University of California

Los Angeles

Analysis and Construction of Nonregular Fractional Factorial Designs

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Frederick Kin Hing Phoa

2009

The dissertation of Frederick Kin Hing Phoa is approved.

Rick Paik Schoenberg

Yingnian Wu

Weng Kee Wong

Hongquan Xu, Committee Chair

University of California, Los Angeles

2009

To my family and my love

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply appreciated from my heart to many people who have played important roles in the completion of my Ph.D. program.

I would like to express my gratitude to my committee members Professors Hongquan Xu, Weng Kee Wong, Yingnian Wu and Rick Schoenberg, for their concerns and supports. Special thanks go to Professors Xu and Wong for their excellent mentorships, fruitful discussions and constructive suggestions in my research. In additon, Professor Xu's positive encouragement keeps pushing me to work hard.

Sincere gratitude is also extended to my teachers: Mr. Lo Man-Chuen (Kwun Tong Maryknoll College), Dr. Emily A. Carter (Princeton), Dr. Jamey Anderson (Santa Monica College) and Dr. Yung-Ya Lin (UCLA). All of them provided me special inspirations when I was still their student. These inspirations eventually led to my choice to enter graduate school, and taught me how to be a good researcher. Especially, Mr. Lo at Kwun Tong Maryknoll College was the first teacher to inspire me to choose my own way and reach my own career goal. I want to thank him very much.

Last but not least, I would like to thank all of my friends and family. For my good friends, especially Elaine Lam, Kathy Kam and Annie Wong, thank them for their prays and unconditional supports. They are very special friends to me. For my little brother, Gilbert, who always concerns me and provides me instant helps whenever necessary. For my parents, who for 27 years have provided me

every opportunity to make this PhD a reality. And finally for my love, Peggy Pan, who is always my source of strength, support and happiness, I cannot express how blessed I am to be with you.

To God be all glory and honor.

Chapter Two is a version of Xu, Phoa and Wong (2008) accepted by *Statistics Surveys*. Chapter Three is a version of Phoa, Wong and Xu (2009) submitted to *Journal of Chemometrics*. Chapter Four is a version of Phoa, Pan and Xu (2008) accepted by *Journal of Statistical Planning and Inference*. Chapter Five is a version of Phoa and Xu (2008) accepted by *Annals of Statistics*.

Vita

1981	Born in Hong Kong.
1999	High School Certificate, Kwun Tong Maryknoll College, Hong Kong.
2000	A.A. (Physical Science), Santa Monica College, Santa Monica, California.
2002	B.S. (Physical Chemistry), UCLA, Los Angeles, California.
2002	B.S. (Applied Mathematics), UCLA, Los Angeles, California.
2002–2005	Teaching Assistant, Chemistry Department, UCLA. Research Assistant, Chemistry Department, UCLA, under direction of Professor Yung-Ya Lin.
2006	M.S. (Statistics), UCLA, Los Angeles, California
2005–2008	Teaching Assistant, Statistics Department, UCLA.
2005–2008	Research Assistant, Statistics Department, UCLA, under direction of Professor Hongquan Xu.
2007–2008	Research Assistant, Biostatistics Department, UCLA, under direction of Professor Weng Kee Wong.

Publications and Presentations

Phoa, F.K.H. and Xu, H. (2008) Quarter-Fraction Factorial Design Constructed via Quaternary Codes. Accepted by *Annals of Statistics*.

Phoa, F.K.H., Pan, Y.H. and Xu, H. (2008) Analysis of Supersaturated Designs via the Dantzig Selector. Accepted by *Journal of Statistical Planning and Inference*.

Walls, J.D., Phoa, F.K.H. and Lin, Y.Y. (2004) Spin Dynamics at very High Spin Polarization. *Physical Review B*, 70, 174410.

Xu, H., Phoa, F.K.H. and Wong, W.K. (2008) Recent Developments in Nonregular Fractional Factorial Designs. Accepted by *Statistics Surveys*.

Applications of using quaternary codes to nonregular designs I - Resolution and Aberration of $2^{2(n-1)}$ Designs. *Joint Research Conference on Statistics in Quality, Industry and Technology*, Knoxville, Tennessee, June 2006.

Applications of using quaternary codes to nonregular designs II - Resolution and Aberration of $2^{2(n-2)}$ Designs. *International Conference on Design of Experiments and Its Applications*, Tianjin, China, July 2006.

Some results on the projectivity of two-level nonregular designs constructed from quaternary codes. *Joint Statistical Meeting*, Salt Lake City, Utah, July 2007.

Analysis of the Supersaturated Designs using Dantzig Selector. *Design and Analysis of Experiments Conference*, Memphis, Tennessee, November 2007.

Analysis and Construction of Nonregular Fractional Factorial Designs. *Seminar*, Department of Statistics, National Tsing-Hua University, Taiwan, February 2009.

Analysis and Construction of Nonregular Fractional Factorial Designs. *Seminar*, Institute of Statistical Science, Academia Sinica, Taiwan, February 2009.

Abstract of the Dissertation

Analysis and Construction of Nonregular Fractional Factorial Designs

by

Frederick Kin Hing Phoa

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2009

Professor Hongquan Xu, Chair

Nonregular fractional factorial designs (FFDs) are widely used in various screening experiments for their run size economy and flexibility. In the beginning of this thesis, a brief review is given on the development of nonregular FFDs since the groundbreaking work by Hamada and Wu (1992). Then this thesis focuses on two major directions of using nonregular FFDs: analysis and construction. Some real-life examples are provided to show the correct ways to analyze nonregular FFDs, so that three potential pitfalls can be avoided. In addition, the Dantzig selector method is introduced for identifying active factors in both nonregular FFDs and supersaturated designs. Real-life examples and simulations show how efficient the Dantzig selector is when it is compared to the existing method in the literature. The quaternary code method is introduced for the construction of nonregular FFDs. The properties and uses of quaternary codes toward the construction of nonregular FFDs are explored, and optimal designs are constructed under some common criteria.

CHAPTER 1

An Introduction

There is considerable scope for reducing resources used in research by designing more efficient studies. Giles (2006) in a foreword in a recent issue in the journal Nature observed that some toxicology studies seemed to lack sophisticated thinking in their designs and wondered whether that had led to many inconclusive studies. The importance of a well designed study cannot be over-emphasized. Experiments are increasingly complex, in addition to rising experimental cost and competing resources. In the extreme case, a poorly-designed study may not be able to answer the posited scientific hypotheses. Careful design considerations even with only minor variation in traditional designs can lead to a more efficient study in terms of more precise estimates or able to estimate more effects in the study at the same cost.

In many scientific investigations, the main interest is in the study of effects of many factors simultaneously. Factorial designs, especially two-level or three-level factorial designs, are the most commonly used experimental plans for this type of investigation. They can be used to detect interactions between two or more factors in an experiment. Such designs were suggested by the US Environmental Protection Agency as one valuable statistical approach for risk assessment of chemical mixtures (Svensgaard and Hertzberg 1994). A full factorial experiment allows all factorial effects to be estimated independently and is commonly used in practice. However, it is often too costly to perform a full factorial experiment.

For example, if we have 8 factors to investigate and each factor has two levels, we need to have $2^8 = 256$ runs. Instead, a fractional factorial design, which is a subset or fraction of a full factorial design, is often preferred because much fewer runs are required. When this fraction is properly selected, the resulting design has optimal properties.

Fractional factorial designs (FFDs) are classified into two broad types: *regular* FFDs and *nonregular* FFDs. Regular FFDs are constructed through defining relations among factors and are described in many textbooks such as Box, Hunter and Hunter (2005), Dean and Voss (1999), Montgomery (2005) and Wu and Hamada (2000). These designs have a simple aliasing structure in that any two effects are either orthogonal or fully aliased. The run sizes are always a power of 2, 3 or a prime, and thus the "gaps" between possible run sizes are getting wider as the power increases. The concept of *resolution* (Box and Hunter 1961) and its refinement *minimum aberration* (Fries and Hunter 1980) play a pivotal role in the optimal choice of regular FFDs. There are many recent developments on minimum aberration designs; see Wu and Hamada (2000) and Mukerjee and Wu (2006) for further references.

Nonregular FFDs such as Plackett-Burman designs and other orthogonal arrays are widely used in various screening experiments for their run size economy and flexibility (Wu and Hamada, 2000). They fill the gaps between regular FFDs in terms of various run sizes and are flexible in accommodating various combinations of factors with different numbers of levels. Unlike regular FFDs, nonregular FFDs may exhibit a complex aliasing structure, that is, a large number of effects may be neither orthogonal nor fully aliased, which makes it difficult to interpret their significance. For this reason, nonregular FFDs were traditionally used to estimate factor main effects only but not their interactions. However, in many

practical situations it is often questionable whether the interaction effects are negligible.

Hamada and Wu (1992) went beyond the traditional approach and proposed an analysis strategy to demonstrate that some interactions could be entertained and estimated through their complex aliasing structure. They pointed out that ignoring interactions can lead to (i) important effects being missed, (ii) spurious effects being detected, and (iii) estimated effects having reversed signs resulting in incorrectly recommended factor levels. Much of the recent studies in nonregular FFDs were motivated from results in Hamada and Wu (1992). They included proposal of new optimality criteria, construction and analysis of nonregular FFDs.

This thesis is organized as follows. In Chapter 2, I briefly review major developments in nonregular FFDs since 1992. Chapter 3 suggests a correct approach of using nonregular FFDs, so that three potential pitfalls are avoided. Three real-life experiments in chemometrics are used for demonstrations. The results show that better and more efficient models are suggested by the correct analyses. In Chapter 4, the Dantzig selector method is introduced for analyzing nonregular and supersaturated designs. A graphical procedure using a *profile plot* and an automatic variable selection procedure, under a new criterion modified from traditional AIC, are suggested. Three real-life experiments and some simulations show how efficient the Dantzig selector method is when it is compared to existing methods in the literature. In Chapters 5 and 6, the quaternary code method is introduced for constructing nonregular FFDs with good properties. The properties and uses of quaternary codes toward the construction of quarter-fraction and $(1/16)^{th}$-fraction nonregular FFDs are explored. Optimal designs are constructed under some common criteria. Chapter 7 gives a conclusion.

CHAPTER 2

A Brief History of Nonregular FFDs

The primary aim of this chapter is to review major developments in nonregular FFDs since 1992.

Here is a brief history of the major developments in nonregular FFDs. Plackett and Burman (1946) gave a large collection of two-level and three-level designs for multi-factorial experiments. These designs are often referred to as the Plackett-Burman designs in the literature. Rao (1947) introduced the concept of orthogonal arrays, including Plackett-Burman designs as special cases. Cheng (1980) showed that orthogonal arrays are universally optimal for main effects model. Hamada and Wu (1992) successfully demonstrated that some interactions could be identified beyond a few significant main effects for Plackett-Burman designs and other orthogonal arrays. Lin and Draper (1992) studied the geometrical projection properties of Plackett-Burman designs while Wang and Wu (1995) and Cheng (1995, 1998) studied the hidden projection properties of Plackett-Burman designs and other orthogonal arrays. The hidden projection properties provide an explanation for the success of the analysis strategy due to Hamada and Wu (1992). Sun and Wu (1993) were the first to coin the term "nonregular FFDs" when studying statistical properties of Hadamard matrices of order 16. Deng and Tang (1999) and Tang and Deng (1999) introduced the concepts of generalized resolution and generalized minimum aberration for two-level nonregular FFDs. Xu and Wu (2001) proposed the generalized minimum aberration for mixed-level

nonregular FFDs. Because of the popularity of minimum aberration, the research on nonregular FFDs has been largely focused on the construction and properties of generalized minimum aberration designs. Our reference list suggests that keen interest in nonregular FFDs began in 1999 and continues to this day as evident by the increasing number of scientific papers on nonregular FFDs in major statistical journals.

2.1 Projection Properties of Plackett-Burman Designs and Other Orthogonal Arrays

We first review the concept of orthogonal arrays due to Rao (1947). An orthogonal array of N runs, m factors, s levels and strength t, denoted by $OA(N, s^m, t)$, is an $N \times m$ matrix in which each column has s symbols or levels and for any t columns all possible s^t combinations of symbols appear equally often in the matrix. Rao (1973) generalized the definition to the asymmetrical case where an orthogonal array is allowed to have variable numbers of symbols, i.e., mixed levels. For example, the 12-run Plackett-Burman design in Table 2.1 is an $OA(12, 2^{11}, 2)$ and the 18-run design in Table 2.2 is an $OA(18, 2^1 3^7, 2)$. Hedayat, Sloane and Stufken (1999) gave a comprehensive description on various aspects of orthogonal arrays.

Plackett-Burman designs are saturated orthogonal arrays of strength two because all degrees of freedom are utilized to estimate main effects. Orthogonal arrays of strength two allow all the main effects to be estimated independently and they are universally optimal for the main effects model (Cheng 1980). A necessary condition for the existence of an $OA(N, s^m, 2)$ is that $N - 1 \geq m(s-1)$. A design is called saturated if $N - 1 = m(s-1)$ and supersaturated if $N - 1 < m(s-1)$.

Table 2.1: 12-run Orthogonal Array (Plackett-Burman Design)

Run	A	B	C	D	E	F	G	8	9	10	11
1	+	+	−	+	+	+	−	−	−	+	−
2	+	−	+	+	+	−	−	−	+	−	+
3	−	+	+	+	−	−	−	+	−	+	+
4	+	+	+	−	−	−	+	−	+	+	−
5	+	+	−	−	−	+	−	+	+	−	+
6	+	−	−	−	+	−	+	+	−	+	+
7	−	−	−	+	−	+	+	−	+	+	+
8	−	−	+	−	+	+	−	+	+	+	−
9	−	+	−	+	+	−	+	+	+	−	−
10	+	−	+	+	−	+	+	+	−	−	−
11	−	+	+	−	+	+	+	−	−	−	+
12	−	−	−	−	−	−	−	−	−	−	−

Table 2.2: 18-run Orthogonal Array

Run	A	B	C	D	E	F	G	H
1	0	0	0	0	0	0	0	0
2	0	0	1	1	1	1	1	1
3	0	0	2	2	2	2	2	2
4	0	1	0	0	1	1	2	2
5	0	1	1	1	2	2	0	0
6	0	1	2	2	0	0	1	1
7	0	2	0	1	0	2	1	2
8	0	2	1	2	1	0	2	0
9	0	2	2	0	2	1	0	1
10	1	0	0	2	2	1	1	0
11	1	0	1	0	0	2	2	1
12	1	0	2	1	1	0	0	2
13	1	1	0	1	2	0	2	1
14	1	1	1	2	0	1	0	2
15	1	1	2	0	1	2	1	0
16	1	2	0	2	1	2	0	1
17	1	2	1	0	2	0	1	2
18	1	2	2	1	0	1	2	0

In the literature, orthogonal arrays of strength two are often called orthogonal designs or orthogonal arrays without mentioning the strength explicitly.

Orthogonal arrays include both regular and nonregular FFDs. For regular FFDs, the concepts of strength and resolution are equivalent because a regular FFD of resolution R is an orthogonal array of strength $t = R - 1$. For a regular FFD of resolution R, the projection onto any R factors must be either a full factorial or copies of a half-replicate of a full factorial. The projection for nonregular FFDs is more complicated.

Plackett-Burman designs are of strength two so that the projection onto any two factors is a full factorial. Lin and Draper (1992) studied the geometrical projection properties of the Plackett-Burman designs onto three or more factors. Their computer searches found all the projections of 12-, 16-, 20-, 24-, 28-, 32- and 36-run Plackett-Burman designs onto three factors. They found that these projections must have at least a copy of the full 2^3 factorial or at least a copy of a 2^{3-1} replicate or both. In particular, any projection onto three factors must contain a copy of a full factorial except for the 16- and 32-run Plackett-Burman designs, which are regular FFDs. The important statistical implication of this finding is that if only at most three factors are truly important, then after identifying the active factors, all factorial effects among these active factors are estimable, regardless which three factors are important.

Box and Tyssedal (1996) defined a design to be of projectivity p if the projection onto every subset of p factors contains a full factorial design, possibly with some points replicated. It follows from these definitions that an orthogonal array of strength t is of projectivity t. Cheng (1995) showed that, as long as the run size N is not a multiple of 2^{t+1}, an $OA(N, 2^m, t)$ with $m \geq t + 2$ has projectivity $t + 1$, even though the strength is only t.

The 12-run Plackett-Burman design given in Table 2.1 is of projectivity three but not of projectivity four. Wang and Wu (1995) found that its projection onto any four factors has the property that all the main effects and two-factor interactions can be estimated if the higher-order interactions are negligible. They referred this estimability of interactions without relying on geometric projection to as having a *hidden projection* property.

More generally, Wang and Wu (1995) defined a design as having a hidden projection property if it allows some or all interactions to be estimated even when the projected design does not have the right resolution or other geometrical/combinatorial design property for the same interactions to be estimated. For the Plackett-Burman designs their hidden projection property is a result of complex aliasing between the interactions and the main effects. For example, in the 12-run Plackett-Burman design given in Table 2.1, any two-factor interaction, say AB, is orthogonal to the main effects A and B, and partially aliased with all other main effects with correlation $1/3$ or $-1/3$. Because no two-factor interaction is fully aliased with any main effects, it is possible to estimate four main effects and all six two-factor interactions among them together.

The general results on hidden projection properties were obtained by Cheng (1995, 1998) and Bulutoglu and Cheng (2003). Cheng (1995) showed that as long as the run size N of an $OA(N, 2^m, 2)$ is not a multiple of 8, its projection onto any four factors allows the estimation of all the main effects and two-factor interactions when the higher-order interactions are negligible. Bulutoglu and Cheng (2003) showed that the same hidden projection property also holds for Paley designs [constructed by a method due to Paley (1933)] of sizes greater than 8, even when their run sizes are multiples of 8. A key result is that such designs do not have defining words of length three or four. Cheng (1998) further showed

that as long as the run size N of an $OA(N, 2^m, 3)$ is not a multiple of 16, its projection onto any five factors allows the estimation of all the main effects and two-factor interactions. Cheng (2006) gave a nice review of projection properties of factorial designs and their role in factor screening.

A few papers studied projection properties of designs with more than two levels. Wang and Wu (1995) studied the hidden projections onto 3 and 4 factors of the popular $OA(18, 3^7, 2)$ given in Table 2.2 (columns B–H). Cheng and Wu (2001) further studied the projection properties of this $OA(18, 3^7, 2)$ and an $OA(36, 3^{12}, 2)$ in terms of their two-stage analysis strategy. They constructed a nonregular $OA(27, 3^8, 2)$ that allows the second-order model to be estimated in all four-factor projections. In contrast, any regular 27-run design with eight 3-level factors does not have this four-factor projection property. They concluded that three-level nonregular FFDs have better projection properties and are more useful than regular FFDs for the dual purposes of factor screening and response surface exploration. Xu, Cheng and Wu (2004) further explored the projection properties of 18-run and 27-run orthogonal arrays and constructed a nonregular $OA(27, 3^{13}, 2)$ that allows the second-order model to be estimated in all of the five-factor projections. Tsai et al. (2000, 2004) and Evangelaras et al. (2005, 2007) also studied projection properties of three-level orthogonal arrays. Dey (2005) studied projectivity properties of asymmetrical orthogonal arrays with all except one factors having two levels.

