Of the three $\boldsymbol{\beta}$ settings that we use, the first corresponds to the situation where only a few elements (four out of eighteen) of $\boldsymbol{\beta}$ are actually non-zero. The second $\boldsymbol{\beta}$ setting also has some elements of $\boldsymbol{\beta}$ which are zero, but there are more non-zero elements (eleven out of eighteen). In these two cases, we expect APLES to have an advantage over MLE and LS in terms of identifying the truly non-zero one. This is because MLE and LS never set any coefficients to 0. This ability to potentially set the truly zero ones to zero can translate into gains in terms of reduced prediction error and also better estimates of the true quality objective functions. An interesting question is how much of an advantage this is. This question is one that is answered using this simulation study.

The third setting for $\boldsymbol{\beta}$ has several large coefficients. In this case, we expect MLE (and LS, for the homoscedastic case) to perform well. However, since the APLES estimates do not penalize truly large coefficients by much, this setting informs us of how competitive APLES can be to these well-established methods.

By equation (4.11), the true mean response surface only depends on the first 10 elements of $\boldsymbol{\beta}$. It does not depend on the coefficients of the noise or interaction terms. It also does not depend on the elements of τ. Thus, only three mean response surfaces are used in our simulations. These correspond to settings 1-5 (sparse effects), settings 6-10 (several medium effects) and settings 11-10 (several large effects). These are shown in figure 4.8.

The first setting for $\boldsymbol{\tau}$ corresponds to the constant residual variance situation and we expect LS to perform well. However, it is interesting to find out what happens when this $\boldsymbol{\tau}$ setting is paired with the various $\boldsymbol{\beta}$ scenarios. APLES and MLE have to estimate an additional 9 τ parameters relative to LS. However, since in the sparse $\boldsymbol{\beta}$ scenario APLES also has the advantage in estimation of the mean response surface, it is of interest to know what the trade-off will be.

The other settings of $\boldsymbol{\tau}$ correspond to increasing extremeness of the residual heteroscedasticity and we expect LS to perform progressively worse. Setting 15 of our simulation study is one which we believe can exhibit the relative performance of APLES and MLE. In this setting all elements of $\boldsymbol{\beta}$ are large and non-zero. In addition, there are several large non-zero elements of $\boldsymbol{\tau}$.

Unlike the true mean response surface, the true variance response surface depends on $\boldsymbol{\tau}$ as well as the last 8 elements of $\boldsymbol{\beta}$ by equation (4.12). Therefore, we obtain a different variance response for every pair of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ settings. Therefore, we have 15 different variance response surfaces for our simulation. These are shown in figures 4.9 - 4.11.

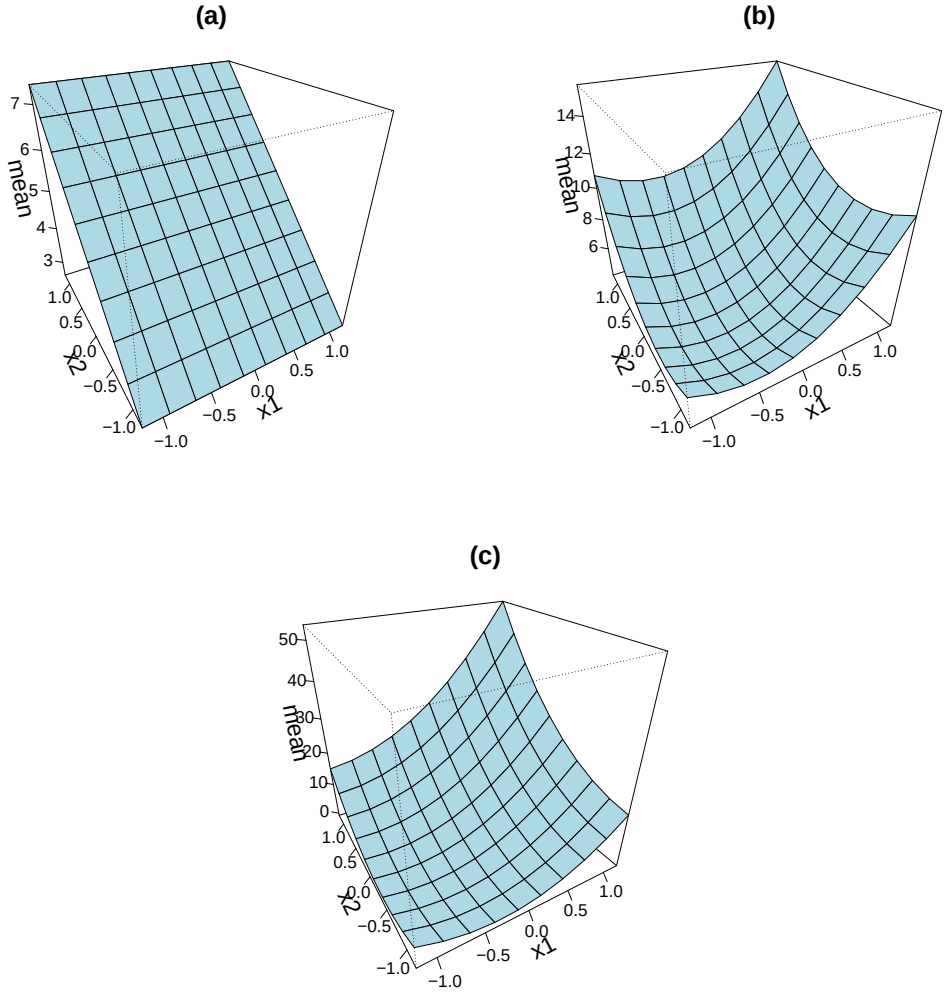

Figure 4.8. Plots of True Mean Response Surfaces Used in Simulation Study. (a) corresponds to settings 1-5, (b) to settings 6-10, and (c) to settings 11-15.

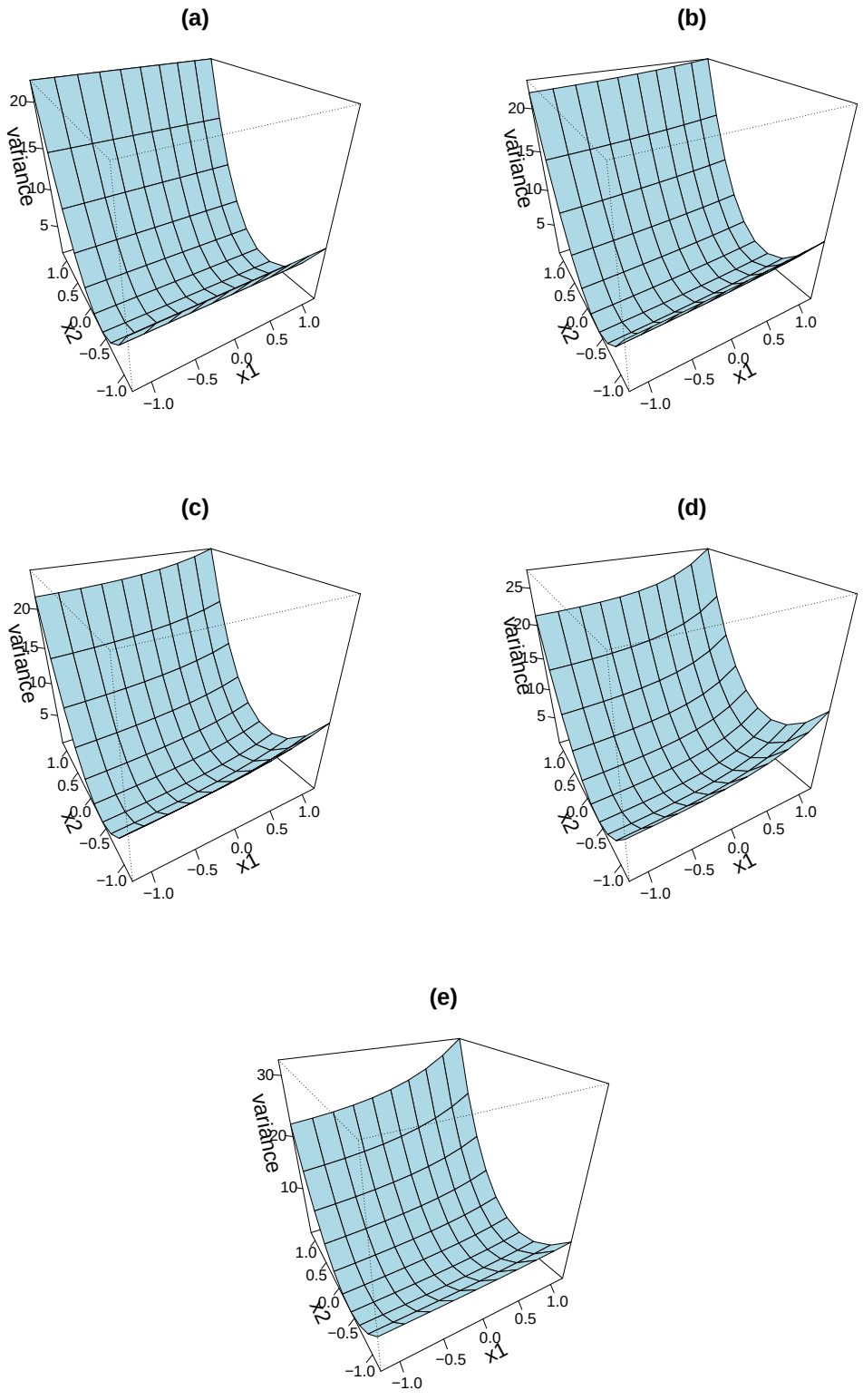

Figure 4.9. Plots of True Variance Response Surfaces Used in Simulation Study. (a) - (e) correspond to settings 1-5 respectively.

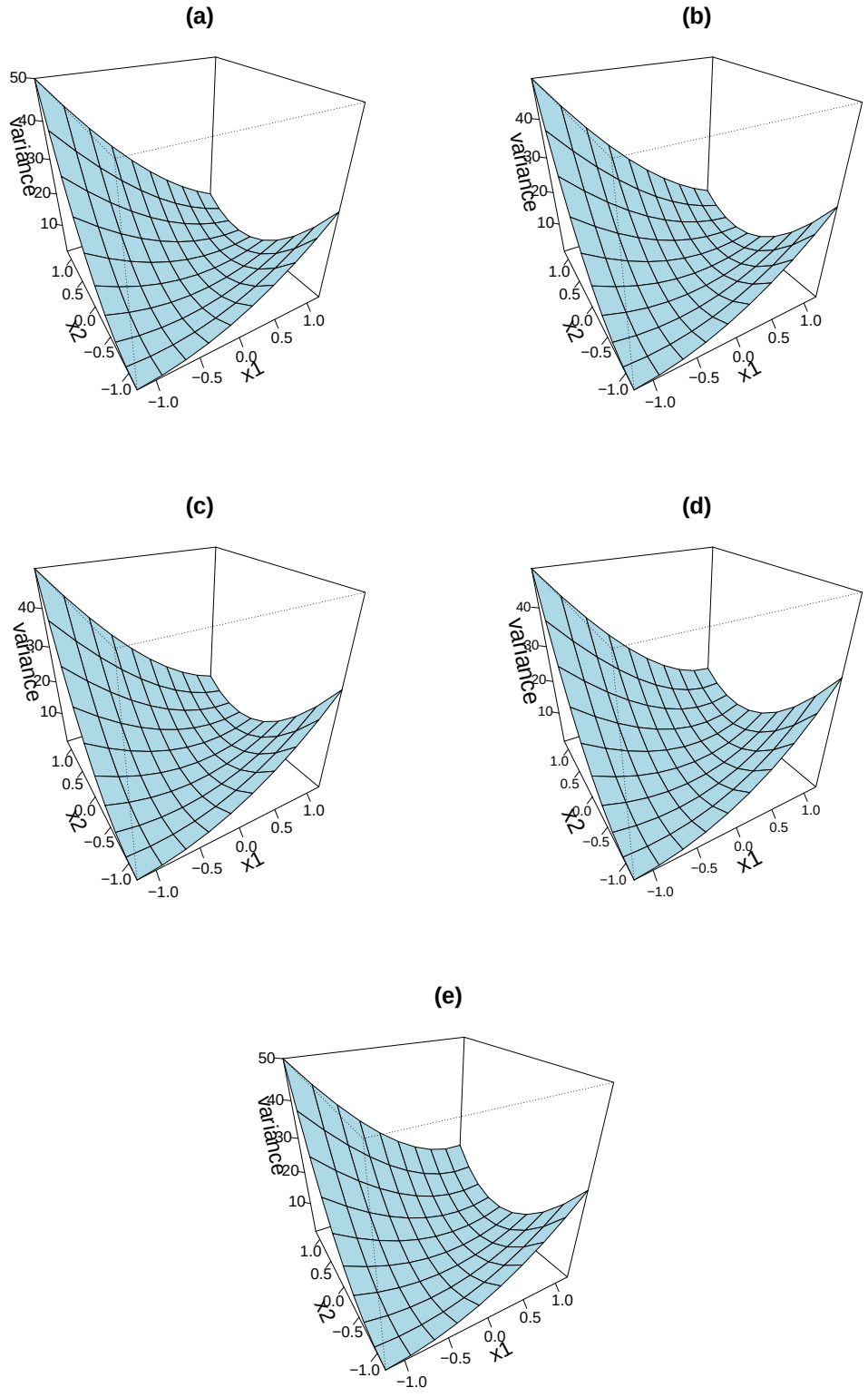

Figure 4.10. Plots of True Variance Response Surfaces Used in Simulation Study. (a) - (e) correspond to settings 6-10 respectively.

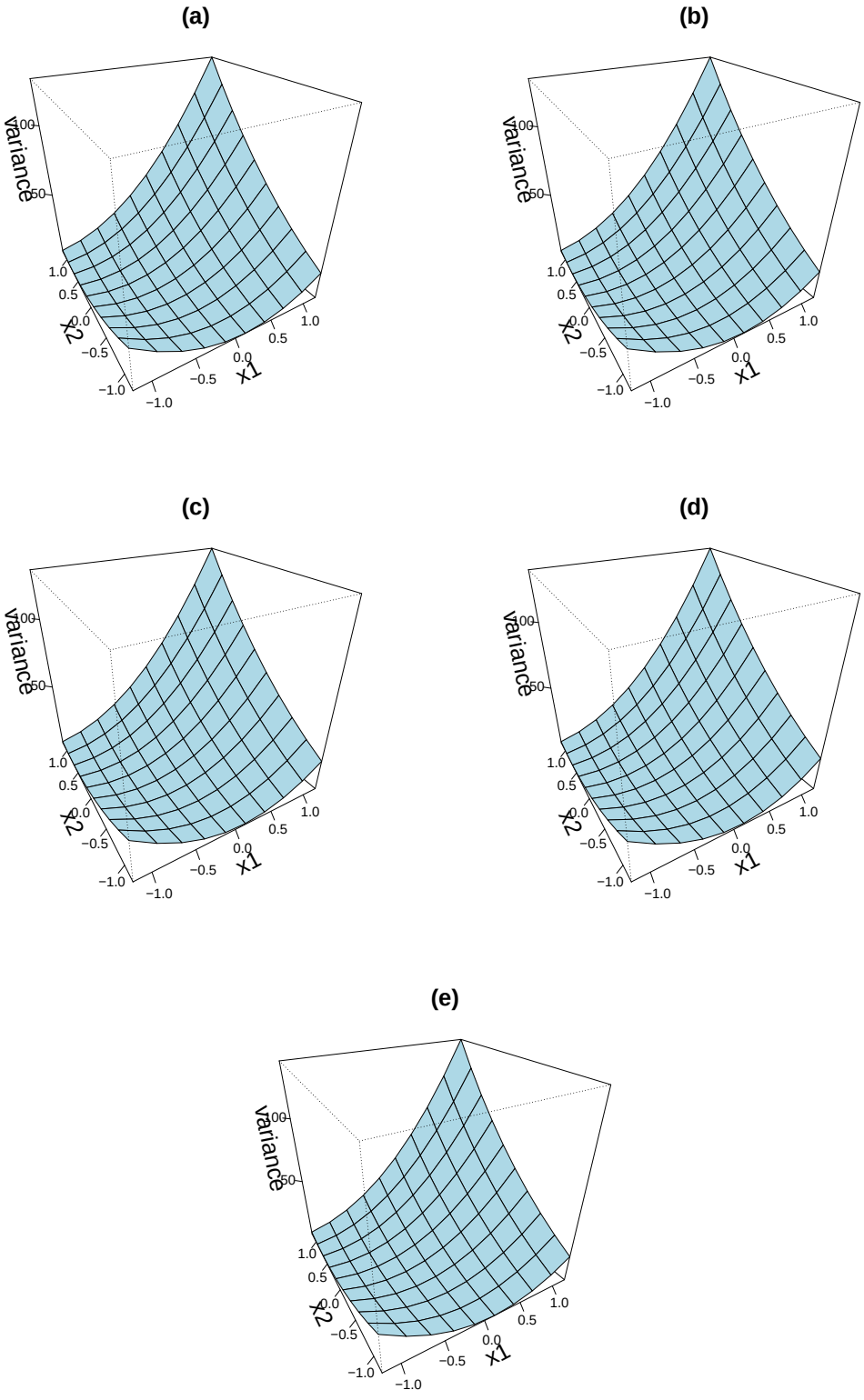

Figure 4.11. Plots of True Variance Response Surfaces Used in Simulation Study. (a) - (e) correspond to settings 11-15 respectively.

4.5.4 Discussion of Simulation Results: Integrated Mean Square Error

In this section, we report the predictive accuracy of the estimated mean and variance functions for the true mean and variance functions respectively. One way to do this is to measure the mean square error of the mean estimate, $\sum(\mu_y(\mathbf{x}) - \hat{\mu}_y(\mathbf{x}))^2/500$, and the mean square error of the variance estimate, $\sum(\sigma^2(\mathbf{x}) - \hat{\sigma}^2_y(\mathbf{x}))^2/500$, over the 500 datasets generated, where $\mathbf{x}$ refers to the original settings of the control factors from the experimental design used to generate the data. However, we want to know how well the estimated response surfaces perform not just at the original design points. We want to assess the performance of the estimated mean and variance responses over the entire design region. For this reason, we report the integrated mean square error.

Following Pickle et al. (2008), we use the simulated integrated mean square error. For each simulated dataset, the simulated integrated mean square error uses the mean square error averaged over a large number of points in $\mathbf{x}$ space.

For the mean response we define the simulated integrated mean square error of the mean (SIMSEM) as

$$SIMSEM = \frac{\sum asem}{500}$$

For each simulated dataset, *asem*, the average squared error for the mean, is defined as the average based on 125000 $\mathbf{x}$ locations (using a $50 \times 50 \times 50$ uniform grid of points in the (x_1, x_2, x_3) space). We have

$$asem = \frac{\sum(\mu_y - \hat{\mu}_y)^2}{125000}$$

Similarly, for the variance response we define the simulated integrated mean square error of the variance (SIMSEV) as

$$SIMSEV = \frac{\sum asev}{500}$$

For each simulated dataset, *asev*, the average squared error for the variance, is defined as the average based on 125000 $\mathbf{x}$ locations (using a $50 \times 50 \times 50$ uniform

grid of points in the (x_1, x_2, x_3) space). We have

$$asev = \frac{\sum(\sigma_y^2 - \hat{\sigma}_y^2)^2}{125000}$$

In tables A.5 - A.7 below, we present the SIMSEM and SIMSEV values for each of the four methods MLO, MLE, LS and APLES. We also illustrate this same information in figures 4.12 and 4.13.

Based on these results, we can make the following conclusions. APLES is a good competitor to the unpenalized MLEs for approximating the true mean and variance responses. For the mean response, the APLES estimator has a smaller SIMSEM than MLE for the three $\boldsymbol{\beta}$ settings that were considered. As is to be expected, the LS estimator performs the worst when we have medium or extreme heteroscedasticity. However, under constant residual variance and under the single small τ effect scenario the LS estimated mean response surface has smaller SIMSEM than APLES or MLE. We observe that since APLES has the ability to set some effects to 0, it performs better than MLE in the first two $\boldsymbol{\beta}$ settings. However, even in the third $\boldsymbol{\beta}$ setting where all of the βs are non-zero, we find in this simulation experiment that APLES still performs better than or equal to the MLE generated mean response surface.