2.2 Generalized Resolution and Generalized Minimum Aberration

Prior to 1999, an outstanding problem was how to assess, compare and rank nonregular FFDs in a systematic fashion. Deng and Tang (1999) and Tang and Deng (1999) were the first to propose generalized resolution and generalized minimum aberration criteria for 2-level nonregular FFDs, which are natural generalizations of the traditional concepts of resolution and minimum aberration for regular FFDs.

2.2.1 Introduction and Definition

To define these two important concepts, generalized resolution and generalized minimum aberration, Deng and Tang (1999) and Tang and Deng (1999) introduced the important notion of J-characteristics and is defined as follows. Given a two-level $N \times m$ design $D = (d_{ij})$, for $s = \{c_1, \ldots, c_k\}$, a subset of k columns of D, define

$$j_k(s; D) = \sum_{i=1}^{N} c_{i1} \cdots c_{ik} \text{ and } J_k(s; D) = |j_k(s; D)|, \tag{2.1}$$

where c_{ij} is the ith component of column c_j. The quantity $j_k(s; D)/N$ can be viewed as an extension of correlation. For illustration, consider the 12-run Plackett-Burman design given in Table 2.1. For $s = \{A, B\}$, $j_2(s; D) = 0$ since A and B are orthogonal. For $s = \{A, B, C\}$, $j_3(s; D)/N = -1/3$ is the correlation between main effect A and two-factor interaction BC. For $s = \{A, B, C, D\}$, $j_4(s; D)/N = -1/3$ is the correlation between two two-factor interactions, say AB and CD. The quantity $\rho_k(s) = J_k(s; D)/N$ is called the normalized J-characteristics by Tang and Deng (1999) or aliasing index by Cheng, Li and Ye (2004) and Phoa and Xu (2008) because $0 \leq \rho_k(s) \leq 1$. It is not difficult to see

that if D is a two-level regular FFD then $\rho_k(s) = 0$ or 1 for all s. Ye (2004) showed that the reserve is also true. Therefore, for a nonregular FFD, there always exist some s such that $0 < \rho_k(s) < 1$.

Suppose that r is the smallest integer such that $\max_{|s|=r} J_r(s; D) > 0$, where the maximization is over all subsets of r columns. Then the *generalized resolution* is defined to be

$$R = r + \delta, \text{ where } \delta = 1 - \max_{|s|=r} \frac{J_r(s; D)}{N}. \tag{2.2}$$

For the 12-run design in Table 2.1, $r = 3$, $\delta = 2/3$ and the generalized resolution is $R = 3.67$. It is easy to see that for an $OA(N, 2^m, t)$, $j_k(s; D) = 0$ for any $k \leq t$ and therefore $r \leq R < r + 1$ where $r = t + 1$. If $\delta > 0$, a subset s of D with r columns contains at least $N\delta/2^r$ copies of a full 2^r factorial and therefore the projectivity of D is at least r (Deng and Tang 1999). For a regular FFD, $\delta = 0$ and the projectivity is exactly $r - 1$.

Two regular FFDs of the same resolution can be distinguished using the minimum aberration criterion, and the same idea can be applied to nonregular FFDs using the *minimum G-aberration* criterion (Deng and Tang 1999). Roughly speaking, the minimum G-aberration criterion always chooses a design with the smallest confounding frequency among designs with maximum generalized resolution. Formally, the minimum G-aberration criterion is to sequentially minimize the components in the confounding frequency vector

$$\text{CFV}(D) = [(f_{11}, \ldots, f_{1N}); (f_{21}, \ldots, f_{2N}); \ldots; (f_{m1}, \ldots, f_{mN})],$$

where f_{kj} denotes the frequency of k-column combinations s with $J_k(s; D) = N + 1 - j$.

Minimum G-aberration is very stringent and it attempts to control J-characteristics in a very strict manner. Tang and Deng (1999) proposed a re-

laxed version of minimum G-aberration and called it the *minimum G_2-aberration* criterion. Let

$$A_k(D) = N^{-2} \sum_{|s|=k} J_k^2(s; D). \tag{2.3}$$

The vector $(A_1(D), \ldots, A_m(D))$ is called the *generalized wordlength pattern*, because for a regular FFD D, $A_k(D)$ is the number of words of length k in the defining contrast subgroup of D. The *minimum G_2-aberration* criterion (Tang and Deng 1999) is to sequentially minimize the generalized wordlength pattern $A_1(D), A_2(D), \ldots, A_m(D)$.

For regular FFDs both minimum G-aberration and minimum G_2-aberration criteria reduce to the traditional minimum aberration criterion. However, these two criteria can result in selecting different nonregular FFDs. We note that minimum G-aberration nonregular FFDs always have maximum generalized resolution whereas minimum G_2-aberration nonregular FFDs may not. This is in contrast to regular case where minimum aberration regular FFDs always have maximum resolution among all regular FFDs.

Tang and Deng (1999) also defined minimum G_e-aberration for any $e > 0$ by replacing $J_k^2(s; D)$ with $J_k^e(s; D)$ in Equation (2.3). However, only the minimum G_2-aberration criterion is popular due to various statistical justifications and theoretical results.

Xu and Wu (2001) proposed the *generalized minimum aberration* criterion for comparing asymmetrical (or mixed-level) designs. The generalized minimum aberration criterion was motivated from ANOVA models and includes the minimum G_2-aberration criterion as a special case. By exploring an important connection between design theory and coding theory, Xu and Wu (2001) showed that the generalized wordlength pattern defined in Equation (2.3) are linear combinations of the distribution of pairwise distance between the rows. This observation plays

a pivotal role in the subsequent theoretical development of nonregular FFDs.

Ma and Fang (2001) independently extended the minimum G_2-aberration criterion for designs with more than two levels. They named their criterion as the *minimum generalized aberration* criterion, which is a special case of the generalized minimum aberration criterion proposed by Xu and Wu (2001).

Ye (2003) redefined the generalized wordlength pattern and generalized minimum aberration for two-level designs using indicator functions. Cheng and Ye (2004) defined generalized resolution and generalized minimum aberration criterion for quantitative factors. The generalized minimum aberration criterion proposed by Xu and Wu (2001) is independent of the choice of treatment contrasts and thus model-free whereas the generalized minimum aberration criterion by Cheng and Ye (2004) depends on the specific model.

2.2.2 Statistical Justification

Deng and Tang (1999) provided a statistical justification for the generalized resolution by showing that designs with maximum generalized resolution minimize the contamination of non-negligible two-factor interactions on the estimation of main effects. Tang and Deng (1999) provided a similar statistical justification for minimum G_2-aberration designs. In a further extension, Xu and Wu (2001) gave a statistical justification for generalized minimum aberration designs with mixed levels.

A common situation that arises in practice is that the main effects are of primary interest but there are uninteresting yet non-negligible interactions that we know will affect the main effects estimates. To fix ideas, consider a two-level $N \times m$ design $D = (d_{ij})$ with columns denoted by $d_1, \ldots, d_m$ and generalized

resolution between 3 and 4. Suppose that one fits a main effects model

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j d_{ij} + \epsilon_i, \tag{2.4}$$

but the true model is

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j d_{ij} + \sum_{k<l}^{m} \beta_{kl} d_{ik} d_{il} + \epsilon_i. \tag{2.5}$$

The least squares estimator $\hat{\beta}_j$ of β_j from the working model Equation (2.4), under the true model Equation (2.5), has expectation given by

$$E(\hat{\beta}_j) = \beta_j + N^{-1} \sum_{k<l}^{m} j_3(d_j, d_k, d_l)\beta_{kl}$$

for $j = 1, \ldots, m$, where $j_3(d_j, d_k, d_l)$ is defined in Equation (2.1). There are many ways to minimize the biases in estimating main effects due to the presence of the interaction effects. A conservative approach is minimizing the maximum bias, $\max_{j<k<l} J_3(d_j, d_k, d_l)$. This is equivalent to maximizing the generalized resolution as defined in Equation (2.2). Therefore, designs with maximum generalized resolution minimize the maximum bias of non-negligible interactions on the estimation of the main effects. A more aggressive approach is minimizing the sum of squared coefficients $\sum_{j=1}^{m} \sum_{k<l}^{m} [j_3(d_j, d_k, d_l)/N]^2 = 3A_3(D)$, where $A_3(D)$ is defined in Equation (2.3). Hence minimum G_2-aberration designs minimize the overall contamination of non-negligible interactions on the estimation of the main effects.

For regular FFDs, Cheng, Steinberg and Sun (1999) justified the minimum aberration criterion by showing that it is a good surrogate for some model-robustness criteria. Following their approach, Cheng, Deng and Tang (2002) considered the situation where (i) the main effects are of primary interest and their estimates are required and (ii) the experimenter would like to have as much

information about two-factor interactions as possible, under the assumption that higher-order interactions are negligible. Without knowing which two-factor interactions are significant, they considered the set of models containing all of the main effects and f two-factor interactions for $f = 1, 2, 3, \ldots$. Let E_f be the number of estimable models and D_f be the average of D-efficiencies. Cheng, Deng and Tang (2002) showed that the minimum G_2-aberration designs tend to have large E_f and D_f values, especially for small f; therefore, the minimum G_2-aberration criterion provides a good surrogate for the traditional model-dependent efficiency criteria. Ai, Li and Zhang (2005) and Mandal and Mukerjee (2005) extended their approach to mixed-level designs.

CHAPTER 3

A Correct Approach of Using Nonregular FFDs

Plackett-Burman (PB) and related FFDs are widely used experimental plans for identifying important factors in screening studies where many factors are involved because of their cost efficiencies. The caveat with the use of PB designs is that interactions among factors are implicitly assumed to be non-existent. However, there are many practical situations where some interactions are significant and ignoring them can result in wrong statistical inferences, including biased estimates, missing out on important factors and detection of spurious factors. In this chapter, data from three chemical experiments are reanalyzed using the frequentist approach discussed in subsection 3.1.1. We are able to identify significant interactions in each of these chemical experiments and improve the overall fit of the model. In addition, a Bayesian approach is used for confirming our findings.

3.1 Analysis Strategies

3.1.1 A Frequentist Approach

The frequentist approach, first proposed by Hamada and Wu (1992), consists of three steps.

Step 1. Entertain all the main effects and interactions that are orthogonal to the main effects. Use standard analysis methods such as ANOVA and half-

normal plots to select significant effects.

Step 2. Entertain the significant effects identified in the previous step and the two-factor interactions that consist of at least one significant effect. Identify significant effects using a forward selection regression procedure.

Step 3. Entertain the significant effects identified in the previous step and all the main effects. Identify significant effects using a forward selection regression procedure.

Iterate between Steps 2 and 3 until the selected model stops changing.

This analysis strategy is based on two assumptions. The first assumption is the validity of the *effect sparsity* principle (Box and Meyer 1986), that is, the number of relatively important effects in an experiment is small. The second assumption is the validity of the *effect heredity* principle (Hamada and Wu 1992), that is, when a two-factor interaction is significant, at least one of the corresponding factor main effects is also significant. This precept is needed because it is often difficult to provide a good physical interpretation for a significant interaction XY without either X or Y being significant.

Note that the traditional approach of analysis using PB or other nonregular FFDs ends at Step 1. If the effect heredity principle is not assumed, it is possible to obtain, in step 2, a statistically good fitting model with only interaction effects and no main effect, which is difficult to be interpreted in physical sense. The addition of Step 3 avoids the possibility of missing main effects in Step 1 because of the existence of interactions. However, this analysis strategy would be unnecessary if one could use an all subsets selection procedure for identifying models, even though this unguided search might lead to several problems including heavy computations and uninterpretable models.

3.1.2 A Bayesian Approach

The Bayesian approach suggested by Box and Meyer (1993) considers all the possible explanations (models including interactions) of the data from a screening experiment and identifies those that fit the data well. The prior assumptions are as follows:

1. Effects calculated for inactive factors may be represented approximately as items from a normal distribution with mean zero and standard deviation σ.

2. For a proportion π of active factors the resulting effects are represented as item from a normal distribution with mean zero and a larger standard deviation $\gamma\sigma$.

The prior information is represented in two parameters: γ, the ratio of the standard deviation of the active to the inactive effects, and π, the percentage of active factors. Box, Hunter and Hunter (2005) suggested to choose γ between 2 and 3 and $\pi = 0.25$, based on a survey of a number of published analyses of factorial designs. Recent study has confirmed that the results are not very sensitive to moderate changes in γ and π when active factors are present.

A Bayesian framework is used to assign posterior probabilities to all models considered. Then these posterior probabilities are accumulated to marginal posterior probabilities for each factor. The technical details of the Bayesian analysis are complicated and given in Box and Meyer (1993) or Box, Hunter and Hunter (2005). In practice, one can use the BsProb function in the R library BsMD, free downloadable from the R project homepage (http://www.R-project.org/), for the calculation. One result of the function is the posterior probabilities for

factor activity. A factor X which has a relatively high posterior probability implies that either the main effect of X or an interaction involving X or both are important.

3.1.3 Other Recent Methods

There are further sophisticated analysis strategies proposed for experiments with complex aliasing. Chipman, Hamada and Wu (1997) proposed a Bayesian approach that employs a Gibbs sampler to perform an efficient stochastic search of the model space. Many other recent variable selection methods can also be used for similar purposes. For example, Yuan, Joseph and Lin (2007) suggested an extension of the general-purpose LARS (least angle regression), first proposed by Efron, Hastle, Johnstone and Tibshirani (2004). Phoa, Pan and Xu (2008) suggested the Dantzig Selector method, first proposed by Candes and Tao (2007), for factor screening. Detailed descriptions are referred to Chapter 4 or Phoa, Pan and Xu (2008).

3.2 Reanalysis of Three Chemometrics Experiments

In this section, we reanalyze data from three real chemical experiments using the analysis strategies introduced in section 3.1. These examples illustrate the opportunity for identifying additional important effects that the traditional approach misses, as well as correcting the estimates of the main effects biased from significant interactions. We use the frequentist approach because of its transparency and use the Bayesian approach to confirm the results. In the latter approach we use $\gamma = 2$ and $\pi = 0.25$ as recommended by Box, Hunter and Hunter (2005).

3.2.1 Example 3.1: High-Performance Liquid Chromatography (HPLC) Experiment

Vander Heyden et al. (1999) used the high-performance liquid chromatography (HPLC) method to study the assay of ridogrel and its related compounds in ridogrel oral film-coated tablet simulations. They chose to use a 12-run PB design to identify the importance of 8 factors on several responses. There were several responses in their study but to fix ideas, we consider only one specific response, which is the percentage recovery of ridogrel. The factors and levels are listed in Table 3.1, and the response and the design matrix are shown in Table 3.2. Their analyses showed there was no significant relationship between any of the factors and this response.

The half-normal plot is a graphical tool that uses the ordered estimated absolute effects to help assess the importance of factors. This method is popular among practitioners because it is easy to use and results are easy to interpret. The basic idea is that the larger the estimate of the factor is, the more important the factor is to the response. Figure 3.1 (left) is the half-normal plot for all factors. The plot identifies two important effects E and F,the percentages of organic solvent $\%B$ in the mobile phase at the start and the end of the gradient. The traditional main effect analysis confirms that these factors are the two most important factors. However, the p-values of E and F are 0.16 and 0.24, respectively. This shows important factors are not necessarily significant factors. The model that consists of only the main effects of E and F has $R^2 = 0.41$.

Using the frequentist analysis strategy introduced in subection 3.1.1, we find a significant EF interaction in step 2. Adding EF to E and F increases R^2 from 0.41 to 0.89. In step 3, we further identify factor H (flow of the mobile phase), which is missed in the traditional approach, as significant at the 5%

level. We repeat steps 2 and 3 iteratively until no more new significant effects are identified and the model does not change anymore. When this happens, we stop the procedure and report the final model, which is

$$\hat{Y} = 101.04 - 0.56E + 0.44F - 0.30H + 0.88EF \tag{3.1}$$

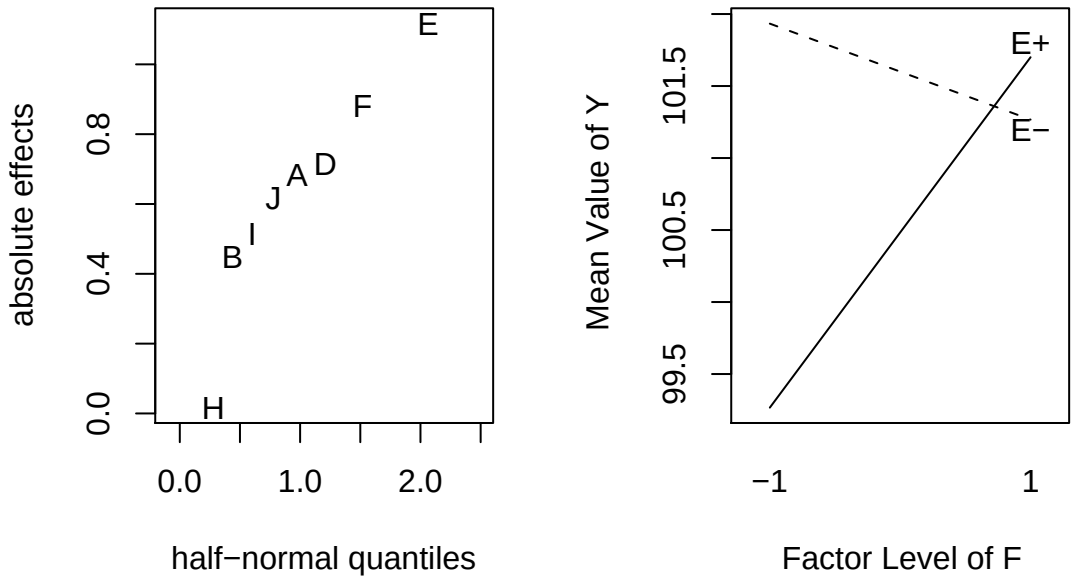
Here E, F and H take on the values either 1 or -1 corresponding to the $+$ and $-$ factor levels respectively. This model has $R^2 = 0.96$, indicating a good fit. In the model (3.1), H is significant at the 5% level (p-value=0.012) and E, F and EF are significant at the 1% level.

Figure 3.1 (right) is the EF interaction plot, which shows the average responses at the level combinations of E and F. The solid (dashed) line represents the change of the mean value of the response at the high (low) level of E when F changes from its low level to its high level. Note that if E is at its high level, the response increases rapidly when F changes from its low level to its high level. On the other hand, if E is at its low level, the response decreases slowly when F changes from its low level to its high level. In other words, F has a large positive effect when E is at its low level whereas F has a small negative effect when E is at its high level. Because the effect of F depends on the level of E, the EF interaction is significant. (If the EF interaction was not significant, the two lines in the interaction plot would be nearly parallel.)

In the main effect model, the estimate of factor H is -0.01 and is not significant at all. This can be explained using the model (3.1). It is easy to verify that the correlation between H and EF is 1/3. Therefore, the estimate of H in the main effect model actually estimates $H + \frac{1}{3}EF$. Note that $-0.30 + \frac{1}{3}(0.88) \approx -0.01$. The main effect of H is not significant in the main effect model because it is partially canceled by the EF interaction.