We now describe the performance of the estimated variance response surfaces as measured using SIMSEV. For each of the three $\boldsymbol{\beta}$ settings, we observe that the methods have equal performance under the first $\boldsymbol{\tau}$ setting (i.e. under homoscedasticity). In the next three $\boldsymbol{\tau}$, we observe an increasing deterioration in the performance of the LS estimator. The MLE and APLES estimator have almost equal performance, with the APLES estimator performing slightly better under the first two $\boldsymbol{\beta}$ settings and the MLE performing better in the third. This pattern is repeated for the fifth $\boldsymbol{\tau}$ setting.

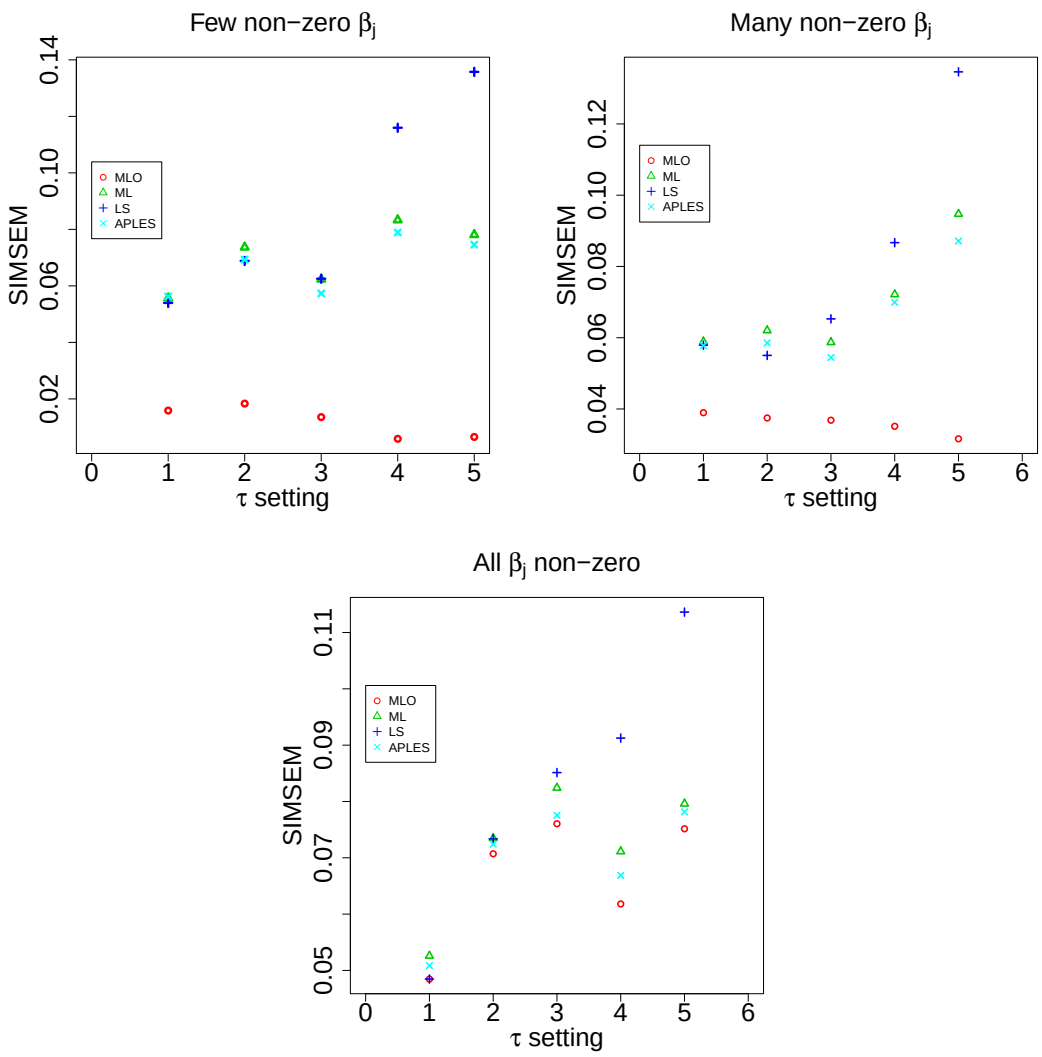

Figure 4.12. Comparison of simulated integrated mean square error of the mean (SIMSEM). Vertical axis shows values of the SIMSEM. The points 1 - 5 on the horizontal axis correspond to the five τ settings.

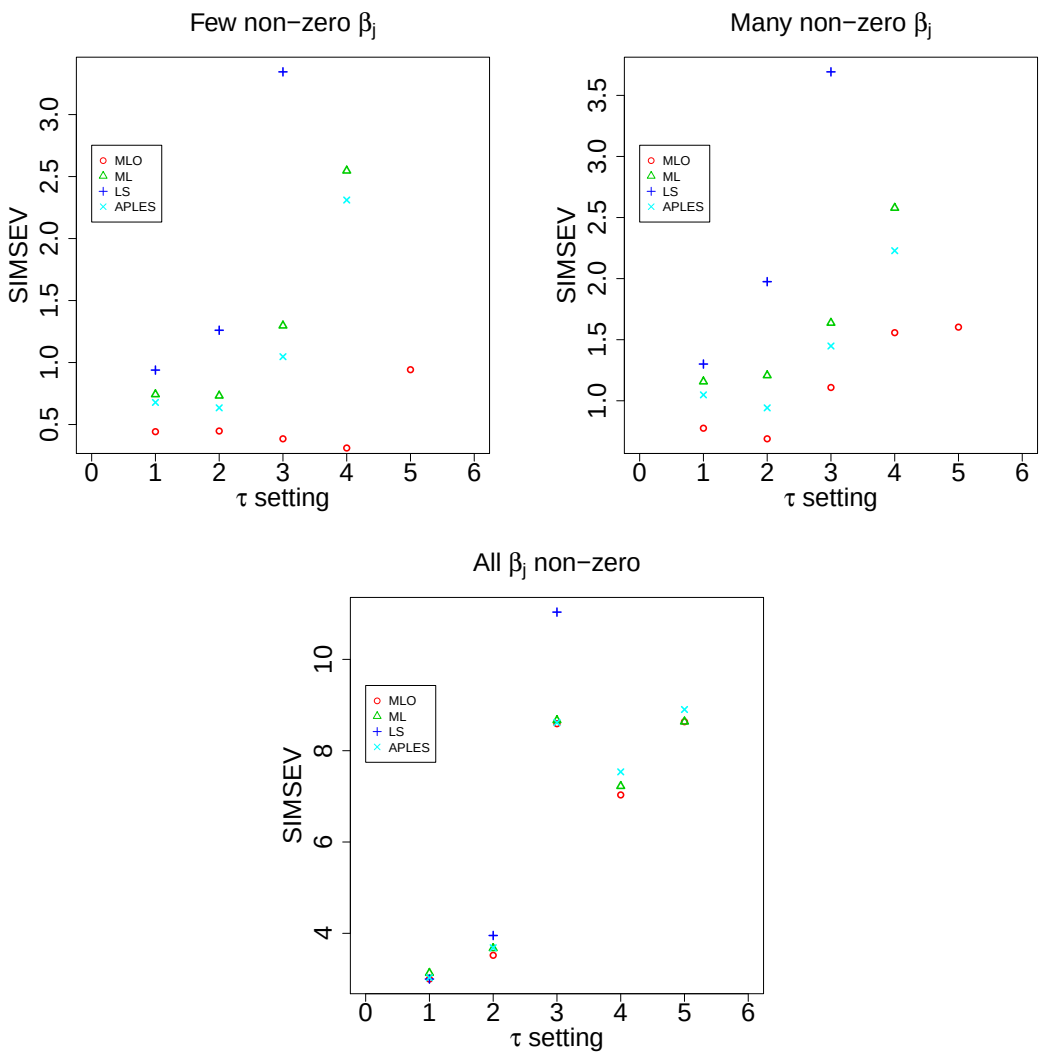

Figure 4.13. Comparison of simulated integrated mean square error of the variance (SIMSEV). Vertical axis shows values of the SIMSEV. The points 1 - 5 on the horizontal axis correspond to the five τ settings.

4.5.5 Discussion of Simulation Results: Predicted Optima

In this section we compare the performance of the estimation approaches (MLE, LS, APLES) in terms of the predicted optima that are obtained when the resulting quality objective function is optimized. We are primarily interested in the criteria MSE-U, TTB-U and MTB-U since in practice we do not have certainty equivalence. However, we also report the results obtained using MSE-CE, TTB-CE and MTB-

CE.

We present fifteen pairs of tables (A.8 - A.22) in appendix A.2. Each of these pairs corresponds to one of the fifteen settings. For a fixed setting (say setting 1), we report the minimum value of the true objective function as well as the predicted value of the true objective function when the optimizing values from each of the four estimates (MLO, MLE, LS, APLES) is plugged into it.

The first column of each table corresponds to the MSE criterion, the second to the TTB criterion and the third to the MTB design criterion. For each method we report the median of all the true quality objective function values when evaluated at the predicted optima for each of the 500 datasets generated. In parentheses, we report a robust estimate of the standard error. For each table, row 1 corresponds to the minimum of the true quality objective function in the region of interest. Row 2 corresponds to the MLO, row 3 to the MLE, row 4 to the LS and row 5 to APLES.

Next, we present fifteen pairs of tables (A.23 - A.37) in appendix A.2. Each of these pairs corresponds to one of the fifteen settings. For a fixed setting (say setting 1), we report the mean distance (L_2) between the true minimum and the predicted minimum (MLE, LS, APLES) within the region $-1 \leq x_i \leq 1$.

The first column of each table corresponds to the MSE criterion, the second to the TTB criterion and the third to the MTB criterion. For each method, we report the mean distance (L_2) between the true minimum and the predicted minimum over the 500 datasets generated. In parentheses, we report the standard error. For each table, row 1 corresponds to the distance for the MLO, row 2 to the MLE, row 3 to LS and row 4 to APLES.

We summarize our findings in the following subsections, each of which answers the following questions respectively.

1. Focusing specifically on the MSE-U and MTB-U objective criteria, how do the three methods compare?

2. Irrespective of the estimation method, how do the MSE optimizing values compare to the TTB optimizing values in terms of distance from the true optimizing values?

4.5.5.1 Performance of methods under MSE-U and MTB-U

We first summarize our results when the objective function is the MSE-U criterion. For **settings 1 - 5** we have a sparse $\boldsymbol{\beta}$, and the values of the τ_j are varied. For setting 1, none of the factors x_1, x_2 or x_3 has a dispersion effect. Under this setting, the methods that we are comparing have very similar values for the true objective function.

Under settings 2, 3 and 4, the MLO clearly performs better than the others and the LS performs much worse than the others. The MLE and APLES perform similarly, with the APLES values being slightly better. Under setting 5, the MLO still performs the best and the LS performs the worst. The MLE and APLES again perform similarly.

For **settings 6 - 10** we have more of the β_j being non-zero than in settings 1 - 5. For setting 6, none of the factors x_1, x_2 or x_3 has a dispersion effect. Under this setting, there is very little to distinguish the methods that we are comparing. This situation also occurred for setting 1.

For settings 7, 8, 9 and 10, we have the LS estimator performing the worst. The MLO still has an advantage in this situation like it did for settings 2, 3 4 and 5. We note that APLES performs better than MLE like in situations 2, 3 and 4. However, it performs much better relative to MLE in settings 7, 8 and 9 than it did in 2, 3 and 4. This is an interesting finding because settings 2, 3 and 4 involve a sparser $\boldsymbol{\beta}$ structure than settings 7, 8 and 9.

For **settings 11 - 15** we have several of the β_j being large and non-zero. All of the methods perform fairly well under all of these settings.

Next, we summarize our results when the objective function is the MTB-U criterion. For **settings 1 - 5** , we observe a similar pattern as for the MSE-U criterion. There is no difference between the methods for setting 1. The LS begins to perform worse as we introduce heteroscedasticity in settings 2 - 5. APLES performs better than MLE in setting 2 and the methods perform similarly in settings 3 - 5.

For **settings 6 - 10** we find that LS still performs worse, but only slightly so and for **settings 11 - 15** all the methods perform the same and actually all hit the minimum of the true objective function.

Our conclusion based on these results is that while various methods may perform differently under different settings of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, their performance is also highly dependent on the objective function being used.

4.5.5.2 Comparison of MSE and TTB criteria

For this comparison, we use tables (A.23 - A.37). We observe that the MSE optimizing solutions are almost always closer to the true optimizing solutions than the TTB optimizing solutions. This is irrespective of the estimation method used.

Our finding is in agreement with the observation made in Lin and Tu (1995) that the TTB optimizing solution may lead to solutions which have larger MSE than one based directly on an MSE criterion. This is because in seeking solutions that bring the mean exactly on target, the TTB criterion may end up with solutions that have high variance thus a larger MSE.

4.6 Discussion

Our contributions in this chapter are threefold. Firstly, we proposed to use a penalized likelihood procedure for obtaining the predicted response surface for use in the heteroscedastic RPD problem. Even though the heteroscedastic RPD problem has been formulated for some time now (Engel and Huele, 1996a), most recent authors have focused on the constant residual variance approach. In this chapter, we have revisited the heteroscedastic RPD problem and applied a relatively modern variable selection procedure to solve it. Our procedure is fast, efficient and has the potential to be used even beyond traditional RPD situations which usually involve a few control and noise factors.

Secondly, we showed through examples and a simulation study that the predicted optima obtained using APLES estimates may result in values of the quality objective function that are more optimal than using traditional methods like MLE and LS. However, the exact performance depends on the actual settings of the parameters that pertain in the process, such as sparsity structure, size and nature of dispersion effects, etc. Our simulation study also encompassed three of the most common quality objective functions that practitioners actually use.

When optimizing estimated responses, one concern is to be able to accurately capture the estimation uncertainty in the objective function being optimized. This

way one can guard against estimated or predicted optima that are too far away from the true ones. Thirdly, we proposed a quality objective function to be used for the heteroscedastic RPD problem when we want to account for parameter estimation uncertainty. This proposed quality objective function is an extension of the quality objective function proposed in Miro-Quesada and Del Castillo (2004) for the constant residual variance situation.

Chapter 5

Analysis of Unreplicated Fractional Factorials using APLES

5.1 Introduction

The importance of identifying a smaller subset of a potentially large number of factors to include as explanatory variables for a response permeates much of statistics. This quest for parsimony accomplishes two goals. First, we obtain models that are interpretable, and second there is less variability in the estimated response. In analyzing designed experiments in the industrial setting, a more immediate goal is practicality. When we screen for important factors in a new process, it may be too expensive or time-consuming to vary the response over all combinations of factors. Two-level unreplicated 2^{k-p} designs are experimental designs which lend themselves well to such screening experiments. Analyzing the resulting data to determine which factors or interactions are active is a challenging task for which a considerable literature exists.

Taguchi (1980) emphasized the need for designing quality into products and processes by identifying both location and dispersion effects. However, the identification of location effects in designed experiments is well known, e.g. Daniel (1959). Box and Meyer (1986a) proposed a Bayesian approach based on effect sparsity. The Pseudo-Standard Error (PSE) approach of Lenth (1989) is still very popular and widely used. More recently, Aguirre-Torres and de la Vara (2011) proposed methods that account for the presence of outliers. Box and Meyer (1986b)

pioneered the analysis of unreplicated 2^{k-p} designs for dispersion effects by showing that the assumption of effect sparsity can be extended to dispersion effects. Other proposals can be found in Bergman and Hynen (1997), Wang (1989), Harvey (1976) and Liao (2000), all of which, like Box and Meyer's approach are based on the squared residuals after an initial location fit. Harvey's method and Wang's method are explicitly based on a log-linear dispersion model. Bursztyn and Steinberg (2006) and Brenneman and Nair (2001) provide reviews on these and other methods proposed for dispersion effects.

All of the dispersion effect identification methods mentioned above suffer from an inherent flaw, in that they perform poorly when there are multiple dispersion effects (McGrath and Lin, 2001b). Even though we are working under the assumption of effect sparsity, it is still dangerous to proceed with these methods unless one has strong a priori information that there is only one dispersion effect. McGrath and Lin (2001b) propose an approach which addresses multiple dispersion effects. However, as noted by Bursztyn and Steinberg (2006), their method also requires pre-identification of the specific triplet that one wants to search among for dispersion effects . If one knows this, then their method is able to identify true dispersion effects without spurious identification. Otherwise, the computational cost to search through the entire set of possible triplets in, for example, a design with fifteen factors can be prohibitive.

Very few methods have been proposed for analyzing both location *and* dispersion effects. The possibility of confounding has to be accounted for as demonstrated by Pan (1999) and McGrath and Lin (2001a). Two such proposals have been suggested by McGrath and Lin (2003) and by Pan and Taam (2002).

Penalized likelihood methods were developed especially for the high dimensional variable setting. They have also been used successfully to analyze designed experiments such as supersaturated designs. For example, Phoa (2009) applies the Dantzig selector and Li and Lin (2009) use the SCAD penalty. In both of these applications, the penalization is applied to location effects in order to achieve parsimony. However, penalized likelihood methods can also be applied to the variance submodel, the main contribution of this chapter. Thus, by jointly penalizing both the mean and the variance submodels, we simultaneously identify and estimate both the location and the dispersion effects.

The rest of this chapter is organized as follows. For the remaining part of this introductory section we formalize the notation and the model. Next, we introduce a popular example to show how our method is applied. In section 5.3 we introduce our proposal and present the algorithm for implementing it. We present the example in section 5.4 and two more examples in sections 5.5 and 5.6 followed by the results of a simulation study in section 5.7. We conclude with some practical considerations and recommendations.

5.1.1 Notation and Model

Let $\mathbf{X} = (\mathbf{X}_0, ..., \mathbf{X}_{n-1})$ be the $n \times n$ matrix of regressors corresponding to an $n = 2^{k-p}$ fractional factorial design. The first column $\mathbf{X}_0 = (1, \ldots, 1)'$ of $\mathbf{X}$ is a column of ones and the remaining columns consist of $+1$ and -1 corresponding to the high and low levels of the k factors and the interactions between them.

Let the $n \times 1$ vector $\mathbf{y} = (y_1, \ldots, y_n)'$ consist of observations of a performance characteristic corresponding to the n runs of the experiment. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ be an $n \times 1$ vector of unobserved errors which are independent, have zero means and do not generally have the same variance. Thus $\text{Var}(\epsilon_i) = \sigma_i^2$. Generally, these will be a function of the design matrix $\mathbf{X}$.

The joint location and dispersion model that we assume for the observations is

$$y_i = \mu_i + \epsilon_i \tag{5.1}$$

where $\epsilon_i \sim N(0, \sigma_i^2)$. The mean submodel is

$$\mu_i = \mathbf{x}_i' \boldsymbol{\beta} \tag{5.2}$$

and so the mean of the response lies in the linear space spanned by the columns of $\mathbf{X}$.

In effect, we have the usual heteroscedastic linear model, which reduces to the homoscedastic case when the σ_i are all equal. We will consider the following log-linear or multiplicative variance submodel: $\sigma_i^2 = e^{\mathbf{x}_i' \boldsymbol{\tau}}$. McGrath and Lin (2001b)

employ the alternative parameterization:

$$\sigma_i^2 = \sigma^2 \prod_{j=1}^{n-1} \Delta_j^{x_{ij}} \tag{5.3}$$

The two formulations are equivalent with $\sigma^2 = e^{x_{i0}\tau_0}$ and $\Delta_j = e^{\tau_j}$.

This is because

$$\begin{aligned}
\sigma_i^2 &= e^{\mathbf{x}_i'\boldsymbol{\tau}} \\
&= e^{x_{i0}\tau_0} e^{x_{i1}\tau_1} \cdots e^{x_{i,n-1}\tau_{n-1}} \\
&= \sigma^2 \Delta_1^{x_{i1}} \cdots \Delta_{n-1}^{x_{i,n-1}}
\end{aligned}$$

The elements $(\beta_0, \beta_1, \ldots, \beta_{n-1})$ of $\boldsymbol{\beta}$ are mean or location parameters and the elements $(\tau_0, \tau_1, \ldots, \tau_{n-1})$ of $\boldsymbol{\tau}$ are variance or dispersion parameters. The jth regressor, corresponding to column j of $\mathbf{X}$, has an active location effect if $\beta_j \neq 0$. Similarly, the jth regressor has an active dispersion effect if $\tau_j \neq 0$.