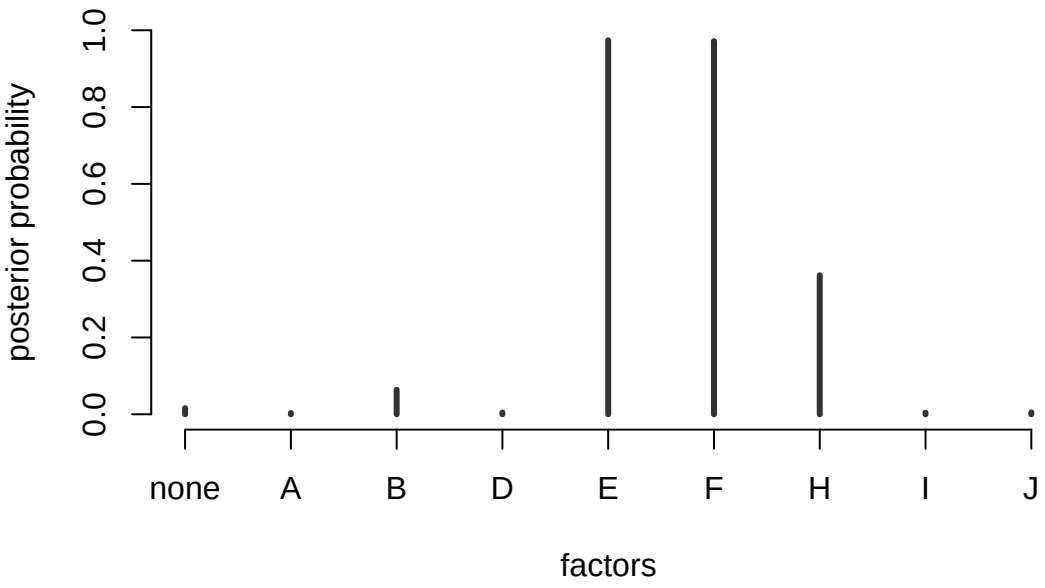
We next analyze the data via the Bayesian approach. The posterior proba-

Figure 3.1: The HPLC Experiment: (Left) Half-normal plot; and (Right) Interaction plot of EF.



bility plot in Figure 3.2 shows that both factors E and F have very high posterior probability, factor H has a moderate posterior probability and the posterior probability of other factors are negligibly small. A high posterior probability of a factor means that the factor main effect and/or its interactions are significant to the response. Therefore, this figure supports the model (3.1) well because the E and F main effects and their interactions are significant and factor H is also significant, but not as significant as factors E and F.

Figure 3.2: The HPLC Experiment: Posterior Probability Plot.



The posterior probability plot in Figure 3.2 is produced by the free and popular statistical software R using the commands

```
library(BsMD)
bs=BsProb(X, y, mInt=2, p=0.25, g=2)
plot(bs, code=F)
```

The first command loads the BsMD package into R and the second and third commands perform the Bayes computation and produce the posterior probability plot. Here X is the 12×8 design matrix and y is the 12×1 response vector given in Table 3.2. The option `mInt=2` means that the Bayesian approach evaluates models with two-factor interactions besides the main effects. The options `p=0.25` and `g=2` specify $\pi = 0.25$ and $\gamma = 2$. Our experience is that the choices of π and γ do not seem to be critical.

3.2.2 Example 3.2: Pressurized Liquid Extraction (PLE) Experiment

Moreda-Pineiro et al. (2007) developed the acetic acid-pressurized liquid extraction (PLE) for the simultaneous extraction of major and trace elements from edible seaweeds. They chose to use a 24-run PB design to study the importance of 8 factors on the mean recovery percentage of released elements, which is defined as:

$$meanrecovery(\%) = \frac{1}{N} \sum ((C_{PLE}/C_{digested}) \times 100)$$

where C_{PLE} and $C_{digested}$ are the element concentrations obtained after the PLE procedure and the acid digestion procedure respectively, and N is the number of elements studied. In their study, they used $N = 10$. Note that Moreda-Pineiro et al. (2007) defined their eighth variable as an imaginary dummy variable for evaluating the possible systematic error and/or the existence of important variables that have not been considered. Our analysis simply treats this dummy variable the same as other 7 variables. The factors and levels are listed in Table 3.3, and the response and the design matrix are shown in Table 3.4.

Following the traditional approach, factors r, s and T, mass/sample ratio, particle size and extraction temperature, are identified as significant, although s and T have much smaller effects; see the half-normal plot in Figure 3.3 (left). The p-value of r, s and T are 0.01, 0.07 and 0.12 respectively. The model that consists of only the main effects of r, s and T has $R^2 = 0.48$.

We treat both particle size and extraction temperature as barely significant and do not include them in step 2. Then using the frequentist analysis strategy in subsection 3.1.1, we find a significant rs interaction in step 2. Figure 3.3 (right) is the rs interaction plot. The solid and dashed lines represent the change of the main value of our response at the high and low level of r respectively when we change the level of s. The rs interaction is significant because there is a large difference between the two slope of these two lines. Adding rs to r, s and T leads to an increase of R^2 to 0.65. When we add rs to the main effect model in step 3, factor T becomes insignificant. The repeat of steps 2 and 3 does not identify new significant effects, so the procedure stops.

Using the latest model, one can obtain the predicted mean recovery percentage of released elements ($\hat{Y}$) by

$$\hat{Y} = 83.42 + 2.83r + 1.92s - 2.67rs \tag{3.2}$$

where r and s take on the values either 1 or -1 corresponding to the $+$ and $-$ factor levels respectively. We further analyze the data via the Bayesian approach. The posterior probability plot in Figure 3.4 supports the model Equation (3.2). Both factors r and s have very high posterior probability because both their main effects and their interaction are significant. Moreover, factor T has a negligibly small posterior probability, suggesting that factor T is insignificant.

The main effect Pareto chart in Moreda-Pineiro et al. (2007) and the half-normal plot in Figure 3.3 (left) arrange the factors in the order from the most

Figure 3.3: The PLE Experiment: (Left) Half-normal plot; and (Right) Interaction plot of rs.

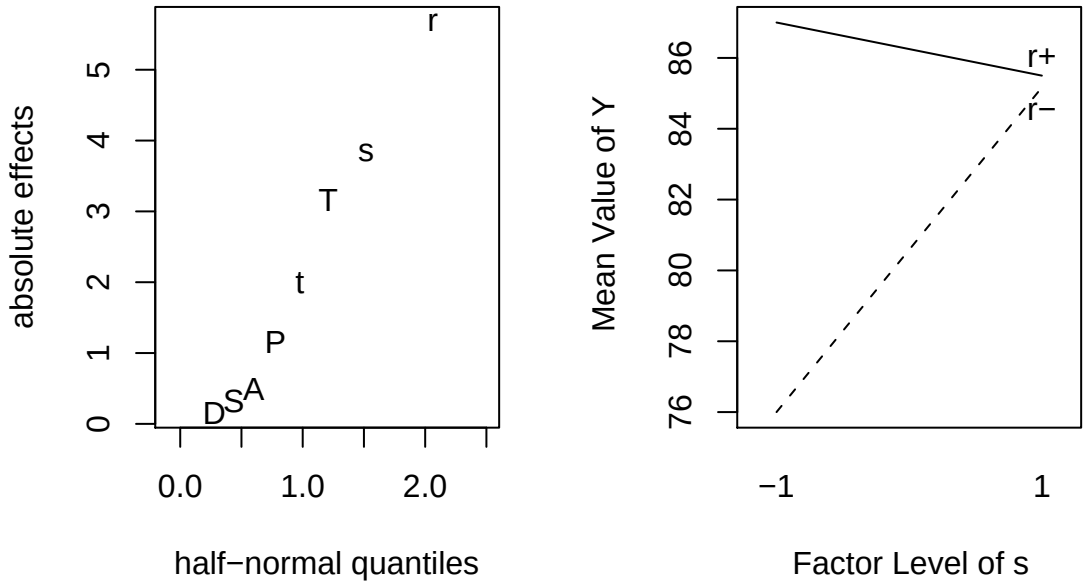
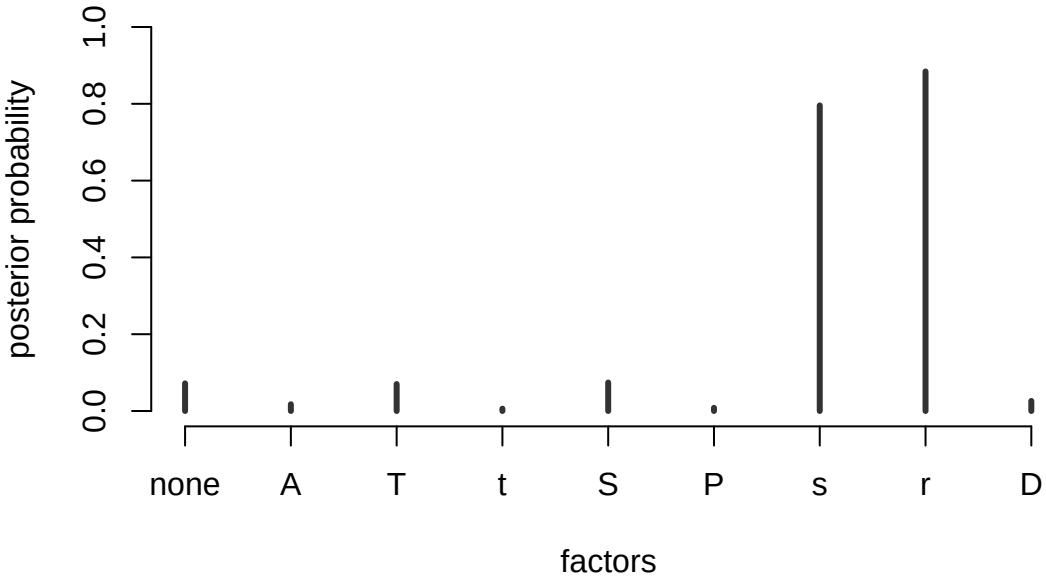


Figure 3.4: The PLE Experiment: Posterior Probability Plot



significant to the least significant as follows: $r > s > T > t > P > A > S > D$. This order of significance to the response is based on the main effect model as follows.

$$\hat{Y} = 83.42 - 0.25A - 1.58T + 1.00t - 0.17S + 0.58P + 1.92s + 2.83r + 0.08D. \quad (3.3)$$

This main effect model does not include the significant rs interaction, and the inclusion leads to a dramatic change in the significance order. The model consisting of all main effects and the rs interaction is as follows.

$$\hat{Y} = 83.42 - 1.97A + 0.14T - 0.72t - 1.89S - 1.14P + 1.92s + 2.83r - 1.64D - 5.17rs. \quad (3.4)$$

The new descending order of factor significance, based on the p-values, becomes $rs > r > s > A > S > D > P > t > T$.

The new order features some major position shifts like factors A and T, and the change is due to the bias from the partial aliasing pattern between main effects and the significant rs interaction. For example, one observes that $\hat{T} = -1.58$ in the main effect model Equation (3.3) actually estimates $T + \frac{1}{3}rs$ because the correlation between T and rs is 1/3. Since $\hat{rs} = -5.17$ is found in the model Equation (3.4), one needs to correct the estimate of T by $\hat{T} - \frac{1}{3}\hat{rs} = -1.58 - \frac{1}{3}(-5.17) = 0.14$; see Equation (3.4). Note that this bias reduction by subtracting rs from $\hat{T}$ changes the sign of T from negative to positive, which means that the effect of changing T from high level to low level is actually opposite to what the main effect model suggested. In addition, the bias reduction leads to a change of p-value of T from 0.12 in the main effect model Equation (3.3) to 0.85 in the model Equation (3.4). As a consequence, T changes from the third most significant factor in the main effect model Equation (3.3) to the least significant factor in the model Equation (3.4). This explains why our final model Equation (3.3) does not include factor T.

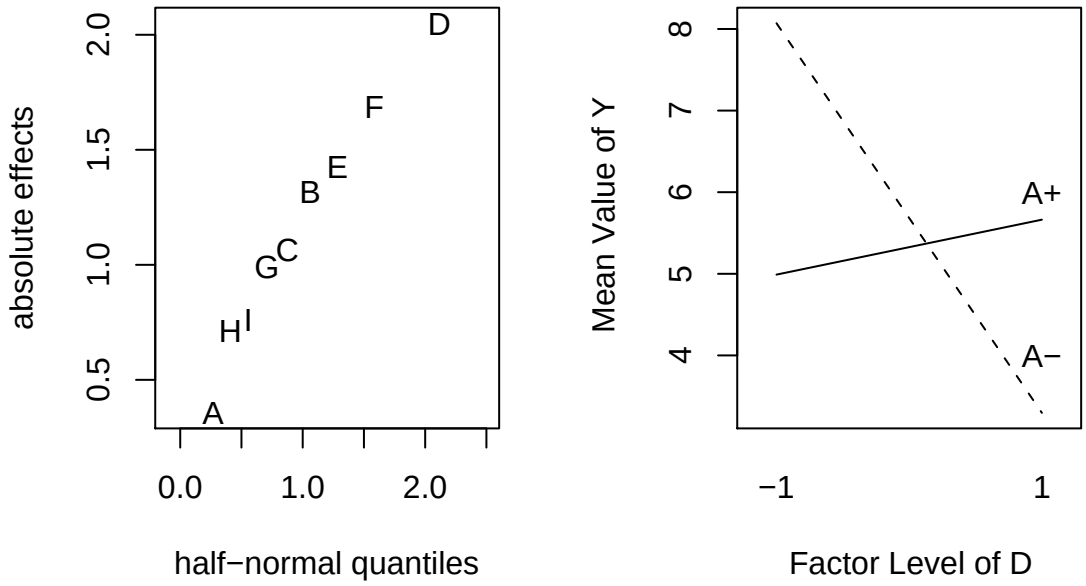
3.2.3 Example 3.3: Compound Extraction Experiment

Dopico-Garcia et al. (2007) developed a two-step analytical methodology that allowed the chemical characterization of white grapes by simultaneously determining their most important phenolic compounds and organic acids. Among all 11 phenolic compounds and organic acids tested in the original experiment, only one phenolic compound, kaempferol-3-O-rutinoside + isorhamnetin-3-O glucoside, will be considered in this example. They used a 12-run PB design to select 8 variables that have influence in both two steps on the system. Note that their ninth variable is an imaginary dummy variable for error evaluation. The factors and levels are listed in Table 3.5, and the response and the design matrix are shown in Table 3.6.

Following the traditional approach, factors D and F, temperature and sorbent type, are identified as significant, although F has a much smaller effect; see the half-normal plot in Figure 3.5 (left). The p-value of D and F are 0.24 and 0.31 respectively. The model that consists of only the main effects of D and F has $R^2 = 0.41$.

Using the frequentist analysis strategy in subsection 3.1.1, we find a significant AD interaction in step 2. Figure 3.5 (right) is the AD interaction plot. The solid and dashed lines represent the change of the main value of our response at the high and low level of A respectively when we change the level of D. The AD interaction is significant because there is a large difference between the two slope of these two lines. Adding AD to D and F leads to an increase of R^2 to 0.71. When we add AD to the main effect model in step 3, factor F becomes insignificant and factor C (extraction time), which is missed in the traditional approach, is identified as significant. The repeat of steps 2 and 3 does not identify new significant effects, so the procedure stops.

Figure 3.5: The Compound Extraction Experiment: (Left) Half-normal plot; and (Right) Interaction plot of rs.

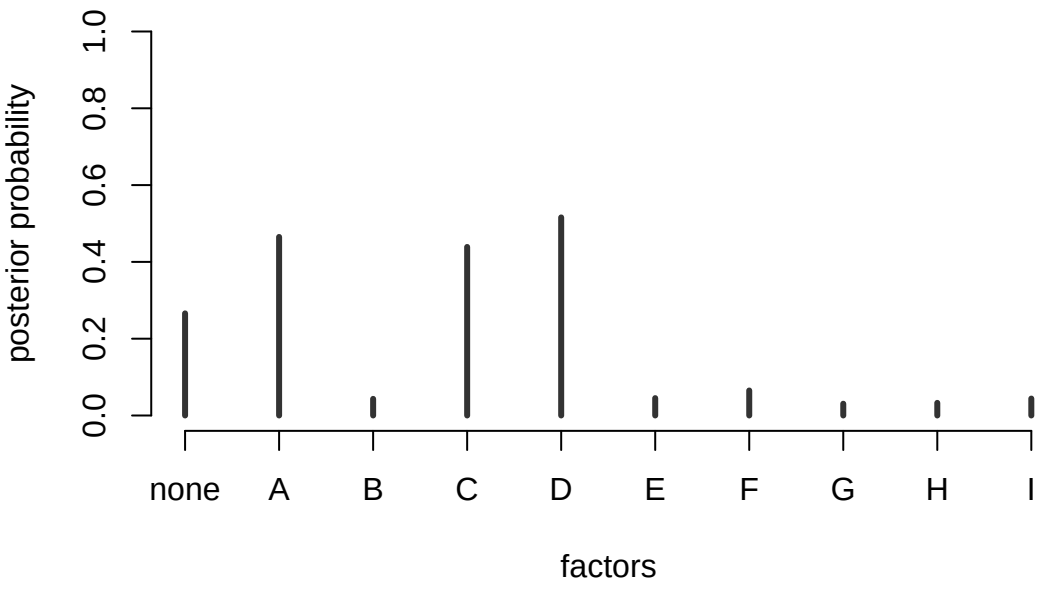


Using the latest model, one can obtain the predicted mean recovery percentage $(\hat{Y})$ by

$$\hat{Y} = 5.51 + 1.11C - 1.03D + 1.73AD \tag{3.5}$$

where A, C and D take on the values either 1 or -1 corresponding to the $+$ and $-$ factor levels respectively. We further analyze the data via the Bayesian approach. The posterior probability plot in Figure 3.6 supports the model Equation (3.5). Factors A, C and D have relatively high posterior probability because either their main effects or their interaction or both are significant. Moreover, factor F has a small posterior probability, suggesting that factor F is insignificant.

Figure 3.6: The Compound Extraction Experiment: Posterior Probability Plot



The main effect model, used in Dopico-Garcia et al. (2007), misidentified the sorbent type (factor F) as significant because of its partial aliasing pattern with the truly significant AD interaction. Factor F is found to be insignificant after being disentangled from AD. In opposite, the main effect model did not identify the extraction time (factor C) as significant because of the effect cancelation with AD. The disentangelement reveals the significance of C to the response.

The misidentification of significant factors may lead to inefficient experiment setting. Assume that the purpose of this experiment is to maximize the response. The maximum response from the model consisting of factors D and F is 7.37 units, obtained by setting both factors D and F at low level. The maximum response from our latest model (Equation 3.5) is 9.38 units, obtained by setting factors D and A at low level and factor C at high level. It is obvious that our latest model increases the response by 2.01 units or 27%.

3.3 Conclusion

This chapter aims at introducing a correct way of using the PB designs in the analysis of experiments. Although traditional approaches always suggested to neglect the interactions, some of these effects might be significant to the response. They might be crucial to be included in the model or inefficient models may be resulted.

Three real-life examples demonstrated how inefficient the model could be if one ignored the existence of significant interaction effects. In the first example, our suggested final model fits much better than the main effect model suggested by the traditional approach. The R^2 increases from 0.41 in the main effect model to 0.96 in our final model. In addition, a newly-found significant factor is iden-

tified after disentangling the partial aliasing from the significant interaction. In the second example, the order of factor significance may not be valid without reducing the biases of the main effects from the significant interaction. It results in assigning incorrect factor levels, which eventually lead to inefficient experiments and inconclusive studies. In the last example, the significance of main effects is meshed due to the partial aliasing between main effects and significant interactions. In particular, a spurious factor is detected and an important factor is missed in the traditional approach. This leads to a 27% decrease in the response maximization process.

In summary, ignoring the existence of significant interaction leads to three pitfalls: (i) some significant interaction effects may be missed, (ii) some insignificant main effects may be misidentified as significant and vice versa, and (iii) the level assignments of some factors may be opposite. The consideration of the interactions may avoid all above errors and a more accurate and efficient model is resulted.