Let us denote the matrix of column vectors of true location effects by $\mathbf{X}_{\mathcal{L}}$ and the corresponding mean parameters vector by $\boldsymbol{\beta}_{\mathcal{L}}$. We always include the first column $\mathbf{X_0}$ in the active set. $\boldsymbol{\beta}_{\mathcal{L}}$ is unknown and interest lies in identifying its members. After identification, the model is given by $\mathbf{y} = \mathbf{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}} + \boldsymbol{\epsilon}$ and interest lies in obtaining estimates $\hat{\boldsymbol{\beta}}_{\mathcal{L}}$ of $\boldsymbol{\beta}_{\mathcal{L}}$.

Methods for identification of location and dispersion effects rely heavily on the assumption of effect sparsity. Prior to identification of location effects, it is assumed that only a small subset of the elements of $\boldsymbol{\beta}$ will be non-zero. Similarly, before identification of dispersion effects, it is assumed that only a small subset of the elements of $\boldsymbol{\tau}$ be non-zero. Thus we think of the true, but unknown $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ as being partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{L}}, \boldsymbol{\beta}_{\mathcal{L}^c})$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_{\mathcal{D}}, \boldsymbol{\tau}_{\mathcal{D}^c})$ where $\boldsymbol{\beta}_{\mathcal{L}}$ and $\boldsymbol{\tau}_{\mathcal{D}}$ are the non-zero or active location and dispersion parameters respectively and $\boldsymbol{\beta}_{\mathcal{L}^c} = \mathbf{0}$ and $\boldsymbol{\tau}_{\mathcal{D}^c} = \mathbf{0}$.

5.2 Example 1: Taguchi's Welding Data

A well-known example of a dataset with both location and dispersion effects is Taguchi's welding dataset Taguchi (1980), shown in Table 5.1. It has been analyzed

by several authors including Box and Meyer (1986b), Wang (1989), McGrath and Lin (2001a). The response is tensile strength and there are nine factors, labeled A through I, studied in a 2^{9-5} experiment with 16 runs. Four interactions (AC, AG, AH, and GH) as well are being considered, so that there are 13 orthogonal contrasts of interest. The two remaining contrasts corresponding to the columns labeled $e1$ and $e2$ measure only experimental error.

	D	H	-e1	G	-F	GH	-AC	A	-E	AH	e2	AG	I	B	-C	y
1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	43.70
2	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1	40.20
3	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	42.40
4	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	44.70
5	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	42.40
6	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	45.90
7	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	42.20
8	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	40.60
9	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	42.40
10	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	45.50
11	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	43.60
12	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	40.60
13	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	44.00
14	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	40.20
15	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	42.50
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	46.50

Table 5.1. Data for Welding Experiment with Factors Arranged in Same Order as Box and Meyer (1986b)

Table 5.2 shows the factors that were identified as active by various authors.

	Authors	Location Model	Dispersion Model
1	Box and Meyer (1986b)	B, C	C
2	Lenth (1989)	B, C	C, H, I
3	Wang (1989)	B, C	C, H, I
4	Bergman and Hynen (1997)	B, C	C, H, I
5	Nelder et al. (1998)	B, C, I	C, I
6	Liao (2000)	B, C	C, H, I
7	Brenneman and Nair (2001)	B, C	C
8	McGrath and Lin (2001a)	B, C	C

Table 5.2. Model Parameters Found to be Non-zero for Taguchi's Welding Data by Various Authors

All authors who have analyzed this data agree that only factors B and C have

active location effects (with the exception of Nelder et al. (1998) who include factor I). This is justified by the normal probability plot as well as by a use of Lenth's procedure. There is less agreement concerning the dispersion effects as seen in Table 5.2. Brenneman and Nair (2001) make a strong case for factor C (and possibly factor I) having a dispersion effect, by considering the projection of the original 16-run unreplicated design on the generators of factors B, C and I. Note that H = (-C)I.

The proposed method described in Chapter 3 is developed in section 5.3 for application to the analysis of unreplicated fractional factorials such as Taguchi's welding dataset.

5.3 The Proposed Method

In order to identify location and dispersion effects as well as estimate the identified location and dispersion effects in one step, we draw on ideas from high dimensional heteroscedastic regression. Let l be the log likelihood of the model (5.1). Daye et al. (2012) and Kolar and Sharpnack (2012) propose solutions to the problem of simultaneous variable selection and estimation of mean and variance parameters in high dimensional settings by maximizing l subject to bounds on the "sizes" of the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. Daye et al. (2012) use the LASSO penalty term, Kolar and Sharpnack (2012) use the SCAD penalty term. Adaptive Penalized Likelihood Effects Selection (APLES) developed in Chapter 3 uses the adaptive LASSO penalty. In this chapter, we apply the method of Chapter 3 to the problem formalized in section 5.1. We implement APLES with a particularly simple algorithm which takes advantage of the orthogonal design matrix and small sample sizes utilized in unreplicated fractional factorials.

The log-likelihood for the model (5.1) is given by:

$$l(\boldsymbol{\beta}, \boldsymbol{\tau}) = -\frac{n}{2} log 2\pi - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau} - \frac{1}{2}\sum_{i=1}^{n} e^{-\mathbf{x}_i'\boldsymbol{\tau}}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \tag{5.4}$$

The estimator proposed in Chapter 3 is the APLES estimator denoted by

$$\hat{\theta}_A = \begin{pmatrix} \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\tau}}_A \end{pmatrix}_{2n\times 1}$$

and defined as

$$\underset{\boldsymbol{\beta},\boldsymbol{\tau}}{\operatorname{argmax}} \left(-\frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\tau} - \frac{1}{2}\sum_{i=1}^{n} e^{-\mathbf{x}_i'\tau}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=0}^{n-1} w_{1j}|\beta_j| + \lambda_2 \sum_{j=0}^{n-1} w_{2j}|\tau_j| \right) \tag{5.5}$$

where λ_1 and λ_2 are tuning parameters for the mean and variance, respectively.

The maximization problem has the same properties as the equivalent one studied in section 3.4. It is not convex in $(\boldsymbol{\beta}, \boldsymbol{\tau})$, but it is biconvex. For fixed $\boldsymbol{\beta}$ it is convex in $\boldsymbol{\tau}$ and for fixed $\boldsymbol{\tau}$, it is convex in $\boldsymbol{\beta}$. For fixed $\boldsymbol{\tau}$ and fixed λ_2, as we increase λ_1 an increasing number of the elements of $\boldsymbol{\beta}$ are set to 0, until reaching λ_1^{max} at which they are all 0. For fixed $\boldsymbol{\beta}$ and fixed λ_1, as we increase λ_2 an increasing number of the elements of $\boldsymbol{\tau}$ are set to 0, until reaching λ_2^{max} at which they are all 0 (excepting the intercept terms which are not penalized). In implementing APLES for this chapter, we found λ_1^{max} and λ_2^{max} by setting λ_1 and λ_2 to two large values and checking to see if the resulting APLES estimators of all non-intercept coefficients are 0. If not, we double the initial λ_1 and λ_2 values and obtain new APLES estimators. By repeating this process a finite number of times, we arrive at two values that can be used as λ_1^{max} and λ_2^{max}.

The weight vectors $\mathbf{w}_1 = (w_{11}, \ldots, w_{1(n-1)})$ and $\mathbf{w}_2 = (w_{21}, \ldots, w_{2(n-1)})$ are chosen to satisfy $w_{1j} = 1/|\tilde{\beta}_j|$ and $w_{2j} = 1/|\tilde{\tau}_j|$ for any consistent estimators $\tilde{\beta}_j$ and $\tilde{\tau}_j$. For all the computations in this chapter, we use the ordinary least squares estimates for $\tilde{\beta}_j$. For $\tilde{\tau}_j$, we use the regression coefficients from a regression of the log of the squared residuals on $\mathbf{X}$. This provides us with consistent estimates for each element of $\boldsymbol{\tau}$ except for τ_0. See Harvey (1976) for derivation of this as well as further details on the log-linear heteroscedastic linear model. Since we do not penalize the intercept (either for the mean model or the variance model), this will not create a problem with the procedure. The adaptive LASSO penalty terms $\left(\lambda_1 \sum_{j=1}^{n-1} w_{1j}|\beta_j| \text{ and } \lambda_2 \sum_{j=1}^{n-1} w_{2j}|\tau_j|\right)$ ensure that some of the elements of $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ will be set to 0, depending on the choice of λ_1 and λ_2. The amount of penalization is data driven to ensure asymptotically unbiased estimates of the large parameter values (See Zou (2006) for more details). Even though desirable properties of the adaptive LASSO such as asymptotic unbiasedness of estimates and consistency in model selection are of little comfort in the small sample settings

we are dealing with here, the results of our simulation study in section 5.7 show that the proposed procedure used in conjunction with an appropriate information criterion performs very well.

We take advantage of the orthogonality of the design matrices encountered in $n = 2^{k-p}$ fractional factorial designs to propose the following easily programmed algorithm.

5.3.1 Algorithm

The following algorithm is an adaptation of the algorithm in Chapter 3 which takes advantage of the orthogonality of design matrix $\boldsymbol{X}$. In the original APLES algorithm (section 3.5), for a given response vector $\mathbf{y} = (y_1, \ldots, y_n)$ or $\mathbf{z} = (z_1, \ldots, z_n)$ (z_i is defined below), we cycle through the coordinates $\beta_0, \ldots, \beta_{n-1}$ and $\tau_0, \ldots, \tau_{n-1}$ and each time perform a regression of the partial residual on $\mathbf{X}$. However, in the algorithm below, we obtain the full least squares estimates. Thus there is no need to cycle through all the coordinates. The details are the same as in Chapter 3.

Consider that we have a pair (λ_1, λ_2) for which we seek estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{n-1})$ and $\hat{\boldsymbol{\tau}} = (\hat{\tau}_0, \ldots, \hat{\tau}_{n-1})$. We first obtain ordinary least squares (OLS) estimates $\hat{\boldsymbol{\beta}}$ of the location vector parameter. Next, we apply a thresholding rule to shrink the coefficients. A good thresholding rule shrinks the trivial effects so severely that they are set to 0 and does not affect the non-trivial effects at all. Next, the current shrunken estimate of $\hat{\boldsymbol{\beta}}$ obtained after applying the thresholding rule is used to compute the vector of squared residuals.

If these squared residuals were obtained from the true location model rather than an estimated one, then they could be treated as belonging to a known distribution belonging to an exponential family (specifically, a Gamma distribution). Assuming this to hold, we can treat the squared residuals as responses in a Gamma error generalized linear model (GLM) to obtain maximum likelihood estimates $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$. Next, we shrink these according to a thresholding rule to obtain new estimates of $\hat{\boldsymbol{\tau}}$. Our algorithm (described in section 5.3) consists of iterating between the location and dispersion models until convergence.

We describe the algorithm below:

Initialization

Fix λ_1 and λ_2. Set $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}, \hat{\boldsymbol{\tau}}^{(0)} = \mathbf{0}$. Choose ϵ_1^* and ϵ_2^* to use in convergence criterion. Define convergence criterion as $\|\hat{\boldsymbol{\beta}}_{(t+1)} - \hat{\boldsymbol{\beta}}_{(t)}\| < \epsilon_1^*$ and $||\hat{\boldsymbol{\tau}}_{(t+1)} - \hat{\boldsymbol{\tau}}_{(t)}|| < \epsilon_2^*$

Implementation: Let $t = 0, 1, 2, \ldots$ until desired convergence criterion is met:

1. Estimate $\hat{\boldsymbol{\beta}}_{temp}$ by weighted least squares regression of $\mathbf{y}$ on $\mathbf{X}$. The weights are given by $\left(e^{-\mathbf{x}_1'\hat{\boldsymbol{\tau}}^{(t)}}, \ldots, e^{-\mathbf{x}_n'\hat{\boldsymbol{\tau}}^{(t)}}\right)$

2. Shrink the elements of $\hat{\boldsymbol{\beta}}_{temp}$ as follows to obtain:

$$\hat{\beta}_j^{(t+1)} = \begin{cases} \hat{\beta}_{temp,j} - \lambda_1 w_{1j} & \text{if } \hat{\beta}_{temp,j} > 0 \text{ and } \lambda_1 w_{1j} < |\hat{\beta}_{temp,j}| \\ \hat{\beta}_{temp,j} + \lambda_1 w_{1j} & \text{if } \hat{\beta}_{temp,j} < 0 \text{ and } \lambda_1 w_{1j} < |\hat{\beta}_{temp,j}| \\ 0 & \text{if } \lambda_1 w_{1j} \geq |\hat{\beta}_{temp,j}| \end{cases}$$

3. Set $d_i = (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(t+1)})^2$ and form the new response (or adjusted dependent variable) for the dispersion submodel $z_i = \mathbf{x}_i'\hat{\boldsymbol{\tau}}^{(t)} + d_i e^{-\mathbf{x}_i'\hat{\boldsymbol{\tau}}^{(t)}} - 1$

4. Estimate $\hat{\boldsymbol{\tau}}_{temp}$ by ordinary least squares of $\mathbf{z}$ on $\mathbf{X}$. That is, $\hat{\boldsymbol{\tau}}_{temp} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$.

5. Shrink the elements of $\hat{\boldsymbol{\tau}}^{(t+1)}$ as follows to obtain:

$$\hat{\tau}_j^{(t+1)} = \begin{cases} \hat{\tau}_{temp,j} - \lambda_2 w_{2j} & \text{if } \hat{\tau}_{temp,j} > 0 \text{ and } \lambda_2 w_{2j} < |\hat{\tau}_{temp,j}| \\ \hat{\tau}_{temp,j} + \lambda_2 w_{2j} & \text{if } \hat{\tau}_{temp,j} < 0 \text{ and } \lambda_2 w_{2j} < |\hat{\tau}_{temp,j}| \\ 0 & \text{if } \lambda_2 w_{2j} \geq |\hat{\tau}_{temp,j}| \end{cases}$$

The above algorithm converges quickly using the convergence criterion defined above.

With the exception of steps 2 and 5, the above algorithm is equivalent to a Fisher scoring procedure for maximum likelihood estimates for $\boldsymbol{\beta}, \boldsymbol{\tau}$ as is well described in Harvey (1976). Steps 2 and 5 are necessary to achieve the adaptive penalization of location and dispersion coefficient estimates. In addition these steps ensure that we obtain an entire continuum of models corresponding to all possible choices of λ_1, λ_2 on the grid $(0, \lambda_1^{max}) \times (0, \lambda_2^{max})$

For Example 1 described in section 5.2, we used $\lambda_1^{max} = 2.5$ and $\lambda_2^{max} = 1.5$. The plots in Figures 5.1-5.6 were obtained by solving for all solutions for $\boldsymbol{\beta}, \boldsymbol{\tau}$ on the grid of values: $\lambda_1 \times \lambda_2 = (0, 0.1, 0.2, \ldots, 2.4, 2.5) \times (0, 0.1, 0.2, \ldots, 1.4, 1.5)$. Suppose we have obtained solutions for $(\lambda_1 = 1.0, \lambda_2 = 0.5)$. When we move onto a new pair which is close (for example, $(\lambda_1 = 1.1, \lambda_2 = 0.5)$), we do not need to begin the algorithm from $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}, \hat{\boldsymbol{\tau}}^{(0)} = \mathbf{0}$. We can use the current estimates as starting values because we expect the solution to be close.

5.3.2 Choice of Tuning Parameters

The proposed approach selects a different model for each pair of the tuning parameters and among all these we select the best model by using an information criterion. The AIC and BIC are known to overfit in small samples. For this reason, we use the corrected Akaike Information Criterion (AICc) (Hurvich and Tsai, 1989) which is a bias corrected version of the AIC for small samples, appropriate because we are working with fixed, small sample sizes. It is defined as:

$$AICc = -2l + 2r + 2(r)(r+1)/(n-r-1) \tag{5.6}$$

where r is the total number of non-zero coefficients (counting both location and dispersion parameters). Phoa (2009) also propose a modified AIC (mAIC) for automatic variable selection in supersaturated designs. It is defined as:

$$mAIC = -2l + 2r^2 \tag{5.7}$$

We find the tuning parameters λ_1, λ_2 which minimizes the AICc or mAIC over a grid of candidate values and choose the model with the smallest AICc or mAIC. The model obtained best explains the data with minimal complexity. We found the mAIC penalty was usually more severe than the AICc because it usually selected smaller models than the AICc.

5.4 Example 1: Analysis of Taguchi's Welding dataset Using APLES

In this section we present the results of an analysis of the Taguchi welding dataset. APLES found active location effects in factors B and C when using either mAIC or AICc. For the dispersion effects, it found an active dispersion effect for factor C when using mAIC and active dispersion effects for factors C and I when using AICc. Tables 5.3 and 5.4 report the coefficient estimates and the standard errors.

Term	APLES with mAIC (Standard Error)	APLES with AICc (Standard Error)
Intercept	42.963 (0.105)	42.963 (0.099)
B	1.033 (0.085)	1.009 (0.074)
C	1.520 (0.104)	1.504 (0.099)

Table 5.3. APLES Estimates and Standard Errors of Location Parameters for Taguchi's Welding Data

Term	APLES with mAIC (Standard Error)	APLES with AICc (Standard Error)
Intercept	-1.953 (0.146)	-2.142 (0.139)
C	0.669 (0.142)	0.808 (0.090)
I	0.000 (0.000)	0.124 (0.136)

Table 5.4. APLES Estimates and Standard Errors of Dispersion Parameters for Taguchi's Welding Data

Figures 5.1-5.6 illustrate our approach. Figures 5.1-5.2 show the location coefficient paths as the value of the tuning parameter λ_1 on the horizontal axis increases and λ_2 is fixed at 0.05 and 3.50 respectively. For each coefficient, as λ_1 increases the coefficient estimate decreases until it is set to 0. Similarly, figures 5.3-5.6 show the dispersion coefficient paths for $\lambda_1 = \{0.05, 0.10, 0.50, 1.50\}$. These plots give the value of the tuning parameter λ_2 on the horizontal axis versus the dispersion coefficient estimates, $\hat{\tau}_j$. For each coefficient, as λ_2 increases we see that the coefficient estimate decreases until it is set to 0. Thus, we obtain a continuous path of

solutions for both location effects as well as for dispersion effects. For each value of λ on the horizontal axis, the plot corresponds to a model. To obtain the final model we perform a grid search over the two-dimensional tuning parameter space and select the best model which minimizes one of the two information criteria given in equations (5.6) and (5.7).

Figures 5.1 and 5.2 illustrate that for this dataset, we have the same location model irrespective of the dispersion model used. In either case, all coefficients except B and C are set to zero when λ_1 equals about 0.1. Figures 5.3-5.6 illustrate how different choices of location model (corresponding to $\lambda_1 = \{0.05, 0.10, 0.50, 1.50\}$) lead to very different dispersion models.

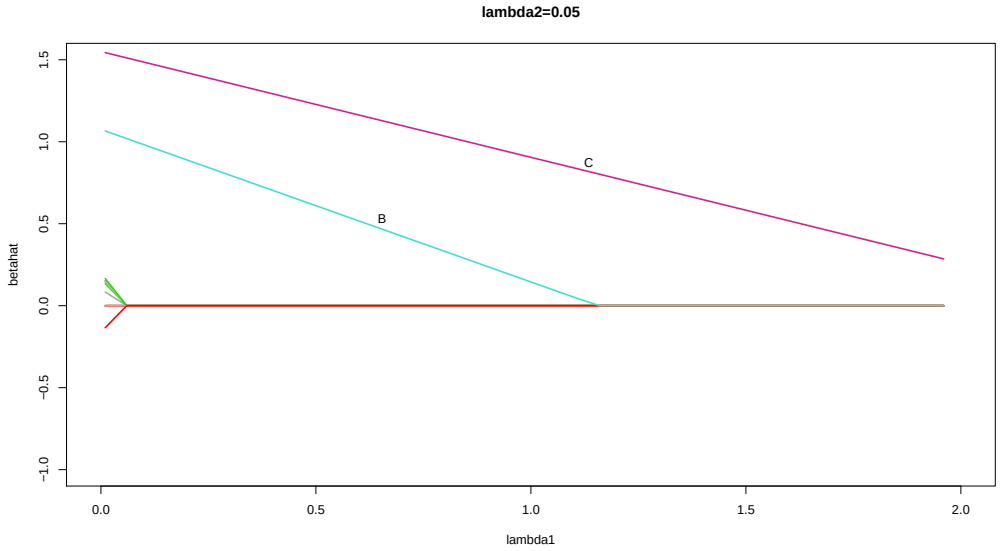

Figure 5.1. Coefficient Path for Location Effects when $\lambda_2 = 0.05$. This plot shows $\hat{\beta}_j$ on the vertical axis and λ_1 on the horizontal axis.