Table 3.1: High-Performance Liquid Chromatography (HPLC) Experiment: Factors and Levels

Symbol	Factor	Unit	Factor Level	
			Low (−)	High (+)
A	pH of the buffer		6.5	7.1
B	Column manufacturer		Alltech	Prodigy
D	Column temperature	^{o}C	23	33
E	$\%B$ in the mobile phase at the start of the gradient	%	24	26
F	$\%B$ in the mobile phase at the end of the gradient	%	41	45
H	Flow of the mobile phase	ml/min	1.4	1.6
I	Detection wavelength	nm	260	270
J	Concentration of the buffer	%, m/v	0.225	0.275

Table 3.2: High-Performance Liquid Chromatography (HPLC) Experiment: Design Matrix and Responses

	Design								Response
Run	A	B	D	E	F	H	I	J	$\%MC$
1	+1	+1	−1	+1	+1	+1	−1	−1	101.6
2	+1	+1	+1	−1	−1	+1	+1	+1	101.7
3	+1	−1	+1	−1	+1	−1	−1	+1	101.6
4	+1	−1	−1	+1	+1	−1	+1	+1	101.9
5	+1	−1	−1	−1	−1	+1	+1	−1	101.8
6	−1	+1	+1	−1	+1	−1	+1	−1	101.1
7	−1	+1	−1	−1	+1	+1	−1	+1	101.1
8	−1	−1	+1	+1	+1	+1	+1	−1	101.6
9	−1	−1	+1	+1	−1	+1	−1	+1	98.4
10	−1	+1	−1	+1	−1	−1	+1	+1	99.7
11	+1	+1	+1	+1	−1	−1	−1	−1	99.7
12	−1	−1	−1	−1	−1	−1	−1	−1	102.3

Table 3.3: Pressurized Liquid Extraction (PLE) Experiment: Factors and Levels

Symbol	Factor	Unit	Factor Level	
			Low (−)	High (+)
A	Acetic acid concentration	M	0.75	1.5
T	Extraction temperature	^{o}C	25	100
t	Extraction time	min	5	10
S	Extraction step		1	3
P	Pressure	MPa	10.3	17.2
s	Particle size	μm	125	200
r	Mass/sample ratio		1	4
D	Dummy factor		-1	$+1$

Table 3.4: Pressurized Liquid Extraction (PLE) Experiment: Design Matrix and Responses

	Design								Response
Run	A	T	t	S	P	s	r	D	% Recovery
1	+1	−1	+1	−1	−1	−1	+1	+1	88
2	+1	+1	−1	+1	−1	−1	−1	+1	69
3	−1	+1	+1	−1	+1	−1	−1	−1	73
4	+1	−1	+1	+1	−1	+1	−1	−1	81
5	+1	+1	−1	+1	+1	−1	+1	−1	86
6	+1	+1	+1	−1	+1	+1	−1	+1	88
7	−1	+1	+1	+1	−1	+1	+1	−1	84
8	−1	−1	+1	+1	+1	−1	+1	+1	83
9	−1	−1	−1	+1	+1	+1	−1	+1	88
10	+1	−1	−1	−1	+1	+1	+1	−1	85
11	−1	+1	−1	−1	−1	+1	+1	+1	85
12	−1	−1	−1	−1	−1	−1	−1	−1	81
13	+1	−1	+1	−1	−1	−1	+1	+1	88
14	+1	+1	−1	+1	−1	−1	−1	+1	71
15	−1	+1	+1	−1	+1	−1	−1	−1	79
16	+1	−1	+1	+1	−1	+1	−1	−1	88
17	+1	+1	−1	+1	+1	−1	+1	−1	88
18	+1	+1	+1	−1	+1	+1	−1	+1	83
19	−1	+1	+1	+1	−1	+1	+1	−1	89
20	−1	−1	+1	+1	+1	−1	+1	+1	89
21	−1	−1	−1	+1	+1	+1	−1	+1	83
22	+1	−1	−1	−1	+1	+1	+1	−1	83
23	−1	+1	−1	−1	−1	+1	+1	+1	87
24	−1	−1	−1	−1	−1	−1	−1	−1	83

Table 3.5: Compound Extraction Experiment: Factors and Levels

Symbol	Factor	Unit	Factor Level	
			Low (−)	High (+)
A	Extraction solvent		Acid Water	MeOH
B	Extraction volume	mL	50	250
C	Extraction time	min	5	20
D	Temperature	^{o}C	40	50
E	Extraction type		Ultrasonic	Stirring
F	Sorbent type		EC	NEC
G	Elution solvent		EtOH	MeOH
H	Elution volume	mL	20	150
I	Dummy factor		−1	+1

Table 3.6: Compound Extraction Experiment: Design Matrix and Responses

	Design									Response
Run	A	B	C	D	E	F	G	H	I	Y
1	+1	−1	+1	−1	−1	−1	+1	+1	+1	6.98
2	+1	+1	−1	+1	−1	−1	−1	+1	+1	5.31
3	−1	+1	+1	−1	+1	−1	−1	−1	+1	9.67
4	+1	−1	+1	+1	−1	+1	−1	−1	−1	6.45
5	+1	+1	−1	+1	+1	−1	+1	−1	−1	5.23
6	+1	+1	+1	−1	+1	+1	−1	+1	−1	5.34
7	−1	+1	+1	+1	−1	+1	+1	−1	+1	4.03
8	−1	−1	+1	+1	+1	−1	+1	+1	−1	3.76
9	−1	−1	−1	+1	+1	+1	−1	+1	+1	2.10
10	+1	−1	−1	−1	+1	+1	+1	−1	+1	2.65
11	−1	+1	−1	−1	−1	+1	+1	+1	−1	7.40
12	−1	−1	−1	−1	−1	−1	−1	−1	−1	7.14

CHAPTER 4

Analysis of Nonregular and Supersaturated Designs

Supersaturated designs are a special class of nonregular FFDs whose run sizes are not enough for estimating all the main effects. A common application of supersaturated designs is *factor screening*. There are usually a large number of factors to be investigated in a screening experiment, but it is believed that only a few of them are active, or explicitly speaking, have significant impact on the response. This phenomenon is commonly recognized as *factor sparsity* (Box and Meyer 1986). The purpose of screening experiments is to identify the active factors correctly and economically. The inactive factors will be discarded, while the active factors will be investigated further in some follow-up experiments. Supersaturated designs are particularly useful in screening experimentation due to their run-size economy (Lin 1999).

Some analysis methods were developed in recent years. Lin (1993) used stepwise regression for selecting active factors. Chipman et al. (1997) proposed a Bayesian variable-selection approach for analyzing experiments with complex aliasing. Westfall et al. (1998) proposed an error control skill in forward selection. Beattie et al. (2002) proposed a two-stage Bayesian model selection strategy for supersaturated experiments. Li and Lin (2002, 2003) proposed a method based on penalized least squares. Holcomb et al. (2003) proposed contrast-based meth-

ods. Lu and Wu (2004) proposed a modified stepwise selection based on an idea of staged dimensionality reduction. Zhang et al. (2007) proposed a method based on partial least squares.

In this chapter, we consider searching active factors in supersaturated designs via the Dantzig selector proposed by Candes and Tao (2007). The Dantzig selector chooses the best subset of variables or active factors by solving a simple convex program, which can be recast as a convenient linear program. Candes and Tao (2007) showed that the Dantzig selector has some remarkable properties under some conditions and has been successfully used in biomedical imaging, analog to digital conversion and sensor networks, where the goals are to recover some sparse signals from some massive data. Our simulation also demonstrates that the Dantzig selector is powerful for analyzing supersaturated designs.

4.1 The Dantzig Selector

Consider a linear regression model $y = X\beta + \epsilon$ where y is an $n \times 1$ vector of observations, X is an $n \times k$ model matrix, β is a $k \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ vector of random errors. Assume that $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ is a vector of independent normal random variables. Candes and Tao (2007) proposed a new estimator called the *Dantzig selector* to estimate the vector of parameters β under the situation of supersaturated experiments (i.e., the number of variables is greater than the number of observations). This estimator is the solution to the l_1-regularization problem

$$\min_{\hat{\beta} \in R^k} \|\hat{\beta}\|_{l_1} \text{ subject to } \|X^t r\|_{l_\infty} \leq \delta \tag{4.1}$$

where r is the residual vector $r = y - X\hat{\beta}$, δ is a tuning parameter and for a vector a, $\|a\|_{l_1} = \sum |a_i|$ and $\|a\|_{l_\infty} = \max |a_i|$. In other words, an estimator with

minimum complexity measured by the l_1-norm is searched among all estimators that are consistent with the data.

According to Candes and Tao (2007), there are some reasons to restrict the correlated residual vector $X^t r$ rather than the size of the residual vector r. One of the reasons is that the estimation procedure using the correlated residual vector is invariant with respect to orthonormal transformations applied to the data vector since the feasible region is invariant. Suppose an orthonormal transformation is applied to the data, giving $y' = Uy$, then $(UX)^t(Uy - UX\hat{\beta}) = X^t(y - X\hat{\beta})$, which shows the invariant. This implies that the estimation of β does not depend upon U.

The Dantzig selector can be recast as a linear program.

$$\min \sum_i u_i \text{ subject to } -u \leq \hat{\beta} \leq u \text{ and } -\delta \mathbf{1}_k \leq X^t(y - X\hat{\beta}) \leq \delta \mathbf{1}_k \qquad (4.2)$$

where the optimization variables are u, $\hat{\beta} \in R^k$ and $\mathbf{1}_k$ is a vector of k ones. This is equivalent to the standard linear program

$$\min c^t x \text{ subject to } Ax \geq b \text{ and } x \geq 0 \qquad (4.3)$$

where

$$c = \begin{pmatrix} \mathbf{1}_k \\ \mathbf{0}_k \end{pmatrix}, A = \begin{pmatrix} X^tX & -X^tX \\ -X^tX & X^tX \\ 2I_k & -I_k \end{pmatrix}, b = \begin{pmatrix} -X^ty - \delta\mathbf{1}_k \\ X^ty - \delta\mathbf{1}_k \\ \mathbf{0}_k \end{pmatrix}, x = \begin{pmatrix} u \\ u + \beta \end{pmatrix}.$$

Candes and Tao (2007) showed that under certain conditions on the model matrix X which roughly guarantee that the model is identifiable, the Dantzig selector can correctly identify the active variables with large probability. Unfortunately, the conditions are too strong and most supersaturated designs in the literature do not satisfy these conditions.

When the columns of X are orthogonal and have unit length, the Dantzig selector $\hat{\beta}$ is the l_1-minimizer subject to the constraint $\|X^t y - \hat{\beta}\|_{l_\infty} \leq \delta$. This implies that $\hat{\beta}$ is simply the soft-thresholded version of $X^t y$ at level δ, thus

$$\hat{\beta}_i = \begin{cases} (X^t y)_i - \delta, & \text{if } (X^t y)_i \geq \delta \\ (X^t y)_i + \delta, & \text{if } (X^t y)_i \leq -\delta \\ 0, & \text{otherwise} \end{cases}$$

where $(X^t y)_i$ is the ith component of $X^t y$. In other words, $X^t y$ is shifted toward the origin. For an arbitrary X, the method continues to exhibit a soft-thresholding type of behavior and as a result, may slightly underestimate the true value of the nonzero parameters.

There are several simple methods to correct for this bias and increase performance in practical settings. Candes and Tao (2007) suggested a two-stage procedure. First, estimate $I = \{i : \beta_i \neq 0\}$ with $\hat{I} = \{i : |\hat{\beta}_i| > \gamma \}$ for some $\gamma \geq 0$ with $\hat{\beta}$ as in the solution to the l_1-regularization problem Equation (4.1). Second, construct the estimator $\hat{\beta}_{\hat{I}} = (X^t_{\hat{I}} X_{\hat{I}})^{-1} X^t_{\hat{I}} y$ and set the other coordinates to zero, where $X_{\hat{I}}$ is corresponding model matrix for model $\hat{I}$. Hence, we rely on the Dantzig selector to estimate the model I by $\hat{I}$, and construct a new estimator by regressing y onto the model $\hat{I}$. Candes and Tao (2007) referred to this estimator as the Gauss-Dantzig selector. This estimator centralizes the estimates and generally yields higher statistical accuracy.

The tuning parameter (δ) in the l_1-regularization problem Equation (4.1) has a significant impact on the results of the estimates. If δ is set to be too high, or in other words, we allow a large range of residuals to take part in the regression equation, the residuals are able to explain all the variations of the response themselves without considering any changes in the predictors. This leads to the insignificance of all predictors towards the change in response, so

we drop all of the predictors. On the other hand, if δ is set to be too low, or in other words, we minimize the variation of the residuals, the variation of the response has to be explained by the predictors, so some inactive factors with small magnitudes of coefficients are falsely included to help explaining the variation of the response. Therefore, an appropriate value of δ is essential in identification of the active factors.

4.2 A Procedure for Analyzing Supersaturated Designs

A proper choice of the tuning parameter δ is crucial for the Dantzig selector. Candes and Tao (2007) suggested the choice of $\delta = \lambda\sigma$ when X is unit length normalized, where $\lambda = \sqrt{2\log k}$ and σ is the standard deviation of the random error. However, we do not know σ in practice and it is a difficult task itself to estimate σ accurately for supersaturated designs. Furthermore, even if we know σ (as in simulation), this choice of δ does not always lead to the best performance. Cai and Lv (2007) argued that it might be possible that $\lambda = \sqrt{2\log k}$ overshrinks the $k\times 1$ vector of unknown parameters β and underestimates the nonzero coordinates when k is much larger than n.

Here we propose a simple approach to choosing δ and analyzing supersaturated designs based on a profile plot of the estimates. The steps are as follows.

1. Standardize data so that y has mean 0 and columns of X have equal lengths.
2. Solve the linear program Equation (4.2) or Equation (4.3) to obtain the Dantzig selector $\hat{\beta}$ for a range of δ.
3. Make a profile plot of the estimates by plotting $\hat{\beta}$ against δ.
4. Choose a proper δ and active effects according to the profile plot.

5. Obtain new estimates by regressing y on the active effects selected in step 4.

We illustrate the approach on three real data in the literature.

Example 1. Consider the cast fatigue experiment (Wu and Hamada 2000, section 7.1), a real data set consisting of 7 two-level factors. The design matrix and the response data are given in Table 4.1. We first consider the main effects model, where each column corresponds to a two-level factor. The profile plot (Figure 4.1) suggests one or two important factors if δ is chosen between 1.9 and 5.4. If δ is chosen between 1.9 and 3.0, both F and D are significant. If δ is chosen between 3.1 and 5.4, only F is significant. The conclusion is that F is very significant and D is marginal. Our result is consistent to the analysis using half-normal plot in Wu and Hamada (2000, Figure 8.1).

We further investigate potential active two-factor interactions. We consider a model with 7 main effects and all 21 two-factor interactions so that the model is supersaturated. The profile plot (Figure 4.2) suggests two or three important effects if δ is chosen between 1.3 and 5.4. Among the three effects, AE is less significant than F and FG, which agrees with the result in Westfall et al. (1998). Note that the significance of AE without its parent main effects violates the effect heredity principle (Wu and Hamada 2000, section 3.5), so one might accept a model with F and FG only, which is recommended by Wu and Hamada (2000, section 8.4).

Example 2. Consider the blood glucose experiment (Wu and Hamada 2000, section 7.1), a real data set consisting of 1 two-level and 7 three-level factors. The design matrix and the response data are given in Table 4.2. We first apply the Dantzig selector to a model with 15 terms. The first column corresponds to the two-level factor A. The next 7 columns correspond to the linear contrasts

Table 4.1: Design Matrix and Response Data, Cast Fatigue Experiment.

Run	A	B	C	D	E	F	G	Response
1	+	+	−	+	+	+	−	6.058
2	+	−	+	+	+	−	−	4.733
3	−	+	+	+	−	−	−	4.625
4	+	+	+	−	−	−	+	5.899
5	+	+	−	−	−	+	−	7.000
6	+	−	−	−	+	−	+	5.752
7	−	−	−	+	−	+	+	5.682
8	−	−	+	−	+	+	−	6.607
9	−	+	−	+	+	−	+	5.818
10	+	−	+	+	−	+	+	5.917
11	−	+	+	−	+	+	+	5.863
12	−	−	−	−	−	−	−	4.809

Figure 4.1: Profile plot for the cast fatigue experiment without interactions. The model includes 7 main effects.

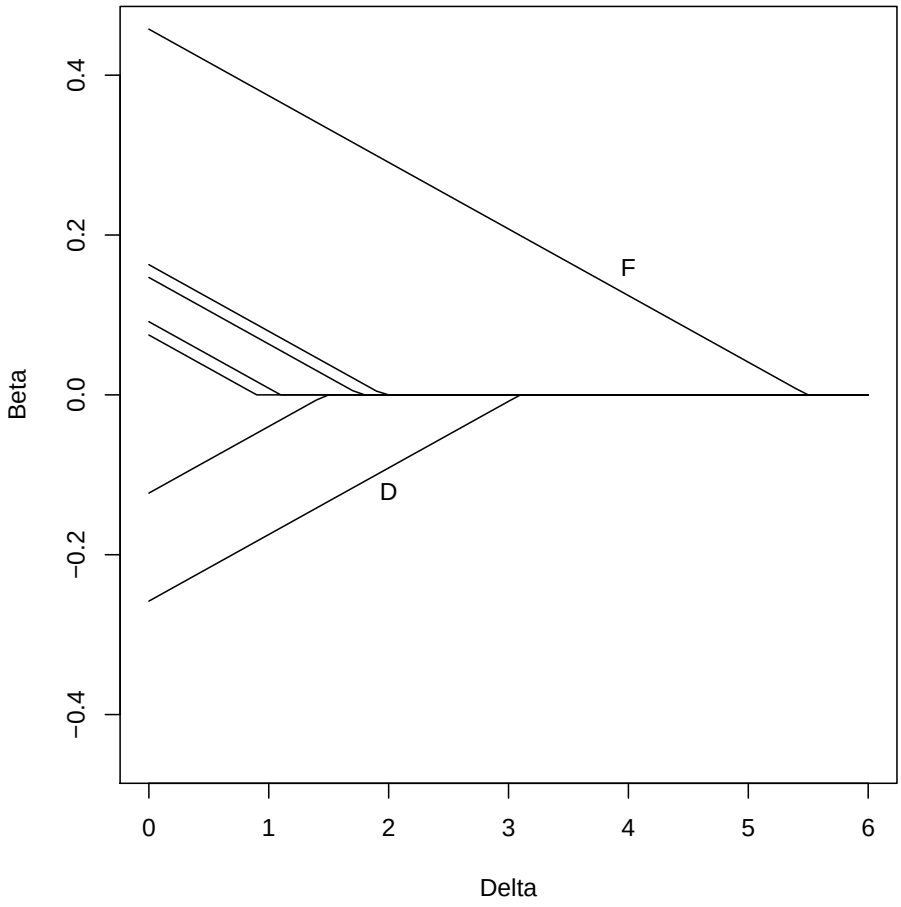
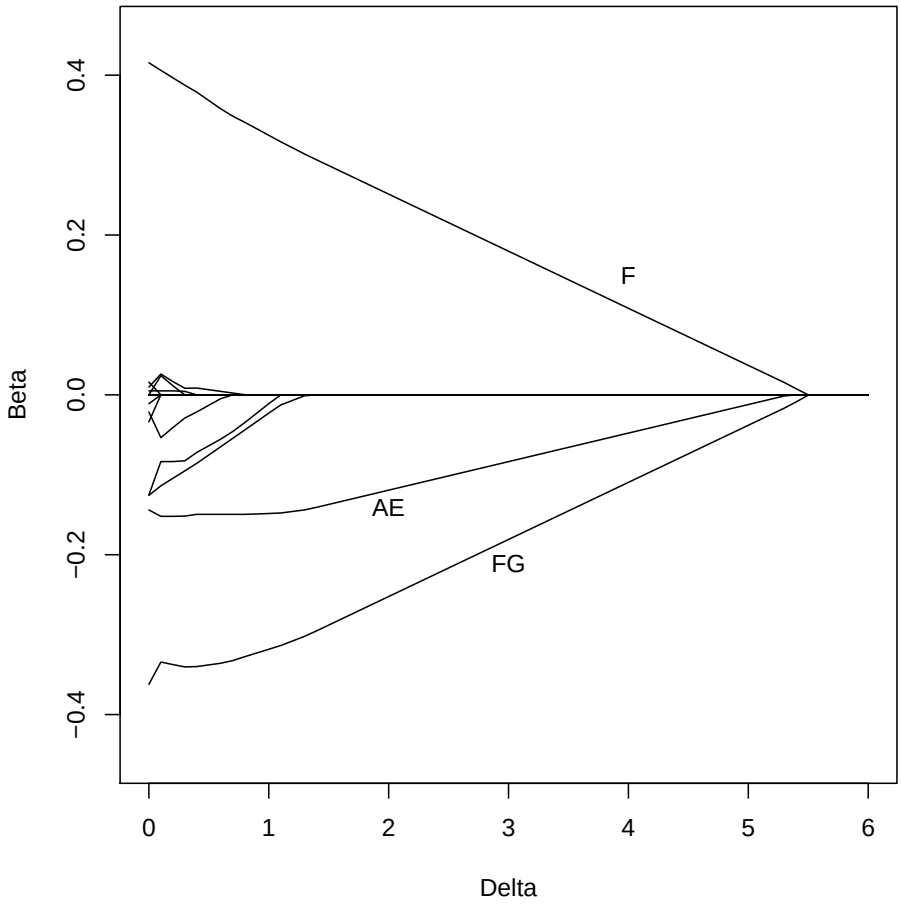


Figure 4.2: Profile plot for the cast fatigue experiment with interactions. The model contains 7 main effects and 21 two-factor interactions.



of the 7 three-level factors from B to H. The last 7 columns correspond to the quadratic contrasts of the 7 three-level factors. The coding of linear and quadratic contrasts is:

$$\text{Linear Contrast: } \begin{pmatrix} 0 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} +1 & 0 & -1 \end{pmatrix}$$
$$\text{Quadratic Contrast: } \begin{pmatrix} 0 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} +1 & -2 & +1 \end{pmatrix}$$

The model matrix X is then normalized to have unit length for each column. The profile plot (Figure 4.3) suggests that E_q and F_q are the only two significant effects if δ is chosen between 12.2 and 17.1. Our result is consistent to the analysis using half-normal plot in Wu and Hamada (2000, Figure 8.2).

We also include two-factor interaction terms in the analysis and consider a model with 15 linear and quadratic terms and 98 two-factor interaction terms. The model matrix X is normalized to have unit length for each column. The profile plot (Figure 4.4) suggests two significant effects, $(BH)_{lq}$ (the interaction between the linear contrast of B and the quadratic contrast of H) and $(BH)_{qq}$ (the interaction between the quadratic contrasts of B and H), if δ is chosen between 18.5 and 23.1. If δ is chosen between 23.2 and 28.2, only $(BH)_{lq}$ is significant. The result does not completely match Equation 8.10 of Wu and Hamada (2000), but is consistent to the top model identified by a Bayesian approach in Wu and Hamada (2000, Table 8.3).

Example 3. In this example, we apply the Dantzig selector to a supersaturated design demonstrated first by Lin (1993). The original dataset has 24 factors but two factors (13 and 16) are identical. As Beattie et al. (2002), we delete factor 13 and rename factors 14–24 as 13–23. The design matrix and response data are given in Table 4.3. The profile plot (Figure 4.5) suggests that only X_{14} appears to be important in this data.

Table 4.2: Design Matrix and Response Data, Blood Glucose Experiment.

Run	A	B	C	D	E	F	G	H	Response
1	0	0	0	0	0	0	0	0	97.94
2	0	1	1	1	1	1	0	1	83.40
3	0	2	2	2	2	2	0	2	95.88
4	0	0	0	1	1	2	1	2	88.86
5	0	1	1	2	2	0	1	0	100.58
6	0	2	2	0	0	1	1	1	89.57
7	0	0	1	0	2	1	2	2	91.98
8	0	1	2	1	0	2	2	0	98.41
9	0	2	0	2	1	0	2	1	87.56
10	1	0	1	2	1	1	0	0	88.11
11	1	1	2	0	2	2	0	1	83.81
12	1	2	0	1	0	0	0	2	98.27
13	1	0	2	2	0	2	1	1	115.52
14	1	1	0	0	1	0	1	2	94.89
15	1	2	1	1	2	1	1	0	94.70
16	1	0	2	1	2	0	2	1	121.62
17	1	1	0	2	0	1	2	2	93.86
18	1	2	1	0	1	2	2	0	96.10

Figure 4.3: Profile plot for the blood glucose experiment without interactions. The model contains 15 main effects.

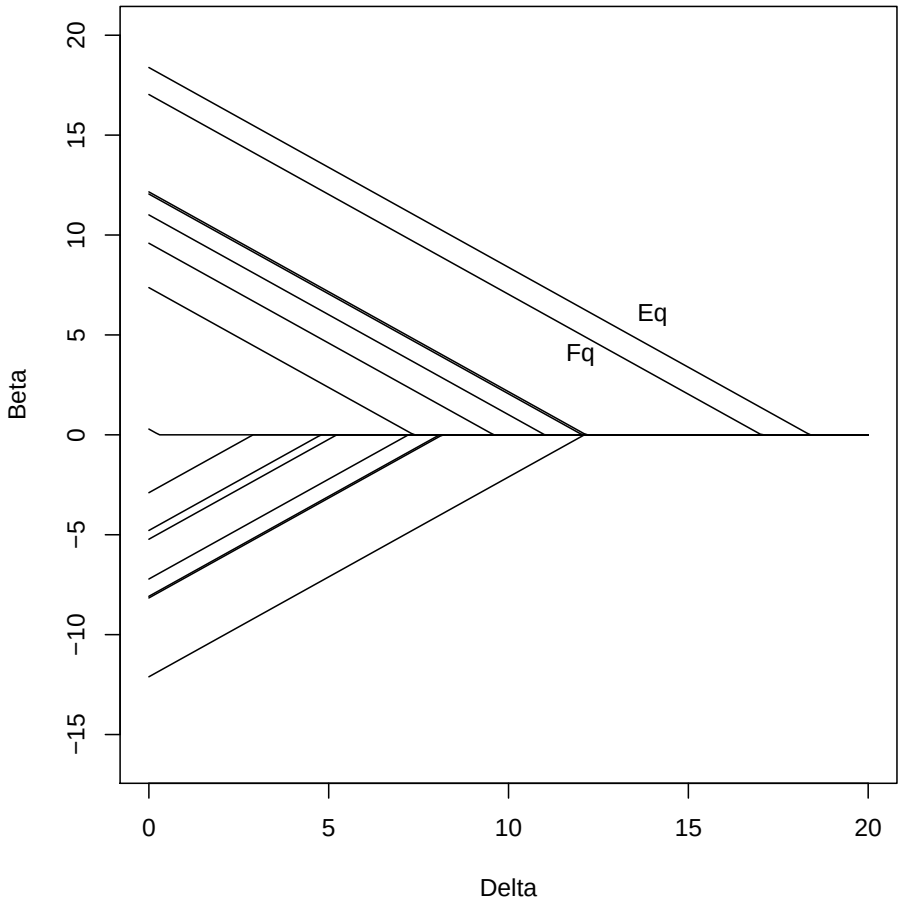


Figure 4.4: Profile plot for the blood glucose experiment with interactions. The model contains 15 main effects and 98 two-factor interaction effects.

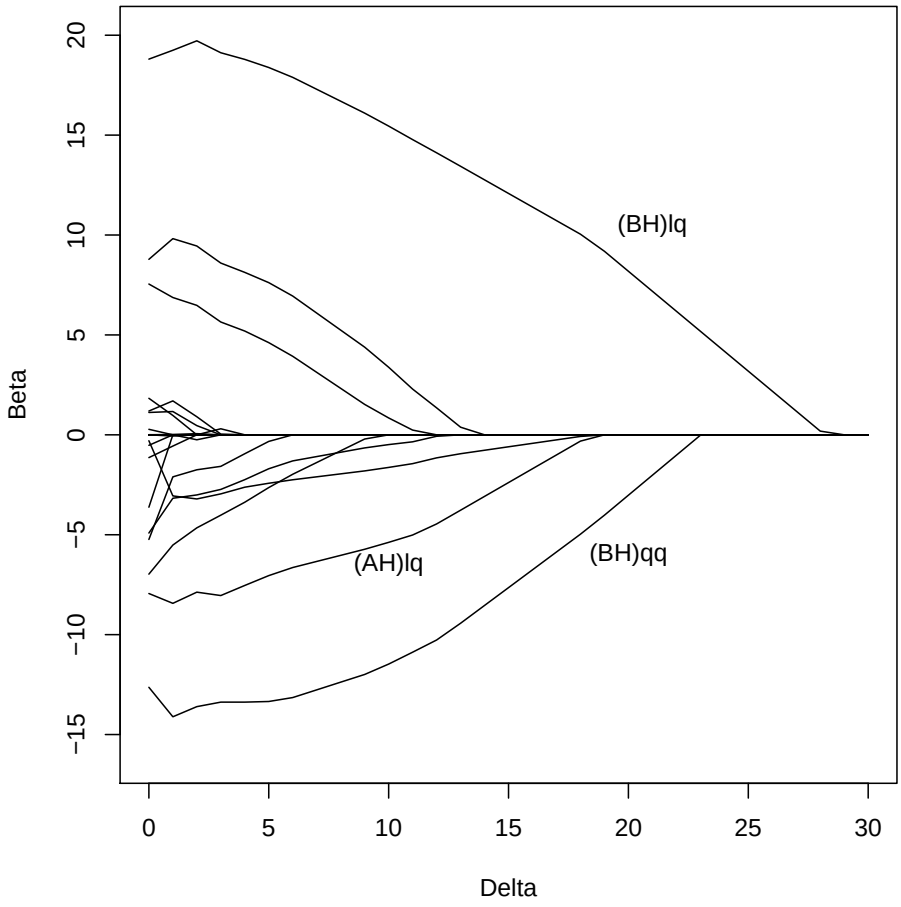


Figure 4.5: Profile plot for the Lin (1993) data. The model contains 23 main effects.

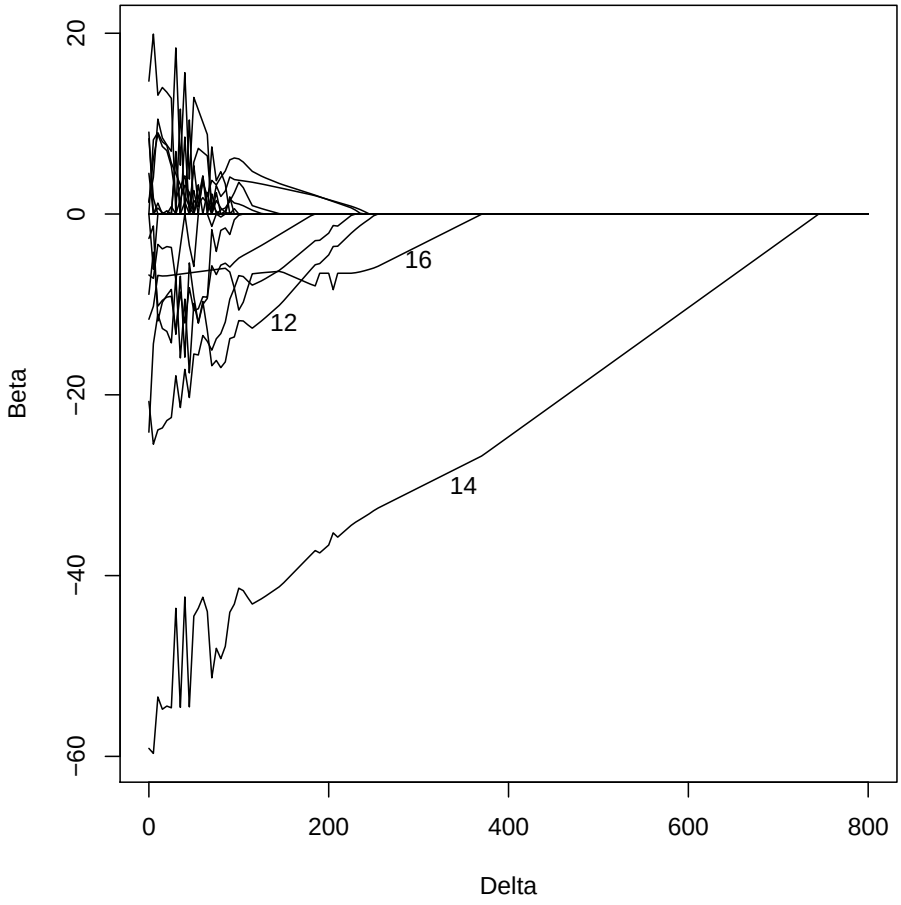


Table 4.3: A Two-level Supersaturated Design (Lin 1993).

	Factors																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Y
1	+	+	+	−	−	−	+	+	+	+	+	−	−	−	+	+	−	−	+	−	−	−	+	133
2	+	−	−	−	−	−	+	+	+	−	−	−	+	+	+	−	+	−	−	+	+	−	−	62
3	+	+	−	+	+	−	−	−	−	+	−	+	+	+	+	+	−	−	−	−	+	+	−	45
4	+	+	−	+	−	+	−	−	−	+	+	−	−	+	+	−	+	+	+	−	−	−	−	52
5	−	−	+	+	+	+	−	+	+	−	−	−	−	+	+	+	−	−	+	−	+	+	+	56
6	−	−	+	+	+	+	+	−	+	+	+	−	+	+	−	+	+	+	+	+	+	−	−	47
7	−	−	−	−	+	−	−	+	−	+	−	+	+	−	+	+	+	+	+	+	−	−	+	88
8	−	+	+	−	−	+	−	+	−	+	−	−	−	−	−	−	−	+	−	+	+	+	−	193
9	−	−	−	−	−	+	+	−	−	−	+	+	−	+	−	+	+	−	−	−	−	+	+	32
10	+	+	+	+	−	+	+	+	−	−	−	+	+	+	−	+	−	+	−	+	−	−	+	53
11	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	+	+	−	−	−	+	+	276
12	+	−	−	−	+	+	+	−	+	+	+	+	−	−	+	−	−	+	−	+	+	+	+	145
13	+	+	+	+	+	−	+	−	+	−	−	+	−	−	−	−	+	−	+	+	−	+	−	130
14	−	−	+	−	−	−	−	−	−	−	+	+	+	−	−	−	−	−	+	−	+	−	−	127

The same data were previously analyzed by several authors. Westfall et al. (1998) highlighted X_{14}, X_{12}, X_{19}, X_4, X_{10} and X_{11} as important, among which X_{14} is the only significant variable at 5% significance level and X_4 is marginally significant. Beattie et al. (2002) compared several model selection methods; only X_{14} is identified as important in every method. Both Li and Lin (2003) and Zhang et al. (2007) suggested X_{14}, X_{12}, X_{19} and X_4 as active factors.

The difference is not surprising when we look at the trajectories of $\hat{\beta}$ in Figure 4.5. Almost all effects, except X_{14}, are noisy and the magnitudes are small enough to be considered within the noise level. We agree with Abraham et al. (1999) that it is not clear the correct answers on which the active factors are. Different approaches may provide different answers on the list of active factors and X_{14} is probably the only common active factor found in different approaches.

4.3 Automatic Variable Selection

The preceding graphical procedure is simple and easy to use. Nevertheless, it is sometimes desirable, for instance in simulation, to have a procedure for choosing the tuning parameter δ automatically. Here we propose a general procedure for choosing δ based on a model selection criterion. First we obtain the Dantzig selector $\hat{\beta}$ for a range of δ. Then for a fixed $\gamma \geq 0$, we obtain a list of models $\hat{I} = \{i : |\hat{\beta}_i| > \gamma\}$, compare these models according to a criterion and choose a δ that yields the best model.

Akaike information criterion (AIC) is popular for model selection. For linear models, it is defined as

$$\text{AIC} = n \log(RSS/n) + 2p$$

where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares and p is the number of parameters in the model. It is known that AIC tends to overfit the model when the sample size is small. Hurvich and Tsai (1989) proposed a bias correction by adding an additional penalty term to AIC. Their modified AIC is defined as

$$\text{cAIC} = \text{AIC} + 2(p+1)(p+2)/(n-p-2).$$

The cAIC typically chooses a smaller model than AIC. However, it still tends to overfit the model for supersaturated designs.

Factor sparsity is an important assumption for the use of supersaturated designs and the Dantzig selector. Based on this assumption we impose a heavy penalty on the model complexity and propose a new modified AIC for supersaturated designs as follows:

$$\text{mAIC} = n \log(RSS/n) + 2p^2. \tag{4.4}$$

The difference between our modified AIC and AIC or cAIC is the penalty of model complexity p. The penalty on p in mAIC is quadratic whereas that in AIC is linear; therefore, mAIC chooses more parsimonious model than AIC. The penalty in cAIC is complicated. It is nearly quadratic on p when p is close to n and nearly linear when p is close to $n/2$. As will be seen later, this modification in Equation (4.4) works well for our examples and simulations. It remains to be seen whether it works for other situations.

The parameter γ can be viewed as a threshold between signal and noise and a relative small γ should be chosen. One can choose γ according to the profile plot or information on the magnitude of effects or noise. One can always repeat the procedure with a few choices of γ and compare the results. In the simulation, we do a coarse grid search and choose one. We find that the modified AIC defined in Equation (4.4) tends to produce more robust results against different choices of γ than AIC or cAIC.

Example 4. We illustrate the automatic procedure with the cast fatigue experiment in Example 1. We fix $\gamma = 0$ and choose δ according to the three information criteria. When entertaining the main effects only, AIC chooses a 2-factor model (F, D) while both cAIC and mAIC choose a 1-factor model (F). When entertaining the two-factor interactions, AIC chooses a model with 9 terms $(F, FG, AE, D, EF, AD, DG, A, AB)$, cAIC chooses a model with 5 terms (F, FG, AE, D, EF) and mAIC chooses a model with 2 terms (F, FG). Table 4.4 lists RSS, R^2, and the AIC, cAIC and mAIC values for these 5 models and an additional model with 3 terms (F, FG, AE). The model with 3 terms (F, FG, AE) has a slightly larger mAIC value than the model with 2 terms (F, FG). It is evident here that mAIC performs the best among the three criteria. The mAIC works well with other choices of γ. For instance, with $\gamma = 0.1$, both AIC and

Table 4.4: Comparison of information criteria in Example 4

Model	Terms	p	RSS	R^2	AIC	cAIC	mAIC
1	F	1	3.132	44.5%	-14.12	-12.79	-14.12
2	F, D	2	2.333	58.7%	-15.65	-12.65	-11.65
3	F, FG	2	0.6066	89.3%	-31.82	-28.82	-27.82
4	F, FG, AE	3	0.2673	95.3%	-39.65	-33.94	-27.65
5	F, FG, AE, D, EF	5	0.03568	99.4%	-59.82	-43.02	-19.82
6	F, FG, AE, D, EF	9	0.001167	99.98%	-92.86	127.14	51.14
	AD, DG, A, AB						

cAIC choose the 5-term model while mAIC still chooses the 2-term model when entertaining the two-factor interactions.