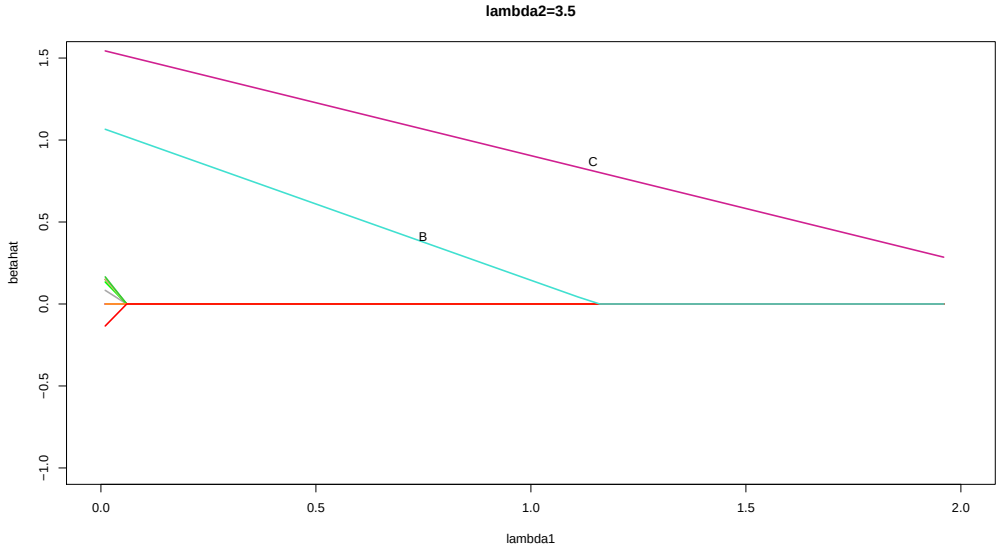

Figure 5.2. Coefficient Path for Location Effects at $\lambda_2 = 3.50$. This plot shows $\hat{\beta}_j$ on the vertical axis and λ_1 on the horizontal axis.

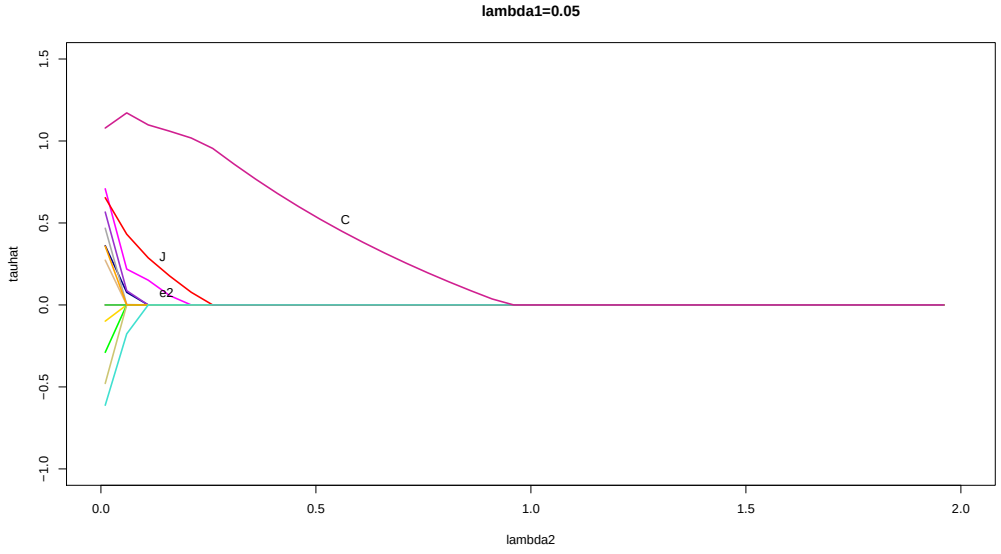

Figure 5.3. Coefficient Path for Dispersion Effects at $\lambda_1 = 0.05$. This plot shows $\hat{\tau}_j$ on the vertical axis and λ_2 on the horizontal axis.

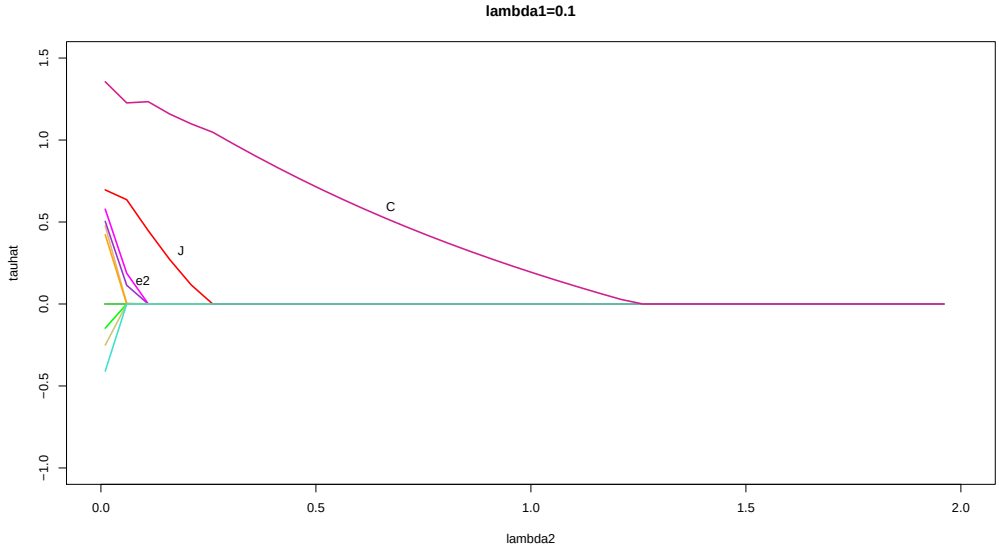

Figure 5.4. Coefficient Path for Dispersion Effects at $\lambda_1 = 0.10$. This plot shows $\hat{\tau}_j$ on the vertical axis and λ_2 on the horizontal axis.

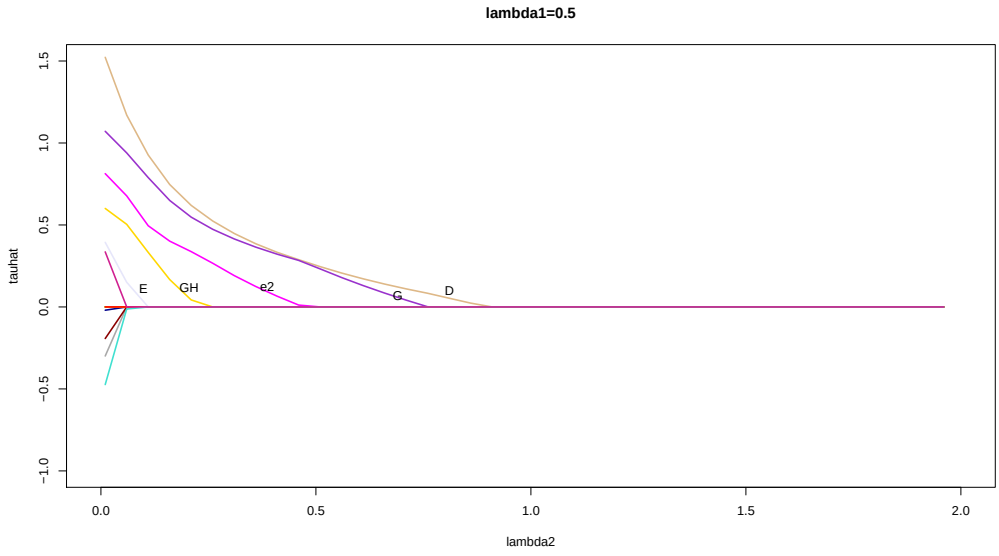

Figure 5.5. Coefficient Path for Dispersion Effects at $\lambda_1 = 0.50$. This plot shows $\hat{\tau}_j$ on the vertical axis and λ_2 on the horizontal axis.

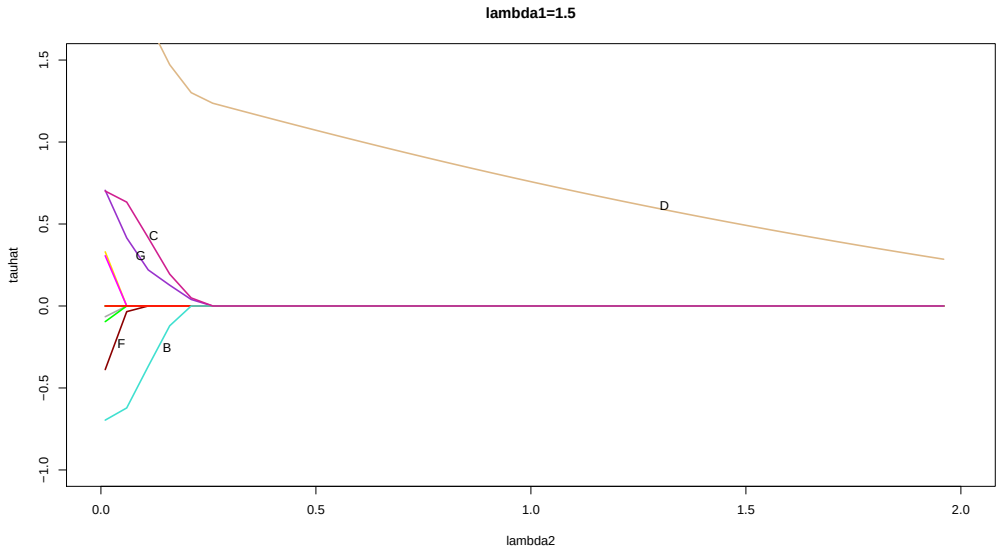

Figure 5.6. Coefficient Path for Dispersion Effects at $\lambda_1 = 1.50$. This plot shows $\hat{\tau}_j$ on the vertical axis and λ_2 on the horizontal axis.

5.5 Example 2: Dyestuff Experiment

In this example, the effect of five factors (temperature (A), starting material (B), reduction pressure (C), oven drying pressure (D), and vacuum leak (E)) on the quality of dyestuff are studied in a 2^{5-1} design (Table 5.5). The quality of the dyestuff was measured by a photoelectric spectrometer that gave a quality characteristic of the "smaller-the-better" type; that is, the lower the value recorded, the better the quality. These data are analyzed in McGrath and Lin (2001b) who show that factor E alone has a dispersion effect. Other methods such as Bergman and Hynen (1997) suggest mild dispersion effects in factor D as well as the DE interaction.

	A	B	C	D	AB	AC	AD	BC	BD	CD	DE.	CE.	BE.	AE.	E.	y
1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1	201.50
2	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	178.00
3	-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1	183.50
4	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	176.00
5	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	188.50
6	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	178.50
7	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	174.50
8	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	196.50
9	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1	255.50
10	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	240.50
11	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	208.50
12	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	244.00
13	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	274.00
14	1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	257.50
15	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1	256.00
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	274.50

Table 5.5. Data for Dyestuff Experiment (Bergman and Hynen, 1997)

Factor D has an active location effect as evidenced by the normal probability plot, Lenth's method and Box and Meyer's posterior probability approach. Given this location model, Bergman and Hynen (1997) studied their residuals for dispersion effects and found a definite dispersion effect for E and somewhat milder effects for D and DE. McGrath and Lin (2001a) used this example to show the potential pitfall of the BH approach when attempting to identify multiple dispersion effects. They demonstrated by the ML method that only factor E has a dispersion effect. However, both the BH and ML approach fit an expanded location model for each new dispersion effect that we want to test for. Another issue with the ML approach, as noted by Bursztyn and Steinberg (2006), is that we need to pre-identify a triplet to test for multiple dispersion effects. Thus it suffers from the disadvantage of having to fit several new expanded location models as well as the pre-identification using a method such as BH.

This example illustrates that APLES reaches the same conclusion as the ML method while taking into account all possible sets of combinations of location - dispersion models, rather than specific ones which have been pre-identified by the investigator. Table 5.6 compares the estimates of β_0, β_D, τ_0 and τ_E using the proposed APLES method and the maximum likelihood estimates obtained with a prior knowledge that D alone has a location effect and E alone has a dispersion effect.

Parameter	Maximum Likelihood Estimate (Standard Error)	Estimate from Proposed APLES Approach (Standard Error)
β_0	219.63 (2.30)	217.97 (2.28)
β_D	33.32 (2.30)	32.80 (1.70)
τ_0	5.03 (0.35)	5.12 (0.33)
τ_E	1.24 (0.35)	1.15 (0.29)

Table 5.6. Maximum Likelihood Estimates and Estimates from Proposed Approach for Dyestuff Experiment

5.6 Example 3: Injection Molding Experiment

Myers et al. (2009) describe an experiment to study the factors that affect shrinkage in an injection molding process. Manufactured parts that exhibit excessive shrinkage are known to cause problems at a later stage of product assembly, and thus it is of importance to the process engineers to identify which factors affect the mean and the dispersion of the shrinkage variable. The experiment is carried out in a 2^{6-2} fractional factorial design with the six factors being mold temperature (A), screw speed (B), holding time (C), cycle time (D), gate size (E) and holding pressure (F). The design matrix and responses are presented in Table 5.7. The aliasing structure is given by E=ABC, and F=BCD.

	one	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD	y
1	1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1	6
2	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	10
3	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1	32
4	1	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	60
5	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	4
6	1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	15
7	1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	26
8	1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	60
9	1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1	8
10	1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	12
11	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	34
12	1	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	60
13	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	16
14	1	1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	5
15	1	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1	37
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	52

Table 5.7. Data for Injection Molding Experiment (Myers et al., 2009)

Based on the normal probability plot of the location effect estimates, Myers et al. (2009) conclude that A, B and AB have active location effects. Assuming this is true, the estimated mean model is given by

$$\boxed{\text{Location Model 1} : \hat{\mu} = 27.313 + 6.938x_A + 17.813x_B + 5.938x_{AB}} \tag{5.8}$$

They proceed to identify factor C as having a dispersion effect using the residuals from this mean model. Figure 5.7 displays the residuals versus the two levels of factor C. One can clearly see that to reduce variation, one merely has to set C to its lower level.

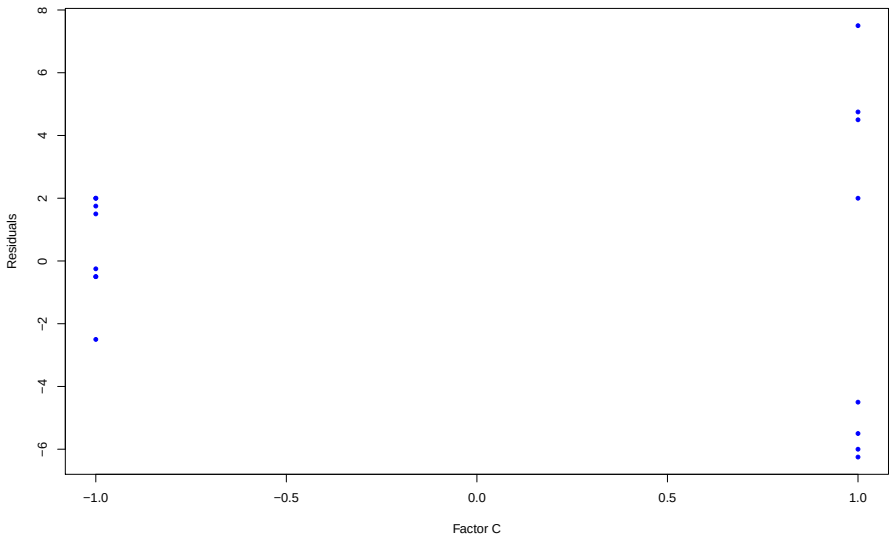

Figure 5.7. Residuals vs Factor C using location model 1 (equation (5.8)) shows that C has a dispersion effect

Liao (2000) also demonstrates that C has a dispersion effect using their method (which is based on the likelihood ratio test), the Bergman Hynen method and the Wang method. We include the Box-Meyer statistics and present these in Table 5.8. The McGrath-Lin statistic also identifies factor C as a dispersion effect. The test statistics for each of these methods is defined in section 2.3.

Factor	BM	BH	L	W
A	-0.38	0.68	0.29	0.28
B	-0.19	0.83	0.07	0.07
C	2.43	11.36	9.70	5.62
D	0.50	1.64	0.49	0.47
AB	0.11	1.11	0.02	0.02
AC	-0.40	0.67	0.31	0.30
AD	0.22	1.25	0.10	0.10
BC	-0.22	0.80	0.10	0.10
BD	-0.19	0.83	0.07	0.07
CD	0.51	1.67	0.52	0.51
ABC	-0.04	0.96	0.00	0.00
ABD	0.52	1.69	0.54	0.52
ACD	0.14	1.15	0.04	0.04
BCD	-0.30	0.74	0.18	0.18
ABCD	0.72	2.05	1.01	0.95

Table 5.8. Dispersion Effect Statistics for Factors and Interactions in the Injection Molding Experiment Based on Location Model 1. BM stands for Box and Meyer (1986b), BH for Bergman and Hynen (1997), L for Liao (2000) and W for Wang (1989).

Note that all of these methods are computed under location model 1 (Equation (5.8)).

Now consider if the location model had included A, B, AB and AD. Then the fitted model is

$$\boxed{\text{Location Model 2} : \hat{\mu} = 27.313 + 6.98x_A + 17.813x_B + 5.938x_{AB} - 2.688x_{AD}} \tag{5.9}$$

Based on this location fit, the plot of the residuals versus the two levels of C no longer show a dispersion effect as seen in Figure 5.8, and confirmed by the various dispersion effect statistics shown in Table 5.9.

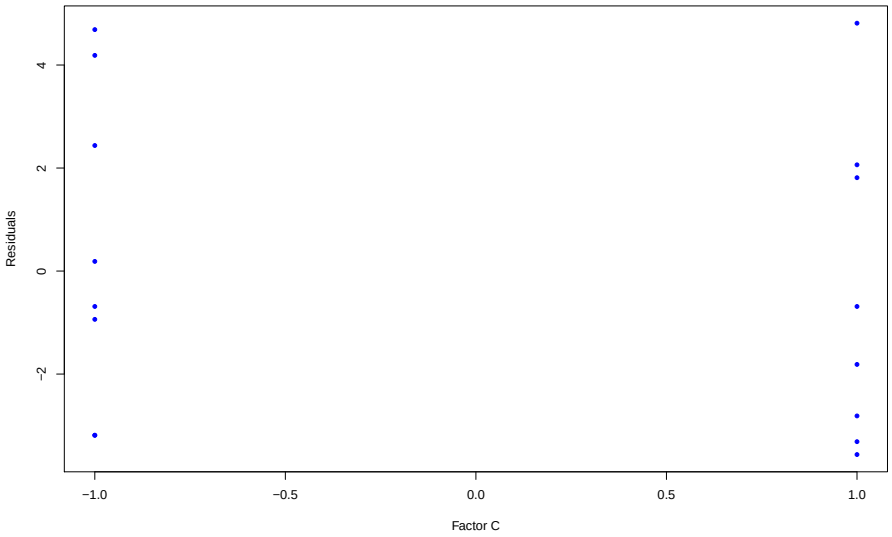

Figure 5.8. Residuals vs factor C using location model 2 (equation (5.9)) does not exhibit a dispersion effect.

Factor	BM	BH	L	ML
A	0.19	1.20	0.07	0.07
B	-0.27	0.76	0.15	0.14
C	-0.02	0.98	0.00	0.00
D	0.98	2.66	1.85	1.65
AB	0.12	1.13	0.03	0.03
AC	-0.86	0.42	1.43	1.31
AD	0.42	1.52	0.35	0.34
BC	-0.01	0.99	0.00	0.00
BD	-0.36	0.70	0.25	0.25
CD	-0.11	0.90	0.02	0.02
ABC	0.18	1.19	0.06	0.06
ABD	1.04	2.82	2.06	1.82
ACD	-0.30	0.74	0.18	0.18
BCD	-0.32	0.72	0.21	0.21
ABCD	0.07	1.08	0.01	0.01

Table 5.9. Dispersion Effect Statistics for Factors and Interactions in Injection Molding Experiment Based on Location Model 2. BM stands for Box and Meyer, BH for Bergman and Hynen, L for Liao and ML for McGrath and Lin.