4.4 Simulations and Results

In this section, we investigate the performance of the Dantzig selector approach via simulation. Example 5 compares the performance of the Dantzig selector method with four different approaches suggested in the literature, and they are (i) SSVS, the Bayesian variable selection procedure proposed by George and McMulloch (1993) and extended for supersaturated designs by Chipman et al. (1997); (ii) SSVS/IBF, the two stage Bayesian procedure by Beattie et al. (2002); (iii) SCAD, the penalized least squares approach proposed by Li and Lin (2003); and (iv) PLSVS, the partial least square regression technique by Zhang et al. (2007). Our simulations are conducted in R using package "lpSolve".

Example 5. To compare the performance of the Dantzig selector method with

that of the four methods by simulation, we consider the same models as Li and Lin (2003) and Zhang et al. (2007). We generate data from the linear model

$$y = X\beta + \epsilon$$

where X is the 14×23 matrix given in Table 4.3 and the random error $\epsilon \sim N(0, 1)$. We consider the following three cases for β:

Case I: One active factor, $\beta_1 = 10$, and all other components of β equal 0;

Case II: Three active factors, $\beta_1 = -15$, β_5=8, $\beta_9 = -2$, and all other components of β equal 0;

Case III: Five active factors, $\beta_1 = -15$, β_5=12, $\beta_9 = -8$, $\beta_{13} = 6$, $\beta_{17} = -2$, and all other components of β equal 0.

We run the simulations 1,000 times by fixing $\gamma = 1$ and choosing δ automatically using mAIC. Table 4.5 compares the Dantzig selector method with the other four methods. In this table, "TMIR" stands for True Model Identified Rate, "SEIR" stands for Smallest Effect Identified Rate, and "Median" and "Mean" are the median and mean sizes of the models.

The Dantzig selector method identifies the true model with the highest probabilities among all five methods. In case I, the Dantzig selector shares 100% perfect identification rates with SCAD and PLSVS in identifying the smallest effect. In cases II and III, the probability of getting the smallest effect with the Dantzig selector method is less than that of SCAD and PLSVS. In terms of the model size, the Dantzig selector method performs the best. The average model size is closer to the true model size than those resulted from the other methods. In this sense our method is more efficient.

We also evaluate the performance of the Dantzig selector with different choices of γ and different criteria. Table 4.6 summarizes simulation results using AIC,

Table 4.5: Comparison of simulation results in Example 5

Case	Method	TMIR	SEIR	Median	Mean
I	SSVS(1/10,500)	40.5%	99.0%	2	3.1
	SSVS(1/10,500)/IBF	61.0%	98.0%	1	2.5
	SCAD	75.6%	100%	1	1.7
	PLSVS (m=1)	61.0%	100%	1	1.5
	Dantzig Selector ($\gamma = 1$)	99.4%	100%	1	1.0
II	SSVS(1/10,500)	8.6%	30.0%	3	4.7
	SSVS(1/10,500)/IBF	8.0%	28.0%	3	4.2
	SCAD	75.6%	98.5%	3	3.3
	PLSVS (m=1)	76.4%	100%	3	3.3
	Dantzig Selector ($\gamma = 1$)	84.4%	85.3%	3	2.9
III	SSVS(1/10,500)	36.4%	84.0%	6	8.0
	SSVS(1/10,500)/IBF	40.7%	75.0%	5	5.6
	SCAD	69.7%	99.4%	5	5.4
	PLSVS (m=1)	73.6%	95.0%	5	5.2
	Dantzig Selector ($\gamma = 1$)	79.1%	91.2%	5	5.1

Table 4.6: Summary of simulation results in Example 5

		TMIR			Average Size		
Case	γ	AIC	cAIC	mAIC	AIC	cAIC	mAIC
I	1.25	99.9%	99.9%	100%	1.001	1.001	1.000
	1.00	99.0%	99.1%	99.4%	1.010	1.009	1.006
	0.75	90.2%	91.3%	94.7%	1.105	1.091	1.054
	0.50	43.9%	50.7%	71.8%	1.843	1.701	1.310
II	1.25	68.9%	68.9%	69.1%	2.769	2.768	2.721
	1.00	79.8%	81.4%	84.4%	3.036	3.015	2.901
	0.75	64.8%	74.9%	85.6%	3.418	3.274	3.012
	0.50	21.3%	42.8%	69.6%	4.538	3.857	3.086
III	1.25	69.4%	79.8%	81.2%	5.263	5.037	4.967
	1.00	54.7%	77.1%	79.1%	5.709	5.263	5.143
	0.75	32.2%	59.9%	63.9%	6.342	5.550	5.372
	0.50	8.1%	32.2%	37.1%	7.573	6.131	5.743

cAIC and mAIC with $\gamma = 1.25$, 1.00, 0.75, and 0.50.

It is evident that mAIC performs the best and AIC performs the worst among all cases; mAIC produces the most stable and accurate results with different choices of γ.

In the next example, we randomly generate some models and evaluate the performance of the Dantzig selector via simulations.

Example 6. We consider five cases. There are i active factors for case i, $1 \leq i \leq 5$. For each case, we generate 500 models where the selection of active factors is random without replacement, the signs of the active factors are randomly selected

Table 4.7: Summary of simulation results in Example 6

Case		Min	1st Quartile	Median	Mean	3rd Quartile	Max
I	TMIR	96%	99%	100%	99.5%	100%	100%
	Size	1.00	1.00	1.00	1.005	1.01	1.04
II	TMIR	78%	99%	100%	99.3%	100%	100%
	Size	1.86	2.00	2.00	2.004	2.01	2.08
III	TMIR	0%	99%	100%	95.6%	100%	100%
	Size	2.22	3.00	3.00	3.001	3.01	3.83
IV	TMIR	0%	88%	98%	84.0%	100%	100%
	Size	1.36	3.97	4.00	3.850	4.01	4.94
V	TMIR	0%	8.8%	86%	64.0%	98%	100%
	Size	1.23	4.05	4.89	4.395	5.00	6.29

from either positive or negative, and the magnitudes are randomly selected from 2 to 10 with replacement. For each model, we generate data from the same linear model as in Example 5 for 100 times and obtain the true model identification rate (TMIR) and the average model size. Table 7 gives the summary statistics of these two quantities among 500 models. In the simulations we fix $\gamma = 1$ and choose δ according to mAIC.

The Dantzig selector method is very effective in identifying 1 active factor; the TIMR ranges from 96% to 100% and the average model size ranges from 1 to 1.04. The method is still effective in identifying 2 active factors; the TMIR ranges from 78% to 100% and the average model size ranges from 1.86 to 2.08. The method is less effective in identifying 3 or more active factors. The median TMIR values are still very high and the median model sizes are still accurate for 3 or 4 active factors. However, the minimum TMIR values are 0%, suggesting

that the method fails to identify a few models. The situation becomes worse for 5 active factors; although the median TMIR is 86%.

The Dantzig selector method does an excellent job in identifying one active factor in both simulations. This is supported by the theory developed by Candes and Tao (2007), which roughly says that the probability of correctly identifying one active factor is high when the correlations between the variables are small. However, their theory says nothing about the performance of the Dantzig selector when there are more than one active factors, because the supersaturated design does not meet the required uniform uncertainty condition. This partially explains why the method fails to identify some models with 3 or more active factors, where factor sparsity would be questionable with 23 factors and only 14 runs.

4.5 Concluding Remarks

This chapter studies the Dantzig selector for selecting active effects in supersaturated designs. We propose a graphical procedure and an automatic variable selection method to accompany with the Dantzig selector. The graphical procedure is recommended in practice and the automatic method, like other automatic methods, should be used with caution. Simulation shows that the Dantzig selector method performs well compared to existing methods in the literature and is more efficient at estimating the model size.

A modified AIC is proposed for model selection. It works well for the examples and simulations conducted here, but may not work well for other situations. Nevertheless, it demonstrates that supersaturated designs are useful when properly analyzed and that the Dantzig selector is a good tool.

The advantages of the Dantzig selector are as follows. First, the Dantzig

selector has a profound theory. Candes and Tao (2007) proved that the Dantzig selector is able to perform an ideal model selection when some uniform uncertainty conditions are fulfilled.

Second, the Dantzig selector is relatively fast, easy and simple to use. It is basically a linear program, which is widely considered as a fast and efficient algorithm to perform massive computation. Linear programming algorithms are available in many software and packages, like R, Matlab, Mathematica, etc., making it easy to program and use the Dantzig selector.

Third, the Dantzig selector is able to handle a large number of factors in two-level, multi-level and mixed-level experiments. Candes and Tao (2007) applied the Dantzig selector to an experiment with up to 200 active factors among 5,000 binary factors and 1,000 observations.

CHAPTER 5

Construction of Two-level Nonregular $(1/4)^{th}$-FFDs

In this chapter we consider the construction of two-level nonregular designs via quaternary codes. A quaternary code is a linear space over $Z_4 = \{0, 1, 2, 3\}$, the ring of integers modulo 4. Quaternary codes have been successfully used to construct good binary codes in coding theory; see Hammons et al. (1994). Xu and Wong (2007) first used quaternary codes to construct two-level nonregular designs. They described a systematic procedure for constructing $2^{2n} \times (2^{2n} - 2^n)$ designs and $2^{2n+1} \times (2^{2n+1} - 2^{n+1})$ designs with resolution 3.5 for any n whereas regular designs of the same size have maximum resolution 3 only. They also presented a collection of nonregular designs with 16, 32, 64, 128, and 256 runs and up to 64 factors. Two obvious advantages of using quaternary codes to construct nonregular designs is its relatively straightforward construction procedure and simple design presentation. Since the designs are constructed via linear codes over Z_4, one can use column indexes to describe these designs. More importantly, many nonregular designs constructed via quaternary codes have better statistical properties than regular designs of the same size in terms of resolution, aberration and projectivity.

The linear structure of a quaternary code makes it possible to study analytically the properties of nonregular designs derived from it. In section 5.2 we study

the properties of quarter-fraction designs which can be defined by a generator matrix that consists of an identity matrix and an additional column. It turns out that the resolution, wordlength and projectivity can be calculated in terms of the frequencies that numbers 1, 2 and 3 appear in the additional column. Applying these results we construct optimal quarter-fraction designs via quaternary codes under the maximum resolution, minimum aberration and maximum projectivity criteria in section 5.3. These designs are often better than regular designs of the same size in terms of the corresponding criterion. It is well known that a regular minimum aberration design has maximum resolution and maximum projectivity among all regular designs. However, different criteria can lead to different non-regular designs. It turns out that we can often, but not always, find a minimum aberration design that has maximum resolution among all possible quaternary code designs. A minimum aberration design has the same aberration as, and often a larger resolution and projectivity than, a regular minimum aberration design. A maximum projectivity design, often differs from a minimum aberration or a maximum resolution design, can have a much larger projectivity than a regular minimum aberration design. It is further shown that some of these designs have generalized minimum aberration and maximum projectivity among all possible designs. We present all the proofs in section 5.4.

5.1 Notations and Definitions

A two-level design D of N runs and m factors is represented by an $N \times m$ matrix, where each row corresponds to a run and each column to a factor, which takes on only two symbols, say -1 and $+1$. For $s = \{c_1, c_2, \ldots, c_k\}$, a subset of k columns of D, we define the *J-characteristics* of design D (Deng and Tang 1999, Tang 2001) in the same way as in Equation (2.1). It is evident that $|j_k(s; D)| \leq N$.

Following Cheng, Li, Ye (2004), we define the *aliasing index* as $\rho_k(s) = \rho_k(s; D) = |j_k(s; D)|/N$, which measures the amount of aliasing among the columns in s. It is obvious that $0 \leq \rho_k(s) \leq 1$. When $\rho_k(s) = 1$, the columns in s are fully aliased with each other and form a *complete word* of length k. When $0 < \rho_k(s) < 1$, the columns in s are partially aliased with each other and form a *partial word* of length k with aliasing index $\rho_k(s)$. A partial word with aliasing index 1 is a complete word. When $\rho_k(s) = 0$, the columns in s do not form a word.

Suppose that r is the smallest integer such that $\max_{|s|=r} \rho_r(s; D) > 0$, where the maximization is over all subsets of r columns of D. The *generalized resolution* (Deng and Tang 1999) of D is defined as the same way in Equation (2.2). For $k = 1, \ldots, m$, we define the elements of wordlength pattern as the same way in Equation (2.3). The vector ($A_1(D)$, $A_2(D)$, ..., $A_m(D)$) is called the generalized wordlength pattern. The *generalized minimum aberration* criterion (Xu and Wu 2001), also called minimum G_2-aberration (Tang and Deng 1999), is to sequentially minimize the components in the generalized wordlength pattern $A_1(D)$, $A_2(D)$, ..., $A_m(D)$. This means that if two designs have $A_k(D)$ to be the first non-equal component in the generalized wordlength pattern, a design with smaller $A_k(D)$ is preferred.

When restricted to regular designs, generalized resolution, generalized wordlength pattern and generalized minimum aberration reduce to the traditional resolution, wordlength pattern and minimum aberration, respectively. For simplicity, we use resolution, wordlength pattern and minimum aberration for both regular and nonregular designs.

A two-level design D is said to have *projectivity* p (Box and Tyssedal 1996) if every p-factor projection contains a complete 2^p factorial design, possibly with

some points replicated. It is evident that a regular design of resolution $R = r$ has projectivity $p = r - 1$. Deng and Tang (1999) showed that a design with resolution $R > r$ has projectivity $p \geq r$.

5.2 Properties of $(1/4)^{th}$-FFDs via Quaternary Codes

5.2.1 Quaternary codes and binary images

A quaternary code takes on values from $Z_4 = \{0, 1, 2, 3\}$ (mod 4). Let G be an $n \times k$ generator matrix over Z_4. All possible linear combinations of the rows in G over Z_4 form a quaternary linear code, denoted by C. The so called *Gray map*, which replaces each element in Z_4 with a pair of two symbols, transforms C into a binary code $D = \phi(C)$, called the binary image of C. For convenience, we use 1 and -1 for the two symbols, instead of the 0 and 1 convention for binary codes. Then the Gray map is defined as:

$$\phi : 0 \to (1, 1), \quad 1 \to (1, -1), \quad 2 \to (-1, -1), \quad 3 \to (-1, 1).$$

Note that C is a $2^{2n} \times k$ matrix over Z_4 and D is a binary $2^{2n} \times 2k$ matrix or a two-level design with 2^{2n} runs and $2k$ factors.

5.2.2 $(1/4)^{th}$-FFDs with Even Number of Factors

To construct quarter-fraction designs, consider an $n \times (n + 1)$ generator matrix $G = (v, I_n)$, where v is an $n \times 1$ column vector over Z_4 and I_n is an $n \times n$ identity matrix. Let D be the $2^{2n} \times (2n + 2)$ two-level design generated by G. It is easy to verify that the identity matrix I_n generates a full $2^{2n} \times 2n$ design; therefore, the property of D depends on the column v only. Throughout the chapter, for $i = 0, 1, 2, 3$, let f_i be the number of times that number i appears in column v.

Theorem 1 characterizes the number of words of D, their lengths and aliasing indexes in terms of the frequency f_i.

Theorem 1. *Consider an $n \times (n+1)$ generator matrix $G = (v, I_n)$. Define $k_1 = f_1 + 2f_2 + f_3 + 1$, $k_2 = 2f_1 + 2f_3 + 2$ and $\rho = 2^{-\lfloor (f_1+f_3)/2 \rfloor}$, where $\lfloor x \rfloor$ is the integer value of x. Then the two-level $2^{2n} \times (2n+2)$ design D generated by G has 1 complete word of length k_2 and $2/\rho^2$ partial words of length k_1 with aliasing index ρ.*

Example 7. Consider a generator matrix

$$G = \begin{pmatrix} v & I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix}.$$

All linear combinations of the three rows of G form a 64×4 linear code C over Z_4. Applying the Gray map, a 64×8 binary image $D = \phi(C)$ is obtained; see Table 5.1.

According to Theorem 1, design D has 1 complete word of length $k_2 = 6$ and 8 partial words of length $k_1 = 5$ with aliasing index $\rho = 0.5$. It is easy to verify that the first six columns form a complete word and that columns $(a, b, c, 7, 8)$ form a partial word with aliasing index 0.5, where $a = 1$ or 2, $b = 3$ or 4, and $c = 5$ or 6. Therefore, by definitions Equation (2.2) and Equation (2.3), the resolution of D is 5.5, and the wordlength pattern of D is $A_5(D) = 2$, $A_6(D) = 1$ and $A_i(D) = 0$ for $i \neq 5, 6$.

For ease of presentation, we say that the ith identity column of I_n in $G = (v, I_n)$ is "associated with" number z if the ith element of v is z, where $z = 0, 1, 2,$ or 3. We also refer to a column of D as associated with number z if it is one of the two columns generated by an identity column that is associated with number z.

Table 5.1: A Quaternary Code C and its Binary Image D

Code C				
Run	1	2	3	4
1	0	0	0	0
2	1	1	0	0
3	2	2	0	0
4	3	3	0	0
5	1	0	1	0
6	2	1	1	0
7	3	2	1	0
8	0	3	1	0
9	2	0	2	0
10	3	1	2	0
11	0	2	2	0
12	1	3	2	0
13	3	0	3	0
14	0	1	3	0
15	1	2	3	0
16	2	3	3	0
17	2	0	0	1
18	3	1	0	1
19	0	2	0	1
20	1	3	0	1
21	3	0	1	1
22	0	1	1	1
23	1	2	1	1
24	2	3	1	1
25	0	0	2	1
26	1	1	2	1
27	2	2	2	1
28	3	3	2	1
29	1	0	3	1
30	2	1	3	1
31	3	2	3	1
32	0	3	3	1
33	0	0	0	2
34	1	1	0	2
35	2	2	0	2
36	3	3	0	2
37	1	0	1	2
38	2	1	1	2
39	3	2	1	2
40	0	3	1	2
41	2	0	2	2
42	3	1	2	2
43	0	2	2	2
44	1	3	2	2
45	3	0	3	2
46	0	1	3	2
47	1	2	3	2
48	2	3	3	2
49	2	0	0	3
50	3	1	0	3
51	0	2	0	3
52	1	3	0	3
53	3	0	1	3
54	0	1	1	3
55	1	2	1	3
56	2	3	1	3
57	0	0	2	3
58	1	1	2	3
59	2	2	2	3
60	3	3	2	3
61	1	0	3	3
62	2	1	3	3
63	3	2	3	3
64	0	3	3	3

Design D								
Run	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	1	−1	1	−1	1	1	1	1
3	−1	−1	−1	−1	1	1	1	1
4	−1	1	−1	1	1	1	1	1
5	1	−1	1	1	1	−1	1	1
6	−1	−1	1	−1	1	−1	1	1
7	−1	1	−1	−1	1	−1	1	1
8	1	1	−1	1	1	−1	1	1
9	−1	−1	1	1	−1	−1	1	1
10	−1	1	1	−1	−1	−1	1	1
11	1	1	−1	−1	−1	−1	1	1
12	1	−1	−1	1	−1	−1	1	1
13	−1	1	1	1	−1	1	1	1
14	1	1	1	−1	−1	1	1	1
15	1	−1	−1	−1	−1	1	1	1
16	−1	−1	−1	1	−1	1	1	1
17	−1	−1	1	1	1	1	1	−1
18	−1	1	1	−1	1	1	1	−1
19	1	1	−1	−1	1	1	1	−1
20	1	−1	−1	1	1	1	1	−1
21	−1	1	1	1	1	−1	1	−1
22	1	1	1	−1	1	−1	1	−1
23	1	−1	−1	−1	1	−1	1	−1
24	−1	−1	−1	1	1	−1	1	−1
25	1	1	1	1	−1	−1	1	−1
26	1	−1	1	−1	−1	−1	1	−1
27	−1	−1	−1	−1	−1	−1	1	−1
28	−1	1	−1	1	−1	−1	1	−1
29	1	−1	1	1	−1	1	1	−1
30	−1	−1	1	−1	−1	1	1	−1
31	−1	1	−1	−1	−1	1	1	−1
32	1	1	−1	1	−1	1	1	−1
33	1	1	1	1	1	1	−1	−1
34	1	−1	1	−1	1	1	−1	−1
35	−1	−1	−1	−1	1	1	−1	−1
36	−1	1	−1	1	1	1	−1	−1
37	1	−1	1	1	1	−1	−1	−1
38	−1	−1	1	−1	1	−1	−1	−1
39	−1	1	−1	−1	1	−1	−1	−1
40	1	1	−1	1	1	−1	−1	−1
41	−1	−1	1	1	−1	−1	−1	−1
42	−1	1	1	−1	−1	−1	−1	−1
43	1	1	−1	−1	−1	−1	−1	−1
44	1	−1	−1	1	−1	−1	−1	−1
45	−1	1	1	1	−1	1	−1	−1
46	1	1	1	−1	−1	1	−1	−1
47	1	−1	−1	−1	−1	1	−1	−1
48	−1	−1	−1	1	−1	1	−1	−1
49	−1	−1	1	1	1	1	−1	1
50	−1	1	1	−1	1	1	−1	1
51	1	1	−1	−1	1	1	−1	1
52	1	−1	−1	1	1	1	−1	1
53	−1	1	1	1	1	−1	−1	1
54	1	1	1	−1	1	−1	−1	1
55	1	−1	−1	−1	1	−1	−1	1
56	−1	−1	−1	1	1	−1	−1	1
57	1	1	1	1	−1	−1	−1	1
58	1	−1	1	−1	−1	−1	−1	1
59	−1	−1	−1	−1	−1	−1	−1	1
60	−1	1	−1	1	−1	−1	−1	1
61	1	−1	1	1	−1	1	−1	1
62	−1	−1	1	−1	−1	1	−1	1
63	−1	1	−1	−1	−1	1	−1	1
64	1	1	−1	1	−1	1	−1	1

Further, we refer to the two columns generated by v as associated with vector v. For example, the first two columns of D in Table 5.1 are associated with vector v, columns 3 to 6 are associated with number 1, and the last two columns are associated with number 2.

Now we can describe more precisely about the words of D in Theorem 1. The complete word of D consists of all the columns associated with vector v, numbers 1 and 3. Each partial word consists of all columns associated with number 2, one of the columns associated with vector v, each number 1 and each number 3. Furthermore, the columns associated with number 0 do not appear in any word.

Recall that a regular design has only complete words. Corollary 1 provides a sufficient and necessary condition for D to be a regular design.

Corollary 1. *Design D is regular if and only if $f_1 + f_3 \leq 1$.*

It is straightforward to complete the resolution of D according to the definition Equation (2.2) and Theorem 1.

Corollary 2. *The resolution of D is k_2 if $k_1 \geq k_2$ or $k_1 + 1 - \rho$ otherwise.*

According to the definition Equation (2.3), when $2/\rho^2$ partial words of length k_1 with aliasing index ρ are summed up, we get $A_{k_1}(D) = 2$. Corollary 3 specifies the wordlength pattern of D.

Corollary 3. *The wordlength pattern of D is:*

(a) *If $k_1 \neq k_2$, then $A_{k_1}(D) = 2$, $A_{k_2}(D) = 1$ and $A_i(D) = 0$ for $i \neq k_1, k_2$.*

(b) *If $k_1 = k_2 = k$, then $A_k(D) = 3$ and $A_i(D) = 0$ for $i \neq k$.*

Next we consider the projectivity of design D generated by $G = (v, I_n)$. Theorem 1 suggests that there is a complete word of length $k_2 = 2(f_1 + f_3) + 2$.

This implies that the projectivity of D is at most $2(f_1+f_3)+1$. The next theorem states that the projectivity of D is indeed $2(f_1+f_3)+1$ if $f_2>0$.

Theorem 2. *Suppose that D is the two-level $2^{2n}\times(2n+2)$ design generated by $G=(v,I_n)$.*

(a) *If $f_2>0$, the projectivity of D is $2(f_1+f_3)+1$.*

(b) *If $f_2=0$ and $f_1+f_3>0$, the projectivity of D is $2(f_1+f_3)-1$.*

Theorem 2 implies that the projectivity of D is not affected by the partial words. As an example, consider design D in Example 1. Theorem 2 suggests that the projectivity of D is 5. This can be verified directly.

5.2.3 $(1/4)^{th}$-FFDs with Odd Number of Factors

Consider a design D generated by $G=(v,I_n)$ has 2^{2n} runs and $2n+2$ factors. To construct quarter-fraction designs with 2^{2n-1} runs, we use the half fraction method, which works as follows. Choose any column of D as a branching column, which divides D into two half-fractions according to the symbols of the branching column. Deleting the branching column yields two $2^{2n-1}\times(2n+1)$ designs. It is easy to verify that the two half-fractions of D are equivalent.

However, the properties of the half-fractions depend on the branching column, which are characterized in Theorem 3.

Theorem 3. *Suppose that D is the two-level $2^{2n}\times(2n+2)$ design generated by $G=(v,I_n)$ and that D' is a half-fraction of D. Define k_1, k_2 and ρ as in Theorem 1.*

(a) *If the branching column is associated with number 1 or 3, D' has 1 complete word of length k_2-1, $1/\rho^2$ partial words of length k_1 with aliasing index ρ and $1/\rho^2$ partial words of length k_1-1 with aliasing index ρ.*

(b) *If the branching column is associated with number* 2, D' *has* 1 *complete word of length* k_2 *and* $2/\rho^2$ *partial words of length* $k_1 - 1$ *with aliasing index* ρ.

It is easy to verify that if the branching column is associated with vector v, this is identical to case (a) when $f_1 + f_3 > 0$ or case (b) when $f_1 + f_3 = 0$ and $f_2 > 0$. If the branching column is associated with number 0, D' and D share the same words because the branching column does not appear in any word of D.

The following four corollaries summarize the resolution and wordlength pattern of D' for cases (a) and (b) separately.

Corollary 4. *The resolution of* D' *derived in Theorem 3(a) is* $k_2 - 1$ *if* $k_1 \geq k_2$ *or* $k_1 - \rho$ *otherwise.*

Corollary 5. *The wordlength pattern of* D' *derived in Theorem 3(a) is*

(a) *If* $k_1 = k_2 = k$, *then* $A_{k-1}(D') = 2$, $A_k(D') = 1$ *and* $A_i(D') = 0$ *for* $i \neq k - 1, k$.

(b) *If* $k_1 = k_2 - 1 = k$, *then* $A_{k-1}(D') = 1$, $A_k(D') = 2$ *and* $A_i(D') = 0$ *for* $i \neq k - 1, k$.

(c) *If* $k_1 \neq k_2$ *or* $k_2 - 1$, *then* $A_{k_1-1}(D') = A_{k_2-1}(D') = A_{k_1}(D') = 1$ *and* $A_i(D') = 0$ *for* $i \neq k_1 - 1, k_1, k_2 - 1$.

Corollary 6. *The resolution of* D' *derived in Theorem 3(b) is* k_2 *if* $k_1 - 1 \geq k_2$ *or* $k_1 - \rho$ *otherwise.*

Corollary 7. *The wordlength pattern of* D' *derived in Theorem 3(b) is*

(a) *If* $k_1 - 1 \neq k_2$, *then* $A_{k_1-1}(D') = 2$, $A_{k_2}(D') = 1$ *and* $A_i(D') = 0$ *for* $i \neq k_1 - 1, k_2$.

(b) *If* $k_1 - 1 = k_2 = k$, *then* $A_k(D') = 3$ *and* $A_i(D') = 0$ *for* $i \neq k$.

The next theorem summarizes the projectivity of a half-fraction of D.

Theorem 4. *Suppose that D is the two-level $2^{2n} \times (2n+2)$ design generated by $G = (v, I_n)$ and that D' is a half-fraction of D.*

(a) *If $f_2 > 0$, $f_1 + f_3 > 0$ and the branching column is associated with number 1 or 3, the projectivity of D' is $2(f_1 + f_3)$.*

(b) *If $f_2 = 0$, $f_1 + f_3 > 0$ and the branching column is associated with number 1 or 3, the projectivity of D' is $2(f_1 + f_3) - 2$.*

(c) *If $f_2 > 1$ and the branching column is associated with number 2, the projectivity of D' is $2(f_1 + f_3) + 1$.*

(d) *If $f_2 = 1$ and the branching column is associated with number 2, the projectivity of D' is $2(f_1 + f_3)$.*

Comparing with Theorem 2, we observe that the projectivity of D' equals to the projectivity of D for case (c) whereas the projectivity of D' equals to the projectivity of D minus one for all other cases.

Example 8. Consider half-fractions of D in Table 5.1. If one of the first six columns is chosen as the branching column, we obtain a 32×7 design D' with resolution 4.5 and wordlength pattern $A_4(D') = 1$, $A_5(D') = 2$ and $A_i(D') = 0$ for $i \neq 4, 5$. Design D' has 1 complete word of length 5, 4 partial words of length 5 with aliasing index 0.5 and 4 partial words of length 4 with aliasing index 0.5. For example, if the first column is chosen as the branching column, then columns 2 to 6 form a complete word and columns $(b, c, 7, 8)$ and $(2, b, c, 7, 8)$ form a partial word with aliasing index 0.5, where $b = 3$ or 4 and $c = 5$ or 6. If one of the last two columns is chosen as the branching column, we obtain a 32×7 design D' with resolution 4.5 and wordlength pattern $A_4(D') = 2$, $A_6(D') = 1$ and $A_i(D') = 0$ for $i \neq 4, 6$. Design D' has 1 complete word of length 6 and 8 partial words of length 4 with aliasing index 0.5. Finally, according to Theorem

4, any half-fraction of D has projectivity 4, which can be verified directly.

5.3 Structure of the Best $(1/4)^{th}$-FFDs

In this section we apply the theorems developed in the section 5.2 to construct the best designs under the maximum resolution, minimum aberration and maximum projectivity criteria. As shown below, different criteria can lead to different best designs.

5.3.1 $(1/4)^{th}$-FFDs with Even Number of Factors

Applying Theorem 1, we have the following results regarding maximum resolution and minimum aberration designs.

Theorem 5. *Among all* $2^{2n} \times (2n+2)$ *designs generated by* $G = (v, I_n)$,

(a) *if* $n = 3k-1, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k-1$ *and* $f_2 = k$ *has maximum resolution* $4k$.

(b) *if* $n = 3k, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k$ *and* $f_2 = k$ *has maximum resolution* $4k + 2 - 2^{-k}$.

(c) *if* $n = 3k+1, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k+1$ *and* $f_2 = k$ *has maximum resolution* $4k + 3 - 2^{-k}$.

Theorem 6. *Among all* $2^{2n} \times (2n+2)$ *designs generated by* $G = (v, I_n)$,

(a) *if* $n = 3k-1, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k-1$ *and* $f_2 = k$ *has minimum aberration and its wordlength pattern is* $A_{4k}(D) = 3$.

(b) *if* $n = 3k, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k$ *and* $f_2 = k$ *has minimum aberration and its wordlength pattern is* $A_{4k+1}(D) = 2$ *and* $A_{4k+2}(D) = 1$.

(c) *if* $n = 3k+1, k \geq 1$, *then a design* D *defined by* $f_1 + f_3 = 2k$ *and* $f_2 = k+1$ *has minimum aberration and its wordlength pattern is* $A_{4k+2}(D) = 1$ *and* $A_{4k+3}(D) = 2$.

When $n = 3k-1$ or $3k$, the minimum aberration design in Theorem 6 coincides with the maximum resolution design in Theorem 5; however, when $n = 3k+1$, the minimum aberration design differs from the maximum resolution design.

Applying Theorem 2, we have the following result regarding maximum projectivity designs.

Theorem 7. *Among all* $2^{2n} \times (2n+2)$ *designs generated by* $G = (v, I_n)$, *a design* D *defined by* $f_1 + f_3 = n-1$ *and* $f_2 = 1$ *has maximum projectivity* $2n-1$, *so does a design* D *defined by* $f_1 + f_3 = n$ *and* $f_2 = 0$.

The maximum projectivity designs in Theorem 7 are different from designs in Theorems 5 and 6 when $n > 4$. According to Corollary 2, a design defined by $f_1 + f_3 = n-1$ and $f_2 = 1$ has resolution $n + 3 - 2^{-\lfloor (n-1)/2 \rfloor}$ for $n \geq 2$ and a design defined by $f_1 + f_3 = n$ and $f_2 = 0$ has resolution $n + 2 - 2^{-\lfloor n/2 \rfloor}$; therefore, the former design is recommended.

5.3.2 $(1/4)^{th}$-FFDs with Odd Number of Factors

To find optimal designs with 2^{2n-1} runs, we consider all possible designs generated by $G = (v, I_n)$ and all possible half-fractions. It turns out that it is sufficient to consider only half-fractions of the minimum aberration designs in Theorem 6 and the maximum projectivity designs in Theorem 7.

Theorem 8. *Suppose that* D' *is a half-fraction of a design* D *given in Theorem 6. Among all* $2^{2n-1} \times (2n+1)$ *designs that are half-fractions of designs generated by* $G = (v, I_n)$, D' *has maximum resolution and minimum aberration*

(a) *if* $n = 3k-1, k \geq 1$, *and the branching column is associated with number* 2. *The resolution of* D' *is* $4k - 2^{-(k-1)}$ *and the wordlength pattern is* $A_{4k-1}(D') = 2$ *and* $A_{4k}(D') = 1$.

(b) *if* $n = 3k, k \geq 1$, *and the branching column is associated with number* 1. *The resolution of* D' *is* $4k + 1 - 2^{-k}$ *and the wordlength pattern is* $A_{4k}(D') = 1$ *and* $A_{4k+1}(D') = 2$.

(c) *if* $n = 3k + 1, k \geq 1$, *and the branching column is associated with number* 2. *The resolution of* D' *is* $4k + 2$ *and the wordlength pattern is* $A_{4k+2}(D') = 3$.

Theorem 9. *Any half-fraction of a design* D *in Theorem 7 has maximum projectivity* $2n - 2$ *among all* $2^{2n-1} \times (2n + 1)$ *designs that are half-fractions of designs generated by* $G = (v, I_n)$.

5.3.3 Table of the Best $(1/4)^{th}$-FFDs

For easy reference, we provide some optimal designs and their properties in Table 5.2. Following the convention on regular designs, we use the notation 2^{m-2} to represent a quarter-fraction design with m factors and 2^{m-2} runs. The second column of Table 5.2 specifies the three optimality criteria: maximum resolution (r), minimum aberration (a) and maximum projectivity (p). The third column is the vector v in the generator matrix $G = (v, I_n)$ and the letter at the end denotes the branching column, which is either the first (f) or last (l) column. The first column is associated with vector v while the last column is associated with number 2. Choosing the first column or a column associated with number 1 as the branching column yields an equivalent design. The next three columns, under the category of "quaternary-code designs", are the wordlength pattern (WLP), resolution (R) and projectivity (pr) of the design generated by $G = (v, I_n)$. The last two columns, under the category of "regular", are the resolution

and projectivity of a regular minimum aberration design with the same size.

Table 5.2 shows that the maximum resolution designs and the minimum aberration designs are similar but they often differ from the maximum projectivity designs. Specifically, the "r" design coincides with the "a" design when $m \neq 6k+4$, $k > 0$, whereas the "p" design differs from the "r" or "a" design when $m = 9$ or $m > 10$.

According to Corollary 1, all designs in Table 5.2 are nonregular designs except for design 2^{6-2}, which is equivalent to the regular minimum aberration design. Design 2^{8-2} is considered in Example 7 and given explicitly in Table 5.1. Design 2^{7-2} is a half-fraction of design 2^{8-2} and illustrated in Example 8.

It is of great interest to compare the quaternary-code designs with regular minimum aberration 2^{m-2} designs, which were given by Chen and Wu (1991). A regular minimum aberration 2^{m-2} design has resolution $R = \lfloor 2m/3 \rfloor$, projectivity $R - 1$ and wordlength pattern $A_R = 3R - 2m + 3$ and $A_{R+1} = 2m - 3R$. All of the "r" designs in Table 5.2 have the same or larger resolution than regular minimum aberration designs; in particular, when $m = 3k+1$ or $3k+2$, all of the "r" designs have larger resolution and therefore larger projectivity. All of the "a" designs have the same wordlength pattern as regular minimum aberration designs and have the same or larger resolution and projectivity. Indeed, Xu (2005) showed that regular minimum aberration 2^{m-2} designs have minimum aberration among all possible designs. Except for design 2^{6-2}, all of the "p" designs have higher projectivity than regular minimum aberration designs, but they may have smaller resolution. Indeed all of the "p" designs have maximum projectivity among all possible designs. The next theorem summarizes these results.

Theorem 10. (a) *The designs given in Theorems 6 and 8 have minimum aberration among all possible designs.*

(b) *The designs given in Theorems 7 and 9 have maximum projectivity among all possible designs.*

It is of interest to know whether the designs given in Theorems 5 and 8 have maximum resolution among all possible designs. We do not have an answer yet. The compete catalogs of Sun et al. (2002) and Bulutoglu and Margot (2008) suggest that designs 2^{6-2}, 2^{7-2} and 2^{8-2} given in Table 5.2 have maximum resolution among all possible designs. This can also be verified analytically using Proposition 2 of Deng and Tang (1999).

5.4 Proofs

In this section, we are going to prove Theorems 1 to 10, which are the theorems related to quarter-fraction designs constructed via quaternary codes. Some lemmas are introduced in order to prove those theorems.

5.4.1 Some Lemmas for $(1/4)^{th}$-FFDs

Consider an $n\times(n+1)$ generator matrix $G_n = (v_n, I_n)$, where v_n is an $n\times 1$ column vector over Z_4 and I_n is an $n \times n$ identity matrix. Let D_n be the $2^{2n} \times (2n + 2)$ binary design generated by G_n.

Let v_{n-1} be the vector consisting of the first $n - 1$ components of v_n and let D_{n-1} be the $2^{2n-2} \times 2n$ binary design generated by the $(n - 1) \times n$ generator matrix $G_{n-1} = (v_{n-1}, I_{n-1})$. Denote $D_{n-1} = (a, b, E)$ where a and b are column vectors generated by v_{n-1} and E is a $2^{(2n-2)} \times (2n - 2)$ full factorial generated by I_{n-1}.