Thus the addition of a single location effect removes the dispersion effect that was previously found. This phenomenon is well known. In order to satisfy effect hierarchy, the main effect of factor D should also be included in the location model, along with AD. When we include factor D however, the conclusion does not change in terms of dispersion effects, and C does not show a dispersion effect.

While noting that it is possible for the same dataset to have two or more possible sets of explanations, we found using APLES that it is unlikely that there are any dispersion effects in this dataset. APLES identified A, B, AB, AD and ACD as having active location effects. The corresponding dispersion model included no dispersion effects. Lenth's method reached the same conclusion for the active location effects. Thus if one uses a robust and adaptive method for identifying location effects then the identification reached by our method is more likely than the one in both Myers et al. (2009) and Liao (2000).

To confirm this analysis, consider maximum likelihood estimation under log-linear dispersion with prior knowledge of the true location and dispersion effects. We compute the deviance and the AICc for each of the following scenarios:

	Location	Dispersion	Deviance	AICc
i:	(A, B, AB)	(C)	73.08	94.42
ii:	(A, B, AB)	(Null)	89.31	105.31
iii:	(A, B, AB, AD, ACD)	(C)	56.56	93.13
iv:	(A, B, AB, AD, ACD)	(Null)	59.30	87.30
v:	(A, B, AB, D, AD)	(C)	64.12	100.69
vi:	(A, B, AB, D, AD)	(Null)	78.38	106.38

Table 5.10. Deviance and AICc values for six potential models for injection molding data

Since the entire set of six models we are comparing do not form a nested set, we cannot compare the deviances. The results show us that the model identified by our method has the lowest AICc of all the models and hence explains the data with the greatest parsimony. Our conclusions are in agreement with McGrath and Lin (2001b). For an alternative analysis of this data, see Loughin and Malone (2013) who assume an additive dispersion model.

5.7 Simulation Results

We performed a simulation experiment to compare our method to the following dispersion effect identification methods which have been proposed in the literature: 1. Box -Meyer (BM); 2. Bergman-Hynen (BH); 3. Harvey (H); 4. Wang (W); 5. McGrath-Lin (ML); 6. Liao (L).

For this simulation, the design matrix is a 2^{5-1} fractional factorial (i.e. one half fraction of a full 2^5 design using generator E=ABCD). Thus, there are $n = 16$ observations and 16 columns. The first column corresponds to **1** (the intercept), the next five corresponds to factors A, B, C, D and E and the remaining 10 columns are associated with the 10 two factor interactions. We induce dispersion effects in factors A and C using the approach of McGrath and Lin (2001b). We do this as follows. We generate 2000 sets of 16 standard normal random variates and store it in a (16×2000) matrix. This corresponds to the case of no dispersion effect. For factor A, we multiply all observations by 1.5 in the rows where A = +1. We do the same for factor C. Next, we repeat using 2, 3, 4 and 5. We do not create a dispersion effect in the AC interaction and so any identification of AC will be spurious.

The plots in figures 5.9 - 5.11 show the power curves for the six methods and our method APLES for each of the factor C and the AC interaction. In each of these plots, the horizontal axis indicates the magnitude of the dispersion effects which was applied to factors A and C. The methods are labeled: L for Liao, W for Wang, BM for Box and Meyer, AP for APLES, BH for Bergman and Hynen, ML for McGrath and Lin, and H for Harvey.

In figure 5.9, APLES has the fourth highest power for detecting the dispersion effect in factor A. The power curves for factor C shown in figure 5.10 are similar. In figure 5.11, we observe that all of the methods except AP, ML and H spuriously detect an effect in the AC interaction. This demonstrates that in the presence of multiple dispersion effects the methods, H, W, BM and BH are not able to control the type I error rate.

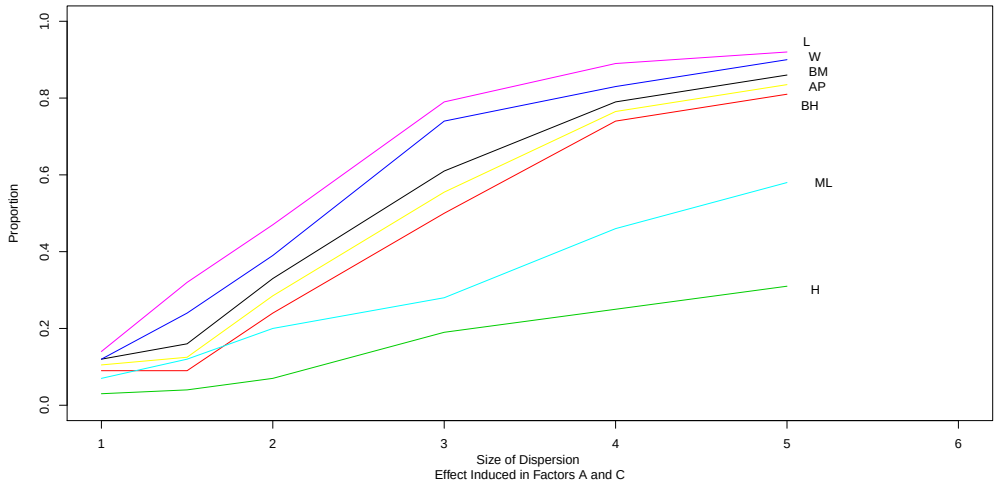

Figure 5.9. Power Curves for the Test for a Dispersion Effect in factor A Using Various Methods. L stands for Liao's method, W for Wang's method, BM for Box and Meyer's method, AP for APLES, BH for Bergman and Hynen's method, ML for McGrath and Lin's method, and H for Harvey's method.

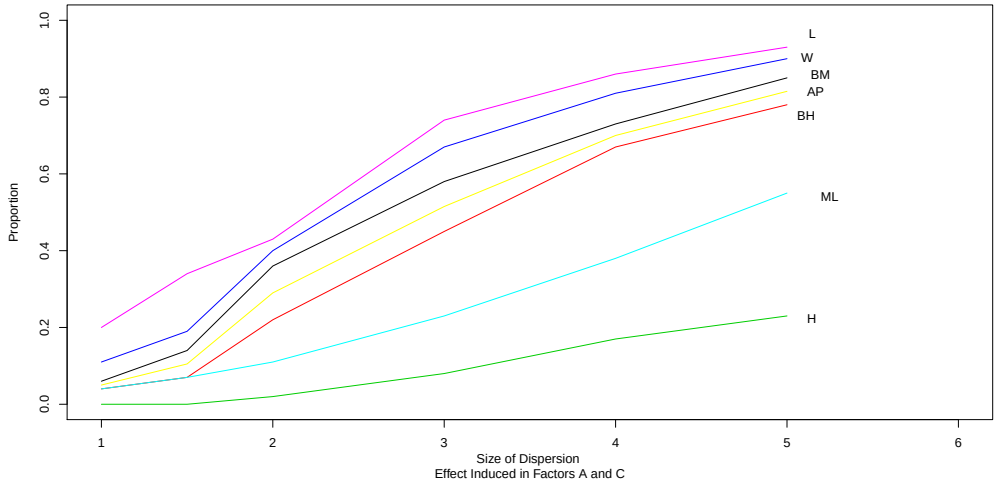

Figure 5.10. Power Curves for the Test for a Dispersion Effect in factor C Using Various Methods. L stands for Liao's method, W for Wang's method, BM for Box and Meyer's method, AP for APLES, BH for Bergman and Hynen's method, ML for McGrath and Lin's method, and H for Harvey's method.

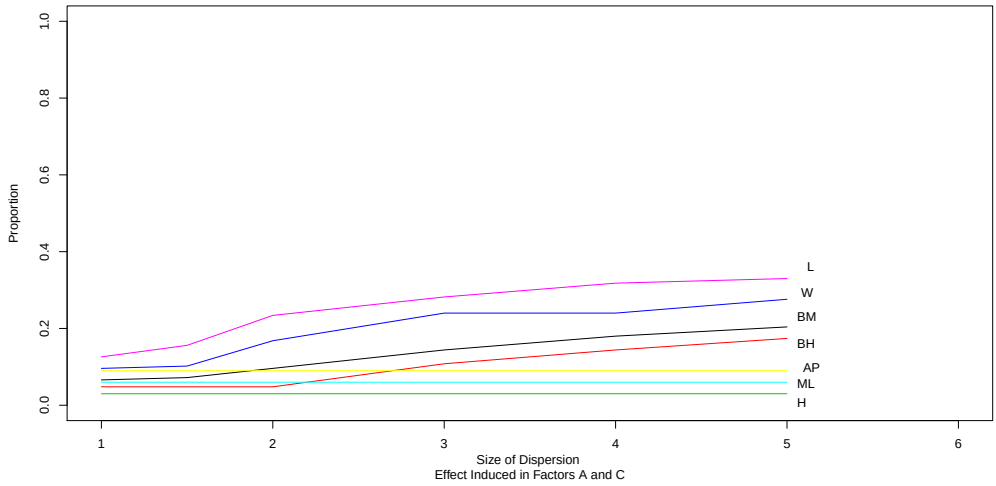

Figure 5.11. Power Curves for the Test for a Dispersion Effect in factor AC Using Various Methods. L stands for Liao's method, W for Wang's method, BM for Box and Meyer's method, AP for APLES, BH for Bergman and Hynen's method, ML for McGrath and Lin's method, and H for Harvey's method.

We conclude that while all of the methods perform well for factors A and C, the high performance of methods BM, BH, W and L is quite deceptive when one considers the AC interaction. While Harvey's method performs the best when there is no effect, it's power for detecting an effect when there is a true one is too weak for it to be useful. This leaves ML and AP. AP has a higher power for detecting a true effect when there is one but is slightly worse at controlling the false positive rate relative to ML. However, note that the ML method was used in this simulation for the triplet ABC. If we did not know before hand where to look, then we would have to search through all $\binom{15}{3}$ possible triplets or otherwise allow some of the methods such as BM to suggest a triplet to us.

5.8 Discussion

In this chapter we have presented a method which is one more resource in the tool box for engineers and other researchers who analyze designed experiments for active screening effects. The analysis of screening experiments for both location and dispersion effects is extremely important in many fields to which statistics is applied because of the overwhelming number of potential predictors for any

phenomenon. The presence of both location and dispersion effects in a particular factor is a natural phenomenon that occurs in several of the datasets that exist in the literature and in actual experiments.

We have used APLES for the analysis of screening experiments and found it to be competitive to the methods that have been previously proposed in the literature particularly for the detection of dispersion effects. By drawing on ideas from high dimensional data analysis, we put the problem in a framework that is quite general relative to some of the earlier methods which inherently require $n < 2p$.

APLES requires minimal assumptions compared to several of the methods in the literature. Unlike the Box and Meyer (1986b) and Bergman and Hynen (1997) approaches, it does not work only under constraints on the null hypothesis of dispersion effects. The statistic proposed in Box and Meyer (1986b) requires all dispersion effects to be null, and that proposed in Bergman and Hynen (1997) requires all to be null except the specific factor being tested for dispersion effects.

We illustrate the actual plots which show the shrinkage and selection operation of the APLES estimator for various values of the tuning parameters λ_1 and λ_2. These plots (5.1 - 5.6) show the coefficient paths. We also compare the performance using two criteria suggested in the literature for small samples, the AICc and the mAIC. We found the mAIC which has a quadratic term for the number of variables selected to be a more severe penalty in terms of the number of variables selected.

In our simulation results, we found that APLES performs better than the BH, ML and H methods in terms of power to detect a true dispersion effect, but not as well as the L, W and BM methods. But, very importantly it performs better than the L, W, BM and BH methods in identifying a non-active effect accurately.

Chapter 6

Contributions and Future Work

In this chapter, we summarize the major contributions of this dissertation in section 6.1 and we describe some future directions in which the ideas presented here may be extended in section 6.2.

6.1 Contributions

Our main contribution in this dissertation is the application of modern penalized likelihood methods that utilize sparsity-inducing penalties to the analysis of datasets that arise from designed experiments. Designed experiments are applied in many different situations with the goal of improving the quality of a process as well as understanding the properties of a process. In robust parameter design (RPD) experiments, we seek to identify those control factors which interact with noise factors so that they may be used to make the process response insensitive to variation. In unreplicated fractional factorial experiments, several potential factors are screened to determine which ones affect the mean and variance of a process response. The identified factors may then be studied further in subsequent experimentation.

6.1.1 Adaptive Penalized Likelihood Effects Selection (APLES)

We introduce the APLES estimation procedure in chapter 3. It is a doubly penalized likelihood procedure with the ability to perform variable selection and estimation simultaneously. We compared APLES to similar proposals in the literature (HIPPO and HHR) which use different penalty functions. We also provided

a justification for the use of APLES within the RPD framework.

We described the choice of the adaptive weights as well as the tuning parameters for APLES. We described and implemented the cyclic coordinate descent (CCD) algorithm of Friedman et al. (2010) for APLES. We also extended the work of Zou (2006) for obtaining the standard errors to APLES estimates.

We showed the finite sample performance of APLES through extensive simulation. We considered four different scenarios that usually arise with real data sets. Our simulation results indicate that the APLES estimation procedure is a good competitor to existing methods as measured both by prediction accuracy and identification of the true support of the mean and variance parameter vectors.

Finally, we described how to apply APLES estimation when the variance submodel is linear in the predictors rather than log-linear.

6.1.2 Robust Parameter Design Using APLES

In chapter 4, we applied APLES for the analysis of RPD. We showed how APLES fits within the RPD and quality improvement paradigm.

One of the main applications for the penalized likelihood methods is in the situation where the sample size is smaller than the number of predictors. We showed that this situation arises even in RPD owing to the interactions between control and noise factors.

We considered three well-known quality objective functions that are well-established in the RPD literature. We extended the objective function proposed by Miro-Quesada and Del Castillo (2004) for the situation when we have parameter estimation uncertainty to our heteroscedastic RPD model.

We described the analysis of two well-known data sets using APLES and compared to previous published analyses.

Finally, we performed a simulation study to show the performance of APLES versus the ordinary least squares and maximum likelihood estimators. We consider fifteen different settings that may arise depending on the structure of the $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ vectors. We based our comparisons on three different criteria: simulated integrated mean square error of the estimated response surfaces, distance between the optima for the true objective functions and the predicted optima based on

estimated objective functions, and values obtained when the predicted optima are plugged into the true objective functions.

6.1.3 Analysis of Screening Experiments Using APLES

In chapter 5, we considered the application of APLES to the analysis of location and dispersion effects in unreplicated fractional factorial experiments. Due to the orthogonality of the design matrix, the proposed CCD algorithm simplifies greatly.

One issue that has been well noted in the literature has to do with the identification of multiple dispersion effects. We performed a simulation study to compare the power of APLES and other methods to identify the truly active dispersion effects. Among all of the methods, APLES performs well in identifying the truly active dispersion effects as well as controlling the level for non-active effects.

6.2 Future Work

In this section, we present three directions in which the methodology proposed in this dissertation may be extended. The first concerns the application of APLES to analysis of RPD. We will consider the use of group penalties which ensure that groups of predictors are either included or excluded in the model together. The second direction is an application of APLES to the analysis of supersaturated designs. Lastly, we consider a Bayesian extension of APLES.

6.2.1 Extension of APLES Using Group Penalty and Application to RPD

In the experimental designs that are used for Robust Parameter Design we usually estimate main effects as well as interaction terms. A well-known principle which has proven to be useful in the analysis of designed experiments is the effect heredity principle discussed in several books on experimental design such as Wu and Hamada (2000). According to this principle, an interaction can only be present in the model if one or both of its parent effects is also included. As an example, the AB interaction is only included in the model if either factor A or factor B or both are also included. Sometimes the principle as stated above is called a *weak* heredity principle (Yuan et al., 2007) and a corresponding *strong* heredity principle

is defined as the requirement that all of the parent effects have to be present in order for an interaction to be included in the model.

The APLES estimation procedure as presented in this dissertation does not enforce the effect heredity principle. However, group penalties have been proposed in the literature (Yuan and Lin, 2006) which have the capability of ensuring that entire groups of terms are included or excluded at a time. As part of our future work arising from this dissertation, we can extend the APLES estimation procedure to a group APLES estimation which will enforce either a strong or weak heredity principle as specified by the researcher.

6.2.2 Analysis of Data from Supersaturated Designs Using APLES

A supersaturated design is one in which the degrees of freedom for all its main effects and the intercept term exceed the total number of distinct factor level combinations of the design. They are used primarily for screening factor main effects. Naturally, analysis of these designs presents a great challenge. One further future application of the APLES methodology is to the analysis of supersaturated designs.

6.2.3 Bayesian Extension of APLES

From a Bayesian point of view, the LASSO or L_1 penalty is equivalent to placing a Laplace or double exponential prior on the coefficients of the linear model. It turns out that using this prior induces the sparsity properties that are enjoyed by LASSO. Therefore, even if we did not begin with the knowledge that L_1 penalty produces sparse coefficient structure, but had consulted with a process expert before analysis and based on this had used the Laplace prior we would arrive at the benefit of sparsity.

This suggests that we may reverse the process. In the analysis of screening experiments, it usually happens that the engineers or process experts may have some prior information on the distribution of the parameters which can then be incorporated into our model formulation. For example, where appropriate we may place both L_1 and L_2 penalties on the parameters leading to a variant of

APLES which acts like the elastic net. The elastic net is a penalized regression method proposed by Zou and Hastie (2005) which linearly combines the L_1 and L_2 penalties.

Appendices

Appendix A

Simulation Results

A.1 Simulation Results for Chapter 3

		β			τ		
Method	PE	MSE	Specificity	Sensitivity	MSE	Specificity	Sensitivity
			$n = 60$				
HHR(AIC)	281.48 (11.62)	8.18 (0.37)	0.22	**1**	**3.39 (0.06)**	0.57	**0.72**
HHR (BIC)	309.92 (12.54)	9.07 (0.4)	0.22	**1**	3.53 (0.06)	0.63	0.66
HHR (AICC)	281.48 (11.62)	8.18 (0.37)	0.22	**1**	**3.39 (0.06)**	0.57	**0.72**
APLES(AIC)	**267.74 (11.7)**	**7.2 (0.37)**	**0.61**	**1**	**3.39 (0.05)**	**0.79**	0.56
APLES (BIC)	**260.53 (11.01)**	**6.49 (0.34)**	**0.7**	0.99	3.47 (0.05)	**0.79**	0.58
APLES (AICC)	**267.74 (11.7)**	**7.2 (0.37)**	**0.61**	**1**	**3.39 (0.05)**	**0.79**	0.56
HIPPO(AIC)	407.61 (15.15)	12.28 (0.5)	0.15	**1**	6.15 (0.14)	0.59	0.65
HIPPO (BIC)	410.24 (17.28)	11.89 (0.53)	0.3	**1**	5.59 (0.14)	0.64	0.62
HIPPO (AICC)	422.04 (15.27)	12.71 (0.49)	0.13	**1**	6.35 (0.15)	0.63	0.63
			$n = 120$				
HHR(AIC)	76.47 (1.57)	0.8 (0.02)	0.1	**1**	1.28 (0.02)	0.31	**0.99**
HHR (BIC)	105.62 (2.87)	0.99 (0.03)	0.75	**1**	1.73 (0.03)	0.84	0.86
HHR (AICC)	76.47 (1.57)	0.8 (0.02)	0.1	**1**	1.28 (0.02)	0.31	**0.99**
APLES(AIC)	61.75 (1.47)	0.63 (0.02)	0.48	**1**	0.78 (0.02)	**0.84**	0.91
APLES (BIC)	**39.53 (1.19)**	**0.36 (0.01)**	**0.84**	**1**	**0.7 (0.02)**	**0.85**	0.9
APLES (AICC)	59.3 (1.49)	0.6 (0.02)	0.52	**1**	0.76 (0.02)	**0.84**	0.91
HIPPO(AIC)	76.64 (2.31)	0.79 (0.02)	0.4	**1**	1.68 (0.04)	0.44	0.96
HIPPO (BIC)	**29.29 (1.29)**	**0.26 (0.01)**	**0.94**	**1**	**0.61 (0.02)**	**0.84**	0.93
HIPPO (AICC)	76.64 (2.31)	0.79 (0.02)	0.4	**1**	1.68 (0.04)	0.44	0.96
			$n = 200$				
HHR(AIC)	50.26 (3.24)	0.28 (0.02)	0.27	**1**	0.46 (0.01)	**0.86**	0.98
HHR (BIC)	75.96 (5.15)	0.37 (0.02)	0.4	**1**	0.65 (0.03)	**0.86**	0.98
HHR (AICC)	48.93 (3.24)	0.27 (0.02)	0.29	**1**	0.46 (0.01)	**0.86**	0.98
APLES(AIC)	42.02 (1.11)	0.19 (0)	**0.93**	**1**	0.65 (0.02)	0.72	0.9
APLES (BIC)	**29.44 (0.94)**	**0.14 (0)**	**0.93**	**1**	0.81 (0.02)	**0.94**	0.83
APLES (AICC)	42.02 (1.11)	0.19 (0)	**0.93**	**1**	0.65 (0.02)	0.72	0.9
HIPPO(AIC)	27.75 (1.05)	0.15 (0.01)	0.76	**1**	**0.25 (0.01)**	0.74	**1**
HIPPO (BIC)	**15.96 (0.49)**	**0.08 (0)**	**1**	**1**	**0.19 (0.01)**	0.69	**1**
HIPPO (AICC)	27.75 (1.05)	0.15 (0.01)	0.76	**1**	**0.25 (0.01)**	0.74	**1**

Table A.1. Simulation scenario 1 when model is correctly specified with $n = 60, 120$ and 200.

		β			τ		
Method	PE	MSE	Specificity	Sensitivity	MSE	Specificity	Sensitivity
			$\alpha = 1$				
HHR(AIC)	96.97 (1.48)	1.38 (0.03)	0.39	**1**	1.02 (0.02)	**0.98**	**1**
HHR (BIC)	122.07 (2.2)	1.38 (0.03)	0.77	**1**	**0.6 (0.02)**	0.97	**1**
HHR (AICC)	96.97 (1.48)	1.38 (0.03)	0.39	**1**	1.02 (0.02)	**0.98**	**1**
APLES(AIC)	70.41 (1.59)	0.9 (0.03)	0.8	**1**	0.82 (0.02)	**0.98**	**1**
APLES (BIC)	**63.59 (1.45)**	**0.75 (0.02)**	**0.91**	**1**	0.73 (0.02)	**0.98**	**1**
APLES (AICC)	**64.07 (1.34)**	**0.78 (0.02)**	0.86	**1**	0.77 (0.02)	**0.99**	**1**
HIPPO(AIC)	67.92 (1.71)	0.88 (0.03)	0.75	**1**	1.17 (0.03)	0.86	**1**
HIPPO (BIC)	71.1 (2.03)	0.8 (0.02)	**0.91**	**1**	**0.69 (0.02)**	**0.98**	**1**
HIPPO (AICC)	66.99 (1.71)	0.87 (0.03)	0.77	**1**	1.13 (0.03)	0.87	**1**
			$\alpha = 3$				
HHR(AIC)	170.94 (3.23)	1.87 (0.04)	0.65	**1**	1.77 (0.04)	0.32	0.72
HHR (BIC)	230.85 (4.34)	2.49 (0.04)	0.66	**1**	**0.32 (0.01)**	**0.82**	0.31
HHR (AICC)	174.67 (3.32)	1.91 (0.04)	0.64	**1**	1.61 (0.04)	0.38	0.68
APLES(AIC)	**114.93 (2.89)**	**1.31 (0.03)**	**0.89**	**1**	1.19 (0.02)	0.67	0.44
APLES (BIC)	148.29 (3.57)	1.6 (0.04)	**0.92**	**1**	**0.45 (0.02)**	**0.86**	0.26
APLES (AICC)	**114.93 (2.89)**	**1.31 (0.03)**	**0.89**	**1**	1.19 (0.02)	0.67	0.44
HIPPO(AIC)	117.16 (3.17)	1.43 (0.04)	0.75	**1**	3.11 (0.04)	0.3	**0.75**
HIPPO (BIC)	193.05 (6.88)	2.11 (0.06)	0.8	**1**	0.98 (0.05)	0.76	0.35
HIPPO (AICC)	117.16 (3.17)	1.43 (0.04)	0.75	**1**	3.11 (0.04)	0.3	**0.75**
			$\alpha = 5$				
HHR(AIC)	263.19 (7.6)	3.74 (0.12)	0.07	**1**	5.99 (0.1)	0.17	0.86
HHR (BIC)	504.45 (18.36)	6.61 (0.24)	0.55	0.99	**2.64 (0.16)**	**0.67**	0.44
HHR (AICC)	263.19 (7.6)	3.74 (0.12)	0.07	**1**	5.99 (0.1)	0.17	0.86
APLES(AIC)	**223.6 (10.71)**	**3.07 (0.16)**	0.53	**1**	3.68 (0.08)	**0.49**	0.59
APLES (BIC)	**192.6 (10.04)**	**2.53 (0.15)**	**0.73**	**1**	**2.94 (0.07)**	**0.49**	0.58
APLES (AICC)	**223.6 (10.71)**	**3.07 (0.16)**	0.53	**1**	3.68 (0.08)	**0.49**	0.59
HIPPO(AIC)	301.12 (10.21)	4.22 (0.16)	0.18	**1**	8.51 (0.17)	0.15	**0.87**
HIPPO (BIC)	337.27 (17.63)	4.4 (0.23)	**0.57**	0.99	5.39 (0.21)	0.3	0.76
HIPPO (AICC)	301.12 (10.21)	4.22 (0.16)	0.18	**1**	8.51 (0.17)	0.15	**0.87**

Table A.2. Simulation scenario 2 when outliers are present with $n = 100$ and $n_{outlier} = 10$. $\alpha = 1, 3$ and 5 indicates increasing extremeness of the outliers.

		β			τ		
Method	PE	MSE	Specificity	Sensitivity	MSE	Specificity	Sensitivity
			$df = 3$				
HHR(AIC)	156.62 (4.13)	1.7 (0.05)	0.77	**1**	**0.48 (0.02)**	**0.73**	**1**
HHR (BIC)	169.77 (3.15)	1.78 (0.03)	0.88	**1**	**0.43 (0.02)**	0.71	**1**
HHR (AICC)	156.62 (4.13)	1.7 (0.05)	0.77	**1**	**0.48 (0.02)**	**0.73**	**1**
APLES(AIC)	55.76 (1.23)	0.61 (0.01)	**1**	**1**	1.04 (0.03)	0.71	**1**
APLES (BIC)	54.63 (1.23)	0.6 (0.01)	**1**	**1**	0.96 (0.03)	**0.74**	**1**
APLES (AICC)	55.76 (1.23)	0.61 (0.01)	**1**	**1**	1.04 (0.03)	0.71	**1**
HIPPO(AIC)	**22.47 (0.87)**	**0.26 (0.01)**	**1**	**1**	2.77 (0.05)	0.44	**1**
HIPPO (BIC)	**20.65 (0.89)**	**0.24 (0.01)**	0.99	**1**	2.41 (0.06)	0.57	**1**
HIPPO (AICC)	**22.47 (0.87)**	**0.26 (0.01)**	**1**	**1**	2.77 (0.05)	0.44	**1**
			$df = 5$				
HHR(AIC)	39.63 (0.81)	0.45 (0.01)	0.53	**1**	1.44 (0.03)	0.53	**1**
HHR (BIC)	61.64 (2.04)	0.68 (0.03)	0.61	**1**	0.94 (0.03)	0.65	**1**
HHR (AICC)	39.63 (0.81)	0.45 (0.01)	0.53	**1**	1.44 (0.03)	0.53	**1**
APLES(AIC)	35.7 (1.06)	0.35 (0.01)	**0.97**	**1**	**0.65 (0.02)**	**0.84**	**1**
APLES (BIC)	**32.92 (0.96)**	**0.32 (0.01)**	**0.98**	**1**	**0.68 (0.02)**	**0.84**	**1**
APLES (AICC)	35.7 (1.06)	0.35 (0.01)	**0.97**	**1**	**0.65 (0.02)**	**0.84**	**1**
HIPPO(AIC)	34.84 (1.08)	0.43 (0.01)	0.66	**1**	2.7 (0.05)	0.49	**1**
HIPPO (BIC)	**20.7 (0.62)**	**0.22 (0.01)**	0.95	**1**	2.17 (0.04)	0.5	**1**
HIPPO (AICC)	34.84 (1.08)	0.43 (0.01)	0.66	**1**	2.7 (0.05)	0.49	**1**
			$df = \infty$				
HHR(AIC)	84.58 (1.83)	0.94 (0.03)	0.71	**1**	1.02 (0.03)	0.47	**1**
HHR (BIC)	118.86 (2.24)	1.18 (0.02)	0.86	**1**	**0.6 (0.02)**	0.61	**1**
HHR (AICC)	84.58 (1.83)	0.94 (0.03)	0.71	**1**	1.02 (0.03)	0.47	**1**
APLES(AIC)	32.99 (0.89)	0.34 (0.01)	**1**	**1**	1.05 (0.02)	**0.69**	**1**
APLES (BIC)	36.25 (1.09)	0.38 (0.01)	0.98	**1**	**0.95 (0.03)**	**0.73**	**1**
APLES (AICC)	32.99 (0.89)	0.34 (0.01)	**1**	**1**	1.05 (0.02)	**0.69**	**1**
HIPPO(AIC)	**21.34 (0.56)**	**0.23 (0.01)**	0.96	**1**	2.44 (0.04)	0.34	**1**
HIPPO (BIC)	29.81 (1.17)	0.31 (0.01)	0.99	**1**	1.61 (0.04)	0.58	**1**
HIPPO (AICC)	**21.34 (0.56)**	**0.23 (0.01)**	0.96	**1**	2.44 (0.04)	0.34	**1**

Table A.3. Simulation scenario 3 with non-normal errors. Errors are sampled from a t distribution with degrees of freedom 3, 5 and 10.

		β			τ		
Method	PE	MSE	Specificity	Sensitivity	MSE	Specificity	Sensitivity
			$n = 60$				
HHR(AIC)	29.12 (0.5)	0.9 (0.02)	0.18	**1**	6.17 (0.05)	**1**	0.14
HHR (BIC)	36.03 (0.65)	0.74 (0.02)	0.58	**1**	5.01 (0.05)	0.98	0.18
HHR (AICC)	29.37 (0.51)	0.87 (0.02)	0.22	**1**	6.03 (0.05)	**1**	0.14
APLES(AIC)	18.24 (0.46)	0.31 (0.01)	**0.95**	**1**	4.91 (0.05)	0.86	0.38
APLES (BIC)	18.2 (0.47)	0.3 (0.01)	**0.98**	**1**	**4.76 (0.04)**	0.91	0.3
APLES (AICC)	18.24 (0.46)	0.31 (0.01)	**0.95**	**1**	**4.91 (0.05)**	0.86	0.38
HIPPO(AIC)	**13.09 (0.53)**	**0.22 (0.01)**	0.93	**1**	6.63 (0.09)	0.43	**0.81**
HIPPO (BIC)	14.17 (0.43)	0.25 (0.01)	**0.95**	**1**	4.98 (0.05)	0.88	0.39
HIPPO (AICC)	**13.11 (0.53)**	**0.22 (0.01)**	0.93	**1**	6.54 (0.09)	0.45	**0.8**
			$n = 120$				
HHR(AIC)	33.69 (0.61)	0.3 (0.01)	0.78	**1**	4.32 (0.02)	**1**	0.16
HHR (BIC)	37.52 (0.65)	0.32 (0.01)	0.83	**1**	4.25 (0.02)	**1**	0.17
HHR (AICC)	33.69 (0.61)	0.3 (0.01)	0.78	**1**	4.32 (0.02)	**1**	0.16
APLES(AIC)	28.21 (0.54)	0.22 (0)	**1**	**1**	**3.31 (0.03)**	0.95	0.61
APLES (BIC)	20.99 (0.51)	0.17 (0)	**1**	**1**	3.93 (0.04)	0.98	0.36
APLES (AICC)	28.21 (0.54)	0.22 (0)	**1**	**1**	**3.31 (0.03)**	0.95	0.61
HIPPO(AIC)	**2.66 (0.12)**	**0.02 (0)**	**1**	**1**	5.93 (0.05)	0.45	**0.99**
HIPPO (BIC)	3.17 (0.14)	**0.03 (0)**	**1**	**1**	5.33 (0.06)	0.57	0.97
HIPPO (AICC)	**2.65 (0.12)**	**0.02 (0)**	**1**	**1**	5.91 (0.05)	0.45	**0.99**
			$n = 200$				
HHR(AIC)	12.56 (0.27)	0.06 (0)	0.93	**1**	3.12 (0.02)	0.25	0.98
HHR (BIC)	12.56 (0.27)	0.06 (0)	0.93	**1**	3.12 (0.02)	0.25	0.98
HHR (AICC)	12.56 (0.27)	0.06 (0)	0.93	**1**	3.12 (0.02)	0.25	0.98
APLES(AIC)	4.76 (0.14)	0.02 (0)	**1**	**1**	**2.93 (0.02)**	**0.86**	0.93
APLES (BIC)	4.76 (0.14)	0.02 (0)	**1**	**1**	**2.93 (0.02)**	**0.86**	0.93
APLES (AICC)	4.76 (0.14)	0.02 (0)	**1**	**1**	**2.93 (0.02)**	**0.86**	0.93
HIPPO(AIC)	**2.75 (0.1)**	**0.01 (0)**	0.97	**1**	3.71 (0.03)	0.25	**1**
HIPPO (BIC)	**2.75 (0.1)**	**0.01 (0)**	0.97	**1**	3.71 (0.03)	0.25	**1**
HIPPO (AICC)	**2.75 (0.1)**	**0.01 (0)**	0.97	**1**	3.71 (0.03)	0.25	**1**

Table A.4. Simulation scenario 4 when the variance model is misspecified with $n = 60, 120$ and 200.

A.2 Simulation Results for Chapter 4

A.2.1 SIMSEM and SIMSEV

Setting	Method	SIMSEM	SIMSEV
1	MLO	0.016	0.442
	MLE	0.056	0.743
	LS	0.054	0.939
	APLES	0.056	0.678
2	MLO	0.018	0.448
	MLE	0.074	0.732
	LS	0.069	1.26
	APLES	0.069	0.635
3	MLO	0.014	0.385
	MLE	0.062	1.297
	LS	0.063	3.345
	APLES	0.057	1.047
4	MLO	0.006	0.311
	MLE	0.083	2.547
	LS	0.116	11.853
	APLES	0.079	2.311
5	MLO	0.007	0.942
	MLE	0.078	4.145
	LS	0.136	21.254
	APLES	0.075	4.105

Table A.5. Values of simulated integrated mean square error of the mean (SIMSEM) and variance (SIMSEV) models based on 500 generated datasets for settings 1 - 5.

Setting	Method	SIMSEM	SIMSEV
6	MLO	0.039	0.775
	MLE	0.059	1.157
	LS	0.058	1.301
	APLES	0.058	1.048
7	MLO	0.037	0.689
	MLE	0.062	1.207
	LS	0.055	1.975
	APLES	0.059	0.942
8	MLO	0.037	1.109
	MLE	0.059	1.637
	LS	0.065	3.692
	APLES	0.054	1.448
9	MLO	0.035	1.557
	MLE	0.072	2.578
	LS	0.087	12.798
	APLES	0.07	2.229
10	MLO	0.032	1.603
	MLE	0.095	6.81
	LS	0.135	25.145
	APLES	0.087	6.09

Table A.6. Values of simulated integrated mean square error of the mean (SIMSEM) and variance (SIMSEV) models based on 500 generated datasets for settings 6 - 10.

Setting	Method	SIMSEM	SIMSEV
11	MLO	0.048	3.062
	MLE	0.053	3.131
	LS	0.048	2.999
	APLES	0.051	3.029
12	MLO	0.071	3.519
	MLE	0.073	3.673
	LS	0.073	3.952
	APLES	0.072	3.685
13	MLO	0.076	8.592
	MLE	0.082	8.665
	LS	0.085	11.035
	APLES	0.078	8.616
14	MLO	0.062	7.031
	MLE	0.071	7.223
	LS	0.091	17.686
	APLES	0.067	7.536
15	MLO	0.075	8.635
	MLE	0.08	8.635
	LS	0.114	27.486
	APLES	0.078	8.902

Table A.7. Values of simulated integrated mean square error of the mean (SIMSEM) and variance (SIMSEV) models based on 500 generated datasets for settings 11 - 15.

A.2.2 Values of True Objective Functions Evaluated at Predicted Optima

The following tables (A.8 - A.22) report values of the true objective function when evaluated at the predicted optima using various estimation methods. The values reported are the medians over 500 generated datasets and in parentheses we report the $\frac{1.486\times \text{ MAD}}{\sqrt{500}}$, where MAD is the median absolute deviation.

Setting 1: CE			
	MSE	TTB	MTB
True	1.01 (0)	1.01 (0)	1.05 (0)
MLO	1.03 (0)	1.05 (0.01)	1.05 (0)
MLE	1.04 (0.01)	1.08 (0.01)	1.05 (0)
LS	1.05 (0)	1.11 (0.02)	1.06 (0)
APLES	1.03 (0)	1.05 (0.01)	1.06 (0)
Setting 1: U			
	MSE	TTB	MTB
True	1.01 (0)	1.01 (0)	1.05 (0)
MLO	1.03 (0)	1.05 (0)	1.05 (0)
MLE	1.04 (0.01)	1.08 (0.01)	1.05 (0)
LS	1.11 (0.02)	1.11 (0.02)	1.06 (0)
APLES	1.03 (0)	1.05 (0.01)	1.05 (0)

Table A.8. Setting 1

Setting 2: CE			
	MSE	TTB	MTB
True	0.61 (0)	0.61 (0)	0.96 (0)
MLO	0.63 (0)	0.63 (0)	0.96 (0)
MLE	0.7 (0.01)	0.76 (0.03)	0.99 (0.01)
LS	1.2 (0.02)	1.34 (0.07)	1.07 (0.02)
APLES	0.69 (0.01)	0.81 (0.03)	0.96 (0)
Setting 2: U			
	MSE	TTB	MTB
True	0.61 (0)	0.61 (0)	0.96 (0)
MLO	0.63 (0)	0.63 (0)	0.96 (0)
MLE	0.7 (0.01)	0.76 (0.03)	0.99 (0.01)
LS	1.26 (0.06)	1.29 (0.08)	1.07 (0.02)
APLES	0.69 (0.01)	0.78 (0.03)	0.96 (0)

Table A.9. Setting 2

Setting 3: CE			
	MSE	TTB	MTB
True	0.37 (0)	0.37 (0)	0.89 (0)
MLO	0.39 (0)	0.39 (0)	0.9 (0)
MLE	0.4 (0.01)	0.42 (0.01)	0.9 (0)
LS	0.84 (0.07)	0.74 (0.08)	1.28 (0)
APLES	0.39 (0)	0.39 (0)	0.9 (0)
Setting 3: U			
	MSE	TTB	MTB
True	0.37 (0)	0.37 (0)	0.89 (0)
MLO	0.39 (0)	0.39 (0)	0.9 (0)
MLE	0.4 (0.01)	0.42 (0.01)	0.9 (0)
LS	0.59 (0.04)	0.71 (0.07)	1.28 (0)
APLES	0.39 (0)	0.39 (0)	0.9 (0)

Table A.10. Setting 3

Setting 4: CE			
	MSE	TTB	MTB
True	0.23 (0)	0.23 (0)	0.84 (0)
MLO	0.23 (0)	0.23 (0)	0.85 (0)
MLE	0.25 (0)	0.27 (0.01)	0.85 (0)
LS	1.33 (0.16)	1.47 (0.2)	1.24 (0.05)
APLES	0.26 (0.01)	0.28 (0.01)	0.85 (0)
Setting 4: U			
	MSE	TTB	MTB
True	0.23 (0)	0.23 (0)	0.84 (0)
MLO	0.23 (0)	0.23 (0)	0.85 (0)
MLE	0.25 (0)	0.28 (0.01)	0.85 (0)
LS	1.54 (0.23)	1.7 (0.18)	1.24 (0.05)
APLES	0.26 (0.01)	0.29 (0.01)	0.85 (0)

Table A.11. Setting 4

Setting 5: CE			
	MSE	TTB	MTB
True	0.22 (0)	0.21 (0)	0.95 (0)
MLO	0.22 (0)	0.22 (0)	0.95 (0)
MLE	0.26 (0.01)	0.29 (0.01)	0.95 (0)
LS	0.37 (0.02)	0.32 (0.02)	0.99 (0.01)
APLES	0.25 (0.01)	0.27 (0.01)	0.95 (0)
Setting 5: U			
	MSE	TTB	MTB
True	0.21 (0)	0.21 (0)	0.95 (0)
MLO	0.21 (0)	0.22 (0)	0.95 (0)
MLE	0.26 (0.01)	0.29 (0.01)	0.95 (0)
LS	0.36 (0.02)	0.32 (0.02)	1 (0.01)
APLES	0.26 (0.01)	0.27 (0.01)	0.95 (0)

Table A.12. Setting 5

Setting 6: CE			
	MSE	TTB	MTB
True	1.01 (0)	1.01 (0)	2.5 (0)
MLO	1.07 (0.01)	1.1 (0.02)	2.5 (0)
MLE	1.15 (0.01)	5.59 (1.18)	2.51 (0)
LS	1.12 (0.01)	5.57 (1)	2.51 (0)
APLES	1.1 (0.01)	5.59 (0.99)	2.5 (0)
Setting 6: U			
	MSE	TTB	MTB
True	1.01 (0)	1.01 (0)	2.5 (0)
MLO	1.07 (0.01)	1.1 (0.02)	2.5 (0)
MLE	1.15 (0.01)	3.17 (1.18)	2.51 (0)
LS	1.12 (0.01)	5.24 (0.53)	2.51 (0)
APLES	1.1 (0.01)	3.17 (0.45)	2.5 (0)

Table A.13. Setting 6

Setting 7: CE			
	MSE	TTB	MTB
True	1.55 (0)	1.55 (0)	2.45 (0)
MLO	1.58 (0)	1.63 (0.02)	2.46 (0)
MLE	1.64 (0)	1.67 (0.02)	2.46 (0)
LS	1.65 (0.01)	1.66 (0.01)	2.47 (0)
APLES	1.61 (0.01)	1.63 (0.01)	2.46 (0)
Setting 7: U			
	MSE	TTB	MTB
True	1.55 (0)	1.55 (0)	2.45 (0)
MLO	1.58 (0)	1.67 (2.78)	2.46 (0)
MLE	1.63 (0)	1.67 (0.01)	2.46 (0)
LS	1.65 (0.01)	9.78 (1.79)	2.47 (0)
APLES	1.61 (0.01)	1.73 (0.03)	2.46 (0)

Table A.14. Setting 7

Setting 8: CE			
	MSE	TTB	MTB
True	2.33 (0)	2.61 (0)	2.4 (0)
MLO	2.4 (0.01)	2.61 (0.04)	2.41 (0)
MLE	2.58 (0.02)	2.8 (2.88)	2.41 (0)
LS	2.52 (0.03)	2.83 (0.09)	2.46 (0)
APLES	2.52 (0.03)	2.8 (0.1)	2.41 (0)
Setting 8: U			
	MSE	TTB	MTB
True	2.33 (0)	2.51 (0)	2.4 (0)
MLO	2.4 (0.01)	2.51 (0.03)	2.41 (0)
MLE	2.58 (0.02)	2.72 (0.07)	2.41 (0)
LS	2.52 (0.03)	2.62 (0.06)	2.46 (0)
APLES	2.52 (0.03)	3.17 (0.17)	2.41 (0)

Table A.15. Setting 8

Setting 9: CE			
	MSE	TTB	MTB
True	3.39 (0)	3.5 (0)	2.36 (0)
MLO	3.53 (0.01)	3.7 (0.03)	2.37 (0)
MLE	3.84 (0.07)	4.07 (0.11)	2.38 (0)
LS	4.16 (0.11)	4.74 (0.28)	2.47 (0)
APLES	3.63 (0.03)	6.3 (0.62)	2.37 (0)
Setting 9: U			
	MSE	TTB	MTB
True	3.39 (0)	3.5 (0)	2.36 (0)
MLO	3.53 (0.01)	3.72 (0.06)	2.37 (0)
MLE	3.83 (0.07)	4.28 (0.16)	2.38 (0)
LS	4.18 (0.11)	4.75 (0.24)	2.47 (0)
APLES	3.63 (0.03)	3.81 (0.06)	2.37 (0)

Table A.16. Setting 9

Setting 10: CE			
	MSE	TTB	MTB
True	2.79 (0)	2.88 (0)	2.29 (0)
MLO	2.94 (0.02)	3.16 (0.05)	2.29 (0)
MLE	3.26 (0.03)	4.85 (0.39)	2.29 (0)
LS	3.21 (0.07)	4.3 (0.29)	2.39 (0)
APLES	3.2 (0.03)	3.72 (0.17)	2.3 (0)
Setting 10: U			
	MSE	TTB	MTB
True	2.79 (0)	2.88 (0)	2.29 (0)
MLO	2.94 (0.02)	3.16 (0.05)	2.29 (0)
MLE	3.26 (0.03)	5.79 (0.6)	2.29 (0)
LS	3.22 (0.07)	5.67 (0.55)	2.39 (0)
APLES	3.19 (0.03)	4.87 (0.41)	2.3 (0)

Table A.17. Setting 10

Setting 11: CE			
	MSE	TTB	MTB
True	12.07 (0)	12.53 (0)	-0.26 (0)
MLO	12.08 (0)	12.53 (0.11)	-0.26 (0)
MLE	12.08 (0)	12.53 (0.04)	-0.26 (0)
LS	12.08 (0)	12.72 (0.06)	-0.26 (0)
APLES	12.07 (0)	12.57 (0.07)	-0.26 (0)
Setting 11: U			
	MSE	TTB	MTB
True	12.07 (0)	12.53 (0)	-0.26 (0)
MLO	12.07 (0)	12.53 (0.11)	-0.26 (0)
MLE	12.08 (0)	12.53 (0.04)	-0.26 (0)
LS	12.08 (0)	12.72 (0.06)	-0.26 (0)
APLES	12.07 (0)	12.93 (0.11)	-0.26 (0)

Table A.18. Setting 11

Setting 12: CE			
	MSE	TTB	MTB
True	11.67 (0)	12.2 (0)	-0.29 (0)
MLO	11.74 (0.01)	12.2 (0.07)	-0.28 (0)
MLE	11.74 (0.01)	12.27 (0.09)	-0.28 (0)
LS	11.74 (0.01)	12.35 (0.07)	-0.28 (0)
APLES	11.74 (0.01)	12.28 (0.09)	-0.28 (0)
Setting 12: U			
	MSE	TTB	MTB
True	11.67 (0)	12.27 (0)	-0.29 (0)
MLO	11.73 (0.01)	12.27 (0.07)	-0.28 (0)
MLE	11.74 (0.01)	12.27 (0.09)	-0.28 (0)
LS	11.73 (0.01)	12.35 (0.07)	-0.28 (0)
APLES	11.74 (0.01)	12.28 (0.09)	-0.28 (0)

Table A.19. Setting 12

Setting 13: CE			
	MSE	TTB	MTB
True	11.43 (0)	12.02 (0)	-0.31 (0)
MLO	11.46 (0.01)	12.02 (0.05)	-0.31 (0)
MLE	11.48 (0.01)	12.02 (0.03)	-0.31 (0)
LS	11.5 (0.01)	12.37 (0.06)	-0.3 (0)
APLES	11.47 (0.01)	12.05 (0.04)	-0.3 (0)
Setting 13: U			
	MSE	TTB	MTB
True	11.43 (0)	12.02 (0)	-0.31 (0)
MLO	11.46 (0.01)	12.02 (0.04)	-0.31 (0)
MLE	11.48 (0.01)	12.02 (0.03)	-0.31 (0)
LS	11.5 (0.01)	12.37 (0.06)	-0.3 (0)
APLES	11.47 (0.01)	12.05 (0.04)	-0.3 (0)

Table A.20. Setting 13

	MSE	TTB	MTB
Setting 14: CE			
True	11.28 (0)	11.6 (0)	-0.33 (0)
MLO	11.87 (0.01)	12.01 (0.03)	-0.33 (0)
MLE	11.87 (0.01)	12.01 (0.04)	-0.33 (0)
LS	11.87 (0.01)	12.16 (0.03)	-0.31 (0)
APLES	11.87 (0.01)	12.01 (0.04)	-0.33 (0)
Setting 14: U			
	MSE	TTB	MTB
True	11.28 (0)	11.6 (0)	-0.33 (0)
MLO	11.87 (0.01)	12.01 (0.03)	-0.33 (0)
MLE	11.87 (0.01)	12.01 (0.04)	-0.33 (0)
LS	11.87 (0.01)	12.16 (0.03)	-0.31 (0)
APLES	11.87 (0.01)	12.06 (0.03)	-0.33 (0)

Table A.21. Setting 14

	MSE	TTB	MTB
Setting 15: CE			
True	13.01 (0)	13.79 (0)	-0.42 (0)
MLO	13.04 (0.02)	13.79 (0.07)	-0.42 (0)
MLE	13.04 (0.01)	13.8 (0.04)	-0.42 (0)
LS	13.07 (0.01)	13.83 (0.13)	-0.37 (0)
APLES	13.08 (0.02)	13.79 (0.04)	-0.42 (0)
Setting 15: U			
	MSE	TTB	MTB
True	13.01 (0)	13.79 (0)	-0.42 (0)
MLO	13.04 (0.02)	13.79 (0.07)	-0.42 (0)
MLE	13.04 (0.01)	13.8 (0.08)	-0.42 (0)
LS	13.07 (0.01)	13.83 (0.13)	-0.37 (0)
APLES	13.08 (0.02)	13.79 (0.08)	-0.42 (0)

Table A.22. Setting 15

A.2.3 Distances from Predicted Optimal Factor Settings to True Optimal Factor Settings

The following tables (A.23 - A.37) we report the mean distance (L_2) between the true minimum and the predicted minimum (MLE, LS, APLES) within the region $-1 \leq x_i \leq 1$ over the 500 datasets generated. In parentheses, we report the standard error.

	Setting 1: CE		
MLO	0.15 (0.01)	0.67 (0.12)	0.07 (0.01)
ML	1.69 (0.17)	1.43 (0.14)	2.11 (0.21)
LS	0.83 (0.08)	0.9 (0.09)	2.15 (0.21)
APLES	0.76 (0.08)	0.67 (0.07)	1.72 (0.17)
	Setting 1: U		
	MSE	TTB	MTB
MLO	0.15 (0.01)	0.67 (0.12)	0.07 (0.01)
ML	1.66 (0.17)	1.43 (0.14)	2.12 (0.21)
LS	0.96 (0.1)	1.12 (0.11)	2.15 (0.22)
APLES	0.85 (0.08)	0.67 (0.07)	1.68 (0.17)

Table A.23. Setting 1

	Setting 2: CE		
MLO	0.07 (0.01)	0.07 (0.01)	0.04 (0)
ML	0.36 (0.04)	0.81 (0.08)	1.29 (0.13)
LS	2.03 (0.2)	1.86 (0.19)	2.14 (0.21)
APLES	0.38 (0.04)	1.01 (0.1)	0.05 (0)
	Setting 2: U		
	MSE	TTB	MTB
MLO	0.07 (0.01)	0.07 (0.01)	0.04 (0)
ML	0.36 (0.04)	0.8 (0.08)	1.29 (0.13)
LS	2.19 (0.22)	1.88 (0.19)	2.11 (0.21)
APLES	0.37 (0.04)	0.96 (0.1)	0.04 (0)

Table A.24. Setting 2

	Setting 3: CE		
MLO	0.08 (0.01)	0.08 (0.01)	0.05 (0.01)
ML	0.09 (0.01)	0.31 (0.03)	0.47 (0.05)
LS	1.2 (0.12)	1.23 (0.12)	2.72 (0.27)
APLES	0.08 (0.01)	0.33 (0.03)	0.47 (0.05)
	Setting 3: U		
	MSE	TTB	MTB
MLO	0.08 (0.01)	0.08 (0.01)	0.05 (0.01)
ML	0.09 (0.01)	0.31 (0.03)	0.47 (0.05)
LS	1.04 (0.1)	1.2 (0.12)	2.72 (0.27)
APLES	0.08 (0.01)	0.33 (0.03)	0.47 (0.05)

Table A.25. Setting 3

	Setting 4: CE		
MLO	0.05 (0)	0.05 (0)	0.04 (0)
ML	0.09 (0.01)	0.26 (0.03)	0.48 (0.05)
LS	1.77 (0.18)	1.67 (0.17)	2.41 (0.24)
APLES	0.09 (0.01)	0.39 (0.04)	0.1 (0.01)
	Setting 4: U		
	MSE	TTB	MTB
MLO	0.05 (0)	0.05 (0)	0.04 (0)
ML	0.09 (0.01)	0.35 (0.04)	0.48 (0.05)
LS	1.81 (0.18)	1.93 (0.19)	2.41 (0.24)
APLES	0.09 (0.01)	0.38 (0.04)	0.05 (0.01)

Table A.26. Setting 4

Setting 5: CE			
	MSE	TTB	MTB
MLO	0.79 (0.08)	0.95 (0.1)	2.03 (0.2)
ML	0.9 (0.09)	1.02 (0.1)	2.09 (0.21)
LS	0.93 (0.09)	1.17 (0.12)	2.27 (0.23)
APLES	0.89 (0.09)	1.32 (0.13)	2.04 (0.2)
Setting 5: U			
	MSE	TTB	MTB
MLO	0.79 (0.08)	0.95 (0.09)	2.03 (0.2)
ML	0.91 (0.09)	1.03 (0.1)	2.09 (0.21)
LS	1.15 (0.12)	1.23 (0.12)	2.29 (0.23)
APLES	0.9 (0.09)	1.33 (0.13)	2.04 (0.2)

Table A.27. Setting 5

Setting 6: CE			
	MSE	TTB	MTB
MLO	0.07 (0)	0.74 (0.08)	0.26 (0)
ML	0.16 (0.03)	1.55 (0.08)	0.58 (0.01)
LS	0.13 (0.07)	1.14 (0.08)	0.71 (0.01)
APLES	0.11 (0.02)	1.09 (0.07)	0.54 (0.01)
Setting 6: U			
	MSE	TTB	MTB
MLO	0.07 (0)	0.7 (0.08)	0.26 (0)
ML	0.16 (0.02)	1.75 (0.08)	0.59 (0.01)
LS	0.13 (0.06)	1.33 (0.08)	0.73 (0.01)
APLES	0.12 (0.02)	0.82 (0.07)	0.54 (0.01)

Table A.28. Setting 6

	Setting 7: CE		
	MSE	TTB	MTB
MLO	0.06 (0.01)	0.47 (0.07)	0.1 (0.03)
ML	0.16 (0.02)	0.78 (0.15)	0.42 (0.06)
LS	0.17 (0.01)	0.5 (0.11)	0.92 (0.07)
APLES	0.12 (0.01)	0.47 (0.11)	0.14 (0.05)
	Setting 7: U		
	MSE	TTB	MTB
MLO	0.06 (0.01)	0.51 (0.07)	0.11 (0.03)
ML	0.16 (0.02)	0.51 (0.17)	0.42 (0.06)
LS	0.17 (0.01)	0.99 (0.13)	1.07 (0.07)
APLES	0.12 (0.01)	0.77 (0.08)	0.27 (0.05)

Table A.29. Setting 7

	Setting 8: CE		
	MSE	TTB	MTB
MLO	0.14 (0.01)	0.94 (0.08)	0.1 (0.01)
ML	0.27 (0.02)	1.44 (0.08)	0.17 (0.04)
LS	0.26 (0.02)	1.48 (0.05)	1.07 (0.09)
APLES	0.22 (0.01)	1.18 (0.05)	0.16 (0.01)
	Setting 8: U		
	MSE	TTB	MTB
MLO	0.14 (0.01)	0.93 (0.12)	0.1 (0.01)
ML	0.27 (0.02)	1.18 (0.05)	0.17 (0.04)
LS	0.26 (0.02)	1.5 (0.1)	1.08 (0.11)
APLES	0.22 (0.01)	1.19 (0.08)	0.16 (0.03)

Table A.30. Setting 8

Setting 9: CE			
	MSE	TTB	MTB
MLO	0.17 (0.01)	0.28 (0.09)	0.13 (0.01)
ML	0.34 (0.03)	0.64 (0.14)	0.34 (0.02)
LS	0.35 (0.03)	0.88 (0.15)	1.63 (0.11)
APLES	0.23 (0.02)	1.31 (0.12)	0.3 (0.02)
Setting 9: U			
	MSE	TTB	MTB
MLO	0.17 (0.01)	0.51 (0.09)	0.13 (0.01)
ML	0.34 (0.03)	0.94 (0.12)	0.49 (0.02)
LS	0.35 (0.03)	0.88 (0.15)	1.64 (0.11)
APLES	0.23 (0.02)	0.69 (0.12)	0.31 (0.02)

Table A.31. Setting 9

Setting 10: CE			
	MSE	TTB	MTB
MLO	0.17 (0.02)	0.6 (0.03)	0.08 (0.01)
ML	0.4 (0.03)	1.02 (0.06)	0.14 (0.03)
LS	0.42 (0.03)	0.92 (0.09)	1.57 (0.16)
APLES	0.38 (0.02)	1.01 (0.13)	0.16 (0.03)
Setting 10: U			
	MSE	TTB	MTB
MLO	0.17 (0.02)	0.6 (0.05)	0.08 (0.01)
ML	0.4 (0.03)	1.28 (0.09)	0.14 (0.05)
LS	0.43 (0.04)	1.1 (0.09)	1.58 (0.16)
APLES	0.38 (0.02)	0.85 (0.07)	0.17 (0.03)

Table A.32. Setting 10

Setting 11: CE			
	MSE	TTB	MTB
MLO	0.01 (0.02)	0.3 (0.06)	0.05 (0.01)
ML	0.01 (0.04)	0.38 (0.1)	0.06 (0.01)
LS	0.01 (0.04)	0.3 (0.09)	0.05 (0.16)
APLES	0.01 (0.04)	0.38 (0.1)	0.05 (0.02)
Setting 11: U			
	MSE	TTB	MTB
MLO	0.01 (0.02)	0.3 (0.06)	0.05 (0.01)
ML	0.01 (0.04)	0.36 (0.13)	0.06 (0.01)
LS	0.01 (0.04)	0.3 (0.11)	0.05 (0.16)
APLES	0.01 (0.04)	0.32 (0.09)	0.05 (0.02)

Table A.33. Setting 11

Setting 12: CE			
	MSE	TTB	MTB
MLO	0.01 (0.01)	0.33 (0.11)	0.06 (0)
ML	0.01 (0.02)	0.33 (0.08)	0.07 (0.01)
LS	0.02 (0.03)	0.33 (0.08)	0.07 (0)
APLES	0.01 (0.02)	0.33 (0.11)	0.08 (0.01)
Setting 12: U			
	MSE	TTB	MTB
MLO	0.01 (0.01)	0.3 (0.11)	0.06 (0)
ML	0.01 (0.02)	0.33 (0.1)	0.07 (0.01)
LS	0.02 (0.03)	0.3 (0.09)	0.07 (0)
APLES	0.01 (0.02)	0.33 (0.07)	0.07 (0.01)

Table A.34. Setting 12

Setting 13: CE			
	MSE	TTB	MTB
MLO	0.08 (0)	0.29 (0.03)	0.05 (0)
ML	0.14 (0)	0.46 (0.04)	0.07 (0.01)
LS	0.08 (0)	0.29 (0.03)	0.07 (0.01)
APLES	0.14 (0)	0.46 (0.04)	0.07 (0.01)
Setting 13: U			
	MSE	TTB	MTB
MLO	0.08 (0)	0.29 (0.03)	0.05 (0)
ML	0.14 (0)	0.46 (0.04)	0.07 (0.01)
LS	0.08 (0)	0.29 (0.03)	0.07 (0.01)
APLES	0.14 (0)	0.46 (0.03)	0.07 (0.01)

Table A.35. Setting 13

Setting 14: CE			
	MSE	TTB	MTB
MLO	0.4 (0.01)	0.41 (0.04)	0.06 (0.01)
ML	0.4 (0)	0.42 (0.03)	0.08 (0.01)
LS	0.4 (0)	0.41 (0.03)	0.1 (0.01)
APLES	0.4 (0)	0.42 (0.03)	0.07 (0.01)
Setting 14: U			
	MSE	TTB	MTB
MLO	0.4 (0)	0.41 (0.04)	0.06 (0.01)
ML	0.4 (0)	0.42 (0.03)	0.08 (0.01)
LS	0.4 (0)	0.41 (0.03)	0.11 (0.01)
APLES	0.4 (0)	0.47 (0.03)	0.07 (0.01)

Table A.36. Setting 14

Setting 15: CE			
	MSE	TTB	MTB
MLO	0.08 (0.01)	0.18 (0.04)	0.05 (0.01)
ML	0.08 (0.01)	0.24 (0.05)	0.05 (0.01)
LS	0.09 (0.01)	0.24 (0.03)	0.11 (0.01)
APLES	0.15 (0.01)	0.24 (0.05)	0.05 (0.01)
Setting 15: U			
	MSE	TTB	MTB
MLO	0.08 (0.01)	0.18 (0.04)	0.05 (0.01)
ML	0.08 (0.01)	0.29 (0.05)	0.05 (0.01)
LS	0.09 (0.01)	0.19 (0.03)	0.12 (0.01)
APLES	0.15 (0.01)	0.29 (0.05)	0.05 (0.01)

Table A.37. Setting 15

Appendix B

R Code for APLES

We provide below R code which implements the cyclic coordinate descent (CCD) algorithm for APLES. In addition, it can also implement a CCD algorithm for the high-dimensional heteroscedastic regression (HHR) or the heteroscedastic iterative penalized pseudo-likelihood operator (HIPPO).

```
soft = function(x, l){xout=NULL
for (j in 1:length(x)){
xout = c(xout, sign(x[j])*max(0,(abs(x[j])-l)))}
xout}

asoft = function(x, l, bh){xout=NULL
for (j in 1:length(x)){
xout = c(xout, soft(x[j], l/(abs(bh[j]))) )}
xout}

scad = function(x, l){xout=NULL
for (j in 1:length(x)){
if (abs(x[j]) > 3.7*l)  xout = c(xout, x[j])
if (abs(x[j]) <= 3.7*l & abs(x[j]) > 2*l) xout = c(xout,
soft( x[j], 3.7*l/2.7 )/(1.7/2.7) )
if (abs(x[j]) <= 2*l) xout = c(xout, soft(x[j], l) ) }
xout}
```

```
pscad = function(x, l){xout=NULL
for (j in 1:length(x)){
if ( abs(x[j]) > 3.7*l)  xout = c(xout, (l^2*(3.7^2 - 1)))/(2*(2.7))
if ( abs(x[j]) <= 3.7*l & abs(x[j]) > l) xout = c(xout,
( 3.7*l*abs(x[j]) - 0.5*( abs(x[j])^2 + l^2 ) ) / 2.7)
if (abs(x[j]) <= l) xout = c( xout, abs(x[j])*l ) }
xout}

penobj1 = function(a, v, E, F, l1, l2){
r = l1*sum( abs(a[[1]]) ) + l2*sum(abs(a[[2]]))
ans=sum( ((v-E%*%a[[1]])^2)/exp(F%*%a[[2]]) ) +
sum( F%*%a[[2]] ) + r
ans}

penobj2 = function(a, v, E, F, l1, l2){
r = l1*sum( abs(a[[1]])/abs(betahat) ) + l2*sum(abs(a[[2]]))
ans=sum( ((v-E%*%a[[1]])^2)/exp(F%*%a[[2]]) ) +
sum( F%*%a[[2]] ) + r
ans}

penobj3 = function(a, v, E, F, l1, l2){
r = sum( pscad(a[[1]], l1) ) + sum( pscad(a[[2]], l2) )
ans=sum( ((v-E%*%a[[1]])^2)/exp(F%*%a[[2]]) ) +
sum( F%*%a[[2]] ) + r
ans}

f = function(lambda1, lambda2, dim1, dim2, betanew=rep(0,dim1),
taunew=rep(0,dim2), iter, type, A, B, betahat, tauhat){
ns=dim(A)[1]; l11=lambda1; l22=lambda2
lambda1=c(0,rep(lambda1,dim1-1)); lambda2=c(0,rep(lambda2,dim2-1));
#bh = lm(y~A[,-1])$coef; tauhat = rep(1, dim2)
for (i in 1:iter){myvec=(as.matrix(B)%*%taunew); w= as.vector(exp(-1*myvec));
for (j in 1:length(betanew)){hold = betanew;
temp=sum(w*as.matrix(A[,j])*(y-(as.matrix(A[,-j])%*%betanew[-j])))/
```

```
sum(w*(as.matrix(A[,j]))^2)
if (type==1){temp=soft(temp, lambda1[j])}
if (type==2){temp=asoft(temp, lambda1[j], betahat[j]+1e-8) }
if (type==3){temp=scad(temp, lambda1[j])}
hold[j]=temp ; betanew=hold}
d = (y-A%*%betanew)^2; d[d<0.005] = 0.005;
myvec=as.matrix(B)%*%taunew; ratio = d/(exp(myvec)); isave=NULL
if (type==1){
for (s in 0:10){ isave = c(isave, penobj1( list(betanew, taunew +
2^(-s)*solve(t(B)%*%B)%*%t(B)%*%(ratio-1) ), y, A, B, l11, l22 )) }}
if (type==2){
for (s in 0:10){ isave = c(isave, penobj2( list(betanew, taunew +
2^(-s)*solve(t(B)%*%B)%*%t(B)%*%(ratio-1) ), y, A, B, l11, l22))}}
if (type==3){
for (s in 0:10){ isave = c(isave, penobj3( list(betanew, taunew +
2^(-s)*solve(t(B)%*%B)%*%t(B)%*%(ratio-1) ), y, A, B, l11, l22 )) }}
s0 = which.min(isave); z0 = myvec + 2^(-(s0-1))*(ratio - 1);
if (dim(B)[2] > 1) {taunew = lm(z0~B[, -1])$coef}
if (dim(B)[2] <= 1) {taunew = lm(z0~1)$coef}
for (j in 1:length(taunew)){hold = taunew;
temp=sum(as.matrix(B[,j])*(z0-(as.matrix(B[,-j])%*%taunew[-j])))/
sum((as.matrix(B[,j]))^2)
if (type==1){temp=soft(temp, lambda2[j])}
if (type==2){temp=asoft(temp, lambda2[j], tauhat[j]+1e-8) }
if (type==3){temp=scad(temp, lambda2[j])}
hold[j]=temp; taunew=hold}
}
list(betanew, taunew)}



sesfun = function(ebeta, etau, abeta, atau, l1, l2, A, B){
abetaf=abeta; atauf=atau; signifbeta=rep(0,16); signiftau=rep(0,16)
A=as.matrix(A[,abeta!=0]); B=as.matrix(B[,atau!=0])
```

```
d1=length(abeta[abeta!=0]); d2=length(atau[atau!=0])
ebeta = ebeta[abeta!=0]; etau = etau[atau!=0]
abeta = abeta[abeta!=0]; atau = atau[atau!=0]

Sigbet = diag(1/(abs(ebeta*abeta)));
if (d2>1) Sigtau = diag((1/(abs(etau*atau))) )
if (d2<=1) Sigtau = (1/(abs(etau*atau)))

myvec=exp(as.matrix(B)%*%atau); Sigma = diag( as.vector(myvec) )

covbeta = solve(t(A) %*% solve(Sigma) %*% A + l1*Sigbet) %*%
(t(A) %*% solve(Sigma) %*% A)%*%solve(t(A) %*% solve(Sigma)%*%A + l1*Sigbet)

covtau = solve(0.5*t(B) %*% B + l2*Sigtau) %*%
(0.5*t(B) %*% B) %*%  solve(0.5*t(B) %*% solve(Sigma) %*% B + l2*Sigtau)

signifbeta = abeta/sqrt(diag(covbeta));
signiftau = atau/sqrt(diag(covtau))
abeta[abs(signifbeta)<2] = 0; atau[abs(signiftau)<2] = 0

abetaf[abetaf!=0] = abeta; atauf[atauf!=0] = atau

list(abetaf, atauf, sqrt(diag(covbeta)), sqrt(diag(covtau)),
signifbeta, signiftau)}
```

Bibliography

Aguirre-Torres, V. and de la Vara, R. (2011). A robust analysis of unreplicated factorials. *Applied Stochastic Models in Business and Industry*, 28:194–205.

Aitkin, M. (1987). Modeling variance heterogeneity in normal regression using glim. *Applied Statistics*, 36:332–339.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Apley, D. W. and Kim, J. (2011). A cautious approach to robust design with model parameter uncertainty. *IIE Transactions*, 43:471–482.

Bartlett, M. S. and Kendall, D. G. (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society Series B*, 8:128–150.

Bergman, B. and Hynen, A. (1997). Dispersion effects from unreplicated designs in the 2^{k-p} series. *Technometrics*, 39:191–198.

Berube, J. and Nair, V. N. (1998). Exploiting the inherent structure in robust parameter design experiments. *Statistica Sinica*, 8:43–66.

Bingham, D. and Nair, V. N. (2012). Noise variable settings in robust design experiments. *Technometrics*, 54(4):388–397.

Borror, C. M., Montgomery, D. C., and Myers, R. H. (2002). Evaluation of statistical designs for experiments involving noise variables. *Journal of Quality Technology*, 34:54–70.

Box, G. and Jones, S. (1990). Robust product designs, part i: first-order models with design x environment interactions. Technical report, Report.

Box, G. and Jones, S. (1992). Designing products that are robust to the environment. *Total Quality Management*, 3(3):265–282.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, 30:1–17.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model Building and Response Surfaces*. John Wiley and Sons, New York, NY.

Box, G. E. P. and Meyer, R. D. (1986a). An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18.

Box, G. E. P. and Meyer, R. D. (1986b). Dispersion effects from fractional designs. *Technometrics*, 28:19–27.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, 13:1–45.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253.

Brenneman, W. A. and Nair, V. N. (2001). Methods for identifying dispersion effects in unreplicated factorial experiments: A criticalanalysis and proposed strategies. *Technometrics*, 43(4):388–405.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Bursztyn, D. and Steinberg, D. M. (2006). Screening experiments for dispersion effects. In Dean, A. M. and Lewis, S. M., editors, *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, pages 21–47. Springer-Verlag, New York.

Charnes, A. and Cooper, W. W. (1963). Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, 11(1):18–39.

Cook, D. R. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70(1).

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82:1079–1091.

Daye, J. Z., Chen, J., and Li, H. (2012). High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics*, 68:316–326.

Del Castillo, E. (2007). *Process optimization: a statistical approach*, volume 105. Springer.

Del Castillo, E., Fan, S. K., and Semple, J. (1997). Computation of global optima in dual response systems. *Journal of Quality Technology*, 29:347–353.

Del Castillo, E. and Montgomery, R. C. (1993). A nonlinear programming solution to the dual response problem. *Journal of Quality Technology*, 25:199–204.

Dellino, G., Kleijnen, J. P. C., and Meloni, C. (2010). Robust optimization in simulation: Taguchi and response surface methodology. *International Journal of Production Economics*, 125:52–59.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Engel, J. (1992). Modeling variation in industrial experiments. *Applied Statistics*, 41:579–593.

Engel, J. and Huele, A. F. (1996a). A generalized linear modeling approach to robust design. *Technometrics*, 38:365–373.

Engel, J. and Huele, A. F. (1996b). Taguchi parameter design by second-order response surfaces. *Quality and Reliability Engineering International*, 12:95–100.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, J., Xue, L., Zou, H., et al. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849.

Frank, I. E. and Friedman, J. H. (1993). A statistical view ofsome chemometrics regression tools (with discussion). *Technometrics*, 35:109–148.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: A review with some new proposals. *Statistica Sinica*, 8:1–41.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, 44(3):461–465.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297307.

Kackar, R. N. and Shoemaker, A. C. (1986). Robust design: A cost-effective method for improving manufacturing processes. *AT & T Technical Journal,*, 65:39–50.

Kim, K. and Lin, D. K. J. (1998). Dual response surface optimization: A fuzzy modeling approach. *Journal of Quality Technology*, 30:1–10.

Koksoy, O. and Doganaksoy, N. (2003). Joint optimization of mean and standard deviation using response surface methods. *Journal of Quality Technology*, 35:239–252.

Kolar, M. and Sharpnack, J. (2012). Variance function estimation in high-dimensions. *Proceedings of the 29 th International Conference on Machine Learning,*.

Lee, Y. and Nelder, J. A. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistics*, 26:95–105.

Lee, Y. and Nelder, J. A. (2003). Robust design via generalized linear models. *Journal of Quality Technology*, 35:2–12.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, 31:469–473.

Leon, R. L., Shoemaker, A. C., and Kacker, R. N. (1987). Performance measures independent of adjustment. *Technometrics*, 29:253–265.

Li, R. and Lin, D. K. J. (2009). Variable selection for screening experiments. *Quality Technology & Quantitative Management*, 6(3):271–280.

Liao, C. T. (2000). Identification of dispersion effects from unreplicated 2^{n-k} fractional factorial designs. *Computational Statistics and Data Analysis*, 33:291–298.

Lin, D. K. J. and Tu, W. (1995). Dual response surface optimization. *Journal of Quality Technology*, 27:34–39.

Linhart, H. and Zucchini, W. (1986). *Subset Selection in Regression*. Wiley, New York, N.Y.

Loughin, T. M. and Malone, C. J. (2013). An adaptation of the bergman-hynn test for dispersion effects in unreplicated two-level factorial designs when the location model may be incorrect. *Journal of Quality Technology*, 45(4):350–359.

Mallows, C. (1973). Some comments on c_p. *Technometrics*, 15:661675.

Matsuura, S., Suzuki, H., Iida, T., Kurec, H., and H., M. (2011). Robust parameter design using a supersaturated design for a response surface model. *Quality and Reliability Engineering Internationa*, 27(4):541–554.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. London England Chapman and Hall 1983.

McGrath, R. N. and Lin, D. K. J. (2001a). Confounding of location and dispersion effects in unreplicated fractional factorials. *Journal of Quality Technology*, 33:129–139.

McGrath, R. N. and Lin, D. K. J. (2001b). Testing multiple dispersion effects in unreplicated fractional factorial designs. *Technometrics*, 43(4):406–414.

McGrath, R. N. and Lin, D. K. J. (2003). Analyzing location and dispersion in unreplicated fractional factorials. *Statistics & Probability Letters*, 65:369–377.

Miller, A. J. (2002). *Subset Selection in Regression*. Chapman and Hall, London, U.K.

Miro-Quesada, G. and Del Castillo, E. (2004). Two approaches for improving the dual response method in robust parameter design. *Journal of Quality Technology*, 36(2):154–168.

Myers, R. (1991). Response surface methodology in quality improvement. *Communications in Statistics-Theory and Methods*, 20(2):457–476.

Myers, R. H. and Carter, W. H. J. (1973). Response surface techniques for dual response systems. *Technometrics*, 15:301–317.

Myers, R. H., Khuri, A. I., and Vining, G. G. (1992). Response surface alternatives to the taguchi robust parameter design approach. *The American Statistician*, 46:131–139.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments (2nd ed.)*. Wiley, New York.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments (3rd ed.)*. Wiley, New York.

Nair, V. N., Abraham, B., MacKay, J., Box, G., Kacker, R. N., Lorenzen, T. J., Lucas, J. M., Myers, R. H., Vining, G. G., Nelder, J. A., et al. (1992). Taguchi's parameter design: a panel discussion. *Technometrics*, 34(2):127–161.

Nair, V. N. and Pregibon, D. (1988). Analyzing dispersion effects from replicated factorial experiments. *Technometrics*, 30:247257.

Nelder, J. A. and Lee, Y. (1991). Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, 7:107–120.

Nelder, J. A., Lee, Y., Bergman, A., Hynen, A., Huele, A. F., and Engel, J. (1998). Joint modeling of mean and dispersion. *Technometrics*, 40(2):168–175.

Pan, G. (1999). The impact of unidentified location effects on dispersion-effects identification from unreplicated factorial designs. *Technometrics*, 41:313–326.

Pan, G. and Taam, W. (2002). On generalized linear model method for detecting dispersion effects in unreplicated factorial designs. *Journal of Statistical Computation and Simulation*, 72(6):431–450.

Park, M. Y. and Hastie, T. (2007). l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B*, 69:659–677.

Park, R. E. (1968). Estimation with heteroscedastic error terms. *Econometrica*, 34:888.

Phoa, F. K. H. (2009). Journal of statistical planning and inference. *Analysis of Supersaturated Designs via the Dantzig Selector*, 139(7):2362–2372.

Pickle, S. M., Robinson, T. J., Birch, J. B., and Anderson-Cook, C. M. (2008). A semi-parametric approach to robust parameter design. *Journal of Statistical Planning and Inference*, 138:114–131.

Robinson, T. J., Borror, C. M., and Myers, R. H. (2004). Robust parameter design: A review. *Quality and Reliability Engineering International*, 20:81–101.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461464.

Shoemaker, A. C., Tsui, K. L., and Wu, C. F. J. (1991). Economical experimentation methods for robust parameter design. *Technometrics*, 33:415–427.

Simpson, W. T., Korte, J. J., Mauery, T. M., and Mistree, F. (1998). Comparison of response surface and kriging models for multidiscipinary design optimization. *AIAA Journal*, 98:1–11.

Steinberg, D. M. and Bursztyn, D. (1994). Dispersion effects in robust-design experiments with noise factors. *Journal of Quality Technology*, 26:16–20.

Taguchi, G. (1986). *Introduction to Quality Engineering*. UNIPUB/Kraus International, White Plains, NY.

Taguchi, G. (1987). *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*. UNIPUB/Kraus International, White Plains, NY.

Taguchi, G.and Wu, Y. (1980). *Quality Engineering Using Robust Design*. Central Japan Quality Control Association, Nagoya, Japan.

Tang, L. C. and Xu, K. (2002). A unified approach for dual response optimization. *Journal of Quality Technology*, 34:437–447.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288.

Vanli, O. A. and Del Castillo, E. (2009). Bayesian approaches for on-line robust parameter design. *IIE Transactions*, 41(4):359–371.

Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2).

Vining, G. G. and Bohn, L. L. (1998). Response surfaces for the mean and variance using a nonparametric approach. *Journal of Quality Technology*, 30(3):282–291.

Vining, G. G. and Myers, R. H. (1990). Combining taguchi and response surface philosophies: A dual response approach. *Journal of Quality Technology*, 22:38–45.

Voss, D. T. and Wang, W. (2006). Analysis of orthogonal saturated designs. In Dean, A. M. and Lewis, S. M., editors, *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, pages 268–286. Springer-Verlag, New York.

Wagener, J. and Dette, H. (2012). Bridge estimators and the adaptive lasso under heteroscedasticity. *Mathematical Methods of Statistics*, 21(2):109–126.

Wang, P. C. (1989). Tests for dispersion effects from orthogonal arrays. *Computational Statistics and Data Analysis*, 8:109–117.

Welch, W. J., Yu, T. K., Kang, S. M., and Sacks, J. (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology*, 22:15–22.

Wu, C. F. J. and Hamada, M. S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

Wu, L. and Huiqiong, L. (2012). Variable selection for joint mean and dispersion models of the inverse gaussian distribution. *Metrica*, 75:795–808.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Xu, D. and Zhang, Z. (2011). Regularized reml for estimation in heteroscedastic regression models. In Li, S., Wang, X., Okazaki, Y., Kawabe, J., Murofushi, T., and Guan, L., editors, *Nonlinear Mathematics for Uncertainty and its Applications*, pages 495–502. Springer-Verlag,Berlin Heidelberg.

Yang, Y. (2005). Can the strengths of aic and bic be shared? *Biometrika*, 92(4):937–950.

Yuan, M., Joseph, V. R., and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4):430–439.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Vita

Kwame A. Kankam

Education

December 2014

Ph.D. in Statistics, The Pennsylvania State University, University Park, PA
Advisor: Prof. James L. Rosenberger

June 2005

BSc. in Mechanical Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Research Interests

Design and Analysis of Experiments; Quality and Productivity Statistics; High-Dimensional Data Analysis

Teaching Experience

May 2011 - December 2014

Instructor for Stat 100, 200, 301, 401, 414, The Pennsylvania State University, University Park, PA

Work Experience

October 2006 - June 2009

Senior Coordinator, Airfreight. Maersk Logistics (Ghana), Accra, Ghana

Internship Experience

June 2013 - August 2013

Statistics Intern at EMD Serono Research and Development Institute, Inc., Billerica, Ma

Statistical Software

R, SAS, SPSS, Minitab

Affiliations

Member of the American Statistical Association (ASA).

Member of the American Society for Quality (ASQ).