We can express D_n in terms of D_{n-1} depending on the last component of v_n,

denoted by z. It is trivial for $z = 0$. It is obvious that $z = 1$ and $z = 3$ produce an equivalent design. Therefore, it is sufficient to consider only $z = 1$ or 2.

When $z = 1$, D_n can be expressed as follows, up to row permutations,

$$D_n = \begin{pmatrix} a & b & E & \mathbf{1} & \mathbf{1} \\ b & -a & E & \mathbf{1} & -\mathbf{1} \\ -a & -b & E & -\mathbf{1} & -\mathbf{1} \\ -b & a & E & -\mathbf{1} & \mathbf{1} \end{pmatrix}, \tag{5.1}$$

where $\mathbf{1}$ is a vector of ones. From this expression and the definition Equation (2.1), we establish the connection between the J-characteristics of D_n and D_{n-1}. Note that the column indexes of D_n are $\{1, 2, \ldots, 2n+2\}$ and that of D_{n-1} are $\{1, 2, \ldots, 2n\}$. For clarify, the s in the notation $j_k(s; D)$ refers to a subset of column indexes of D and we omit k when it is not important.

Lemma 1. *Suppose that the last component of v_n is 1. For any subset $e \subset \{3, 4, \ldots, 2n\}$,*

(a) $j(\{1, 2n+1\} \cup e; D_n) = j(\{2, 2n+2\} \cup e; D_n) = 2j(\{1\} \cup e; D_{n-1}) + 2j(\{2\} \cup e; D_{n-1})$;

(b) $j(\{1, 2n+2\} \cup e; D_n) = -j(\{2, 2n+1\} \cup e; D_n) = 2j(\{1\} \cup e; D_{n-1}) - 2j(\{2\} \cup e; D_{n-1})$;

(c) $j(\{1, 2, 2n+1, 2n+2\} \cup e; D_n) = 4j(\{1, 2\} \cup e; D_{n-1})$;

(d) $j(s \cup e; D_n) = 0$ *for* $s = \{1\}, \{2\}, \{2n+1\}, \{2n+2\}, \{1, 2\}, \{2n+1, 2n+2\}, \{1, 2, 2n+1\}, \{1, 2, 2n+2\}, \{1, 2n+1, 2n+2\},$ *or* $\{2, 2n+1, 2n+2\}$.

When $z = 2$, D_n can be expressed as follows, up to row permutations,

$$D_n = \begin{pmatrix} a & b & E & \mathbf{1} & \mathbf{1} \\ -a & -b & E & \mathbf{1} & -\mathbf{1} \\ a & b & E & -\mathbf{1} & -\mathbf{1} \\ -a & -b & E & -\mathbf{1} & \mathbf{1} \end{pmatrix} \tag{5.2}$$

From this expression and the definition Equation (2.1), we establish the connection between the J-characteristics of D_n and D_{n-1}.

Lemma 2. *Suppose that the last component of v_n is 2. For any subset $e \subset \{3, 4, \ldots, 2n\}$,*

(a) $j(\{1, 2n+1, 2n+2\} \cup e; D_n) = 4j(\{1\} \cup e; D_{n-1})$*;*

(b) $j(\{2, 2n+1, 2n+2\} \cup e; D_n) = 4j(\{2\} \cup e; D_{n-1})$*;*

(c) $j(\{1, 2\} \cup e; D_n) = 4j(\{1, 2\} \cup e; D_{n-1})$*;*

(d) $j(s \cup e; D_n) = 0$ *for* $s = \{1\}, \{2\}, \{2n+1\}, \{2n+2\}, \{1, 2n+1\}, \{1, 2n+2\}, \{2, 2n+1\}, \{2, 2n+2\}, \{2n+1, 2n+2\}, \{1, 2, 2n+1\}, \{1, 2, 2n+2\}$, *or* $\{1, 2, 2n+1, 2n+2\}$.

The next result describes the partial words of D_n and their J-characteristics.

Lemma 3. *Suppose that v_n is a vector of n 1s. For $l = 1, 2$, let $s_l = \{l, x_2, \ldots, x_{n+1}\}$ where $x_i = 2i - 1$ or $2i$ for $i = 2, \ldots, n+1$.*

(a) *If $n = 2t + 1$, either $j_{n+1}(s_1; D_n) = 0$ and $|j_{n+1}(s_2; D_n)| = 2^{3t+2}$ or $|j_{n+1}(s_1; D_n)| = 2^{3t+2}$ and $j_{n+1}(s_2; D_n) = 0$.*

(b) *If $n = 2t$, $|j_{n+1}(s_1; D_n)| = |j_{n+1}(s_2; D_n)| = 2^{3t}$.*

Proof. We prove the lemma by induction. It is trivial to verify that the lemma holds for $n = 1, 2$. Assume the lemma holds for $n = k - 1$. Consider $n = k$.

We have $s_1 = \{1, x_{k+1}\} \cup e$ and $s_2 = \{2, x_{k+1}\} \cup e$, where $e \subset \{x_2, \ldots, x_k\}$ with $x_i = 2i - 1$ or $2i$ for $i = 2, \ldots, k$.

First consider $x_{k+1} = 2k + 1$. By Lemmas 1(a) and 1(b),

$$j_{k+1}(s_1; D_k) = 2j_k(\{1\} \cup e; D_{k-1}) + 2j_k(\{2\} \cup e; D_{k-1}), \quad (5.3)$$

$$-j_{k+1}(s_2; D_k) = 2j_k(\{1\} \cup e; D_{k-1}) - 2j_k(\{2\} \cup e; D_{k-1}), \quad (5.4)$$

where D_{k-1} is the $2^{2k-2} \times 2k$ design generated by $G_{k-1} = (\mathbf{1}, I_{k-1})$.

If $n = k = 2t+1$, the assertion of $k-1 = 2t$ implies that $|j_k(\{1\} \cup e; D_{k-1})| = |j_k(\{2\} \cup e; D_{k-1})| = 2^{3t}$. Then from Equation (5.3) and Equation (5.4), we conclude that either $|j_{k+1}(s_1; D_k)|$ or $|j_{k+1}(s_2; D_k)|$ must be 0 and the other must be 2^{3t+2}.

If $n = k = 2t + 2$, the assertion of $k - 1 = 2t + 1$ implies that either $|j_k(\{1\} \cup e; D_{k-1})|$ or $|j_k(\{2\} \cup e; D_{k-1})|$ must be 0 and the other must be 2^{3t+2}. Then Equation (5.3) and Equation (5.4) together yield $|j_{k+1}(s_1; D_k)| = |j_{k+1}(s_2; D_k)| = 2^{3t+3}$. This proves the results for $x_{k+1} = 2k + 1$.

The proof for $x_{k+1} = 2k + 2$ is similar. Therefore, the lemma holds for $n = k$. The proof is completed by induction. □

The next result describes the complete and partial words of D_n and their aliasing indexes.

Lemma 4. *Suppose that v_n consists of p 1s followed by q 2s, where $p + q = n$. For $l = 1, 2$, let $s_l = \{l, x_2, \ldots, x_{p+1}, 2p + 3, 2p + 4, \ldots, 2n + 2\}$ where $x_i = 2i - 1$ or $2i$ for $i = 2, \ldots, p + 1$.*

(a) *If $p = 2t + 1$, either $\rho_k(s_1; D_n)$ or $\rho_k(s_2; D_n)$ is 0 and the other is 2^{-t} where $k = p + 2q + 1$.*

(b) *If $p = 2t$, $\rho_k(s_1; D_n) = \rho_k(s_2; D_n) = 2^{-t}$ where $k = p + 2q + 1$.*

(c) $\rho_k(s_0; D_n) = 1$ *where* $s_0 = \{1, 2, \ldots, 2p+2\}$ *and* $k = 2p+2$.

(d) $\rho_k(s; D_n) = 0$ *for* s *other than* s_1, s_2 *or* s_0 *considered in* (a), (b) *and* (c).

Proof. (a) and (b), when $q = 0$, it follows from Lemma 3. When $q > 0$, recursively applying Lemmas 2(a) or 2(b) yields the result.

(c) It follows from Lemmas 1(c) and 2(c).

(d) It follows from Lemmas 1(d) and 2(d). □

Now consider half-fractions of D_n. Suppose that one of the last two columns of D_n is chosen as the branching column. Let D'_n be the resulting $2^{2n-1} \times (2n+1)$ design.

When the last component of v_n is 1 and the last column of D_n is chosen as the branching column, following Equation (5.1), we can write D'_n as

$$D'_n = \begin{pmatrix} a & b & E & \mathbf{1} \\ -b & a & E & -\mathbf{1} \end{pmatrix}. \tag{5.5}$$

The following lemma expresses the J-characteristics of D'_n in terms of that of $D_{n-1} = (a, b, E)$.

Lemma 5. *Suppose that the last component of* v_n *is* 1 *and the last column of* D_n *is chosen as the branching column. For any subset* $e \subset \{3, 4, \ldots, 2n\}$,

(a) $j(\{1\} \cup e; D'_n) = -j(\{2, 2n+1\} \cup e; D'_n) = j(\{1\} \cup e; D_{n-1}) - j(\{2\} \cup e; D_{n-1})$;

(b) $j(\{2\} \cup e; D'_n) = j(\{1, 2n+1\} \cup e; D'_n) = j(\{1\} \cup e; D_{n-1}) + j(\{2\} \cup e; D_{n-1})$;

(c) $j(\{1, 2, 2n+1\} \cup e; D'_n) = 2j(\{1, 2\} \cup e; D_{n-1})$;

(d) $j(s \cup e; D'_n) = 0$ *for* $s = \{1, 2\}$, *or* $\{2n+1\}$.

It is easy to verify that choosing the second last column of D_n as the branching column yields a design that is equivalent to D'_n in Equation (5.5).

When the last component of v_n is 2 and the last (or second last) column of D_n is chosen as the branching column, following Equation (5.2), we can write D'_n as

$$D'_n = \begin{pmatrix} a & b & E & \mathbf{1} \\ -a & -b & E & -\mathbf{1} \end{pmatrix}. \tag{5.6}$$

We can also express the J-characteristics of D'_n in terms of that of D_{n-1}.

Lemma 6. *Suppose that the last component of v_n is 2 and the last column of D_n is chosen as the branching column. For any subset $e \subset \{3, 4, \ldots, 2n\}$,*

(a) $j(\{1, 2n+1\} \cup e; D'_n) = 2j(\{1\} \cup e; D_{n-1})$;

(b) $j(\{2, 2n+1\} \cup e; D'_n) = 2j(\{2\} \cup e; D_{n-1})$;

(c) $j(\{1, 2\} \cup e; D'_n) = 2j(\{1, 2\} \cup e; D_{n-1})$;

(d) $j(s \cup e; D'_n) = 0$ *for* $s = \{1\}, \{2\}, \{2n+1\}$, *or* $\{1, 2, 2n+1\}$.

5.4.2 Proofs of Theorems

Proof of Theorem 1. Without loss of generality, assume that v consists of p 1s followed by q 2s, where $p + q = n$. Lemma 4 suggests that all possible words are in forms of s_1, s_2 or s_0. If $p = 2t+1$, by Lemma 4(a), there are 2^p words of length $p + 2q + 1$ with aliasing index $\rho = 2^{-t}$. If $p = 2t$, by Lemma 4(b), there are 2^{p+1} words of length $p+2q+1$ with aliasing index $\rho = 2^{-t}$. By Lemma 4(c), there are 1 complete word of length $2p + 2$. This completes the proof. □

Proof of Theorem 2. Without loss of generality, assume that v consists of p 1s and q 2s, where $p + q = n$.

(a) We prove the result by induction on p. The result is trivial when $p = 0$. Assume that it is true for $p = k - 1$. Consider $p = k$. We can write $D_n = D_{k+q}$ as in Equation (5.1), where a and b are the balanced two-level columns and E

is a full factorial with $2k + 2q - 2$ columns. We need to show that D_{k+q} has projectivity $2k + 1$. Consider any subset s with $2k + 1$ columns of D_{k+q}. There are three possible cases.

(i) Both of the last two columns of D_{k+q} belong to s. Denote $E_1 = (a, b, E)$, $E_2 = (b, -a, E)$, $E_3 = (-a, -b, E)$ and $E_4 = (-b, a, E)$. Clearly the E_i's are isomorphic to each other. The assertion of $p = k - 1$ implies that each E_i has projectivity $2k - 1$. Then the projection onto s contains a full 2^{2k+1} factorial.

(ii) None of the last two columns of D_{k+q} belongs to s. Observe that E is a full factorial with $2k+2q-2 \geq 2k$ columns. It is easy to verify that the projection onto s contains a full 2^{2k+1} factorial whether s includes none, one or both of the first two columns.

(iii) One of the last two columns of D_{k+q} belongs to s and the other does not. Observe that the projection onto the subset consisting of the first two and the last two columns has resolution ≥ 4 and projectivity ≥ 3. Further observe that E is a full factorial. Then it is easy to verify that the projection onto s contains a full 2^{2k+1} factorial whether s includes none, one or both of the first two columns.

The three cases together suggest that D_{k+q} has projectivity $2k + 1$. By induction the proof is completed.

(b) The proof is similar to (a) and omitted. □

Proof of Theorem 3. Without loss of generality, assume that v consists of p 1s and q 2s, where $p + q = n$. Let v_{n-1} be the vector consisting of the first $n - 1$ components of v and let D_{n-1} be the binary design generated by $G_{n-1} = (v_{n-1}, I_{n-1})$.

(a) Without loss of generality, assume that the last component of v is 1 and that the last column of D is chosen as the branching column. If $p = 2t$, by

Lemma 4(a), D_{n-1} has 2^{p-1} words of length $p+2q$ with aliasing index $2^{-(t-1)}$. By Lemma 5(a) and 5(b), these 2^{p-1} words in D_{n-1} generate 2^p words of length $p+2q$ and 2^p words of length $p+2q+1$ with aliasing index $\rho = 2^{-t}$ in D'. If $p = 2t+1$, by Lemma 4(b), D_{n-1} has 2^p words of length $p+2q$ with aliasing index 2^{-t}. By Lemma 5(a) and 5(b), these 2^p words in D_{n-1} generate 2^{p-1} words of length $p+2q$ and 2^{p-1} words of length $p+2q+1$ with aliasing index $\rho = 2^{-t}$ in D'. So in both cases, D' has $1/\rho^2$ words of length $p+2q = k_1 - 1$ and $1/\rho^2$ words of length k_1 with aliasing index $\rho = 2^{-\lfloor p/2 \rfloor}$. By Lemma 4(c), D_{n-1} has 1 complete word of length $2p$, which generates a complete word of length $2p+1$ in D', by Lemma 5(c). This completes the proof.

(b) Without loss of generality, assume that the last component of v is 2 and that the last column of D is chosen as the branching column. By Theorem 1, D_{n-1} has 1 complete word of length $2p+2$ and $2/\rho^2$ words of length $p+2(q-1)+1$ with aliasing index $\rho = 2^{-\lfloor p/2 \rfloor}$. By Lemma 6(a) and 6(b), each partial word in D_{n-1} generates a partial word of length $p+2q = k_1 - 1$ in D' with aliasing index ρ. Lemma 6(c) implies that the complete word in D_{n-1} produces a complete word with the same length $2p+2 = k_2$ in D'. This completes the proof. □

Proof of Theorem 4. Without loss of generality, assume that v consists of p 1s and q 2s, where $p+q = n$.

(a) By Theorem 2(a), D has projectivity $2p+1$. It is obvious that any half-fraction of D has projectivity $\geq 2p$. By Theorem 3(a), D' has a complete word of length $2p+1$ so its projectivity is $2p$.

(b) As in (a), by Theorem 2(b), D has projectivity $2p-1$ so D' has projectivity $\geq 2p-2$.

(c) Without loss of generality, we write D' as Equation (5.6), where a and b

are balanced two-level columns and E is a full factorial with $2p+2q-2$ columns. By Theorem 2(a), $D_{n-1} = (a, b, E)$ has projectivity $2p+1$, so is $(-a, -b, E)$. Then it is clear that D' has projectivity $2p+1$.

(d) By Theorem 2, D has projectivity $2p+1$ so D' has projectivity $\geq 2p$. □

Proof of Theorem 5. Without loss of generality, we assume $f_0 = f_3 = 0$. Then $f_2 = n-f_1$, $k_1 = f_1+2f_2+1 = 2n-f_1+1$ and $k_2 = 2f_1+2$. According to Theorem 1 and Corollary 2 we need to consider whether the condition $k_1 \geq k_2$ holds. It is obvious that the condition $k_1 \geq k_2$ is equivalent to $f_1 \leq (2n-1)/3$. If $k_1 \geq k_2$, the resolution is $k_2 = 2f_1 + 2$ so we shall maximize k_2 and choose $f_1 = \lfloor (2n-1)/3 \rfloor$ since f_1 is an integer. If $k_1 < k_2$, the resolution is $k_1 + 1 - \rho = 2n - f_1 + 2 - \rho$ so we shall maximize k_1 and choose $f_1 = \lfloor (2n+1)/3 \rfloor$, the smallest integer that is greater than $(2n-1)/3$.

(a) When $n = 3k-1$, the first choice leads to $f_1 = 2k-1$, $f_2 = k$, $k_1 = k_2 = 4k$ and $R(D) = 4k$ while the second choice leads to $f_1 = 2k$, $f_2 = k-1$, $k_1 = 4k-1$, $k_2 = 4k+2$ and $R(D) = 4k-2^{-k}$. Therefore, the first choice leads to a maximum resolution design.

(b) When $n = 3k$, the first choice leads to $f_1 = 2k-1$, $f_2 = k+1$, $k_1 = 4k+2$, $k_2 = 4k$ and $R(D) = 4k$ while the second choice leads to $f_1 = 2k$, $f_2 = k$, $k_1 = 4k+1$, $k_2 = 4k+2$ and $R(D) = 4k+2-2^{-k}$. Therefore, the second choice leads to a maximum resolution design.

(c) When $n = 3k+1$, the first choice leads to $f_1 = 2k$, $f_2 = k+1$, $k_1 = 4k+3$, $k_2 = 4k+2$ and $R(D) = 4k+2$ while the second choice leads to $f_1 = 2k+1$, $f_2 = k$, $k_1 = 4k+2$, $k_2 = 4k+4$ and $R(D) = 4k+3-2^{-k}$. Therefore, the second choice leads to a maximum resolution design. □

Proof of Theorem 6. Note that the minimum aberration design must maximize

the integer part of the resolution. As explained in the proof of Theorem 5, we only need to consider two choices: $f_1 = \lfloor (2n-1)/3 \rfloor$ or $f_1 = \lfloor (2n+1)/3 \rfloor$.

(a) When $n = 3k-1$, the first choice leads to a minimum aberration design with $f_1 = 2k-1$, $f_2 = k$, $k_1 = k_2 = 4k$ and $A_{4k}(D) = 3$.

(b) When $n = 3k$, the second choice leads to a minimum aberration design with $f_1 = 2k$, $f_2 = k$, $k_1 = 4k+1$, $k_2 = 4k+2$, $A_{4k+1}(D) = 2$ and $A_{4k+2}(D) = 1$.

(c) When $n = 3k+1$, the first choice leads to $f_1 = 2k$, $f_2 = k+1$, $k_1 = 4k+3$, $k_2 = 4k+2$, $A_{4k+2}(D) = 1$ and $A_{4k+3}(D) = 2$ while the second choice leads to $f_1 = 2k+1$, $f_2 = k$, $k_1 = 4k+2$, $k_2 = 4k+4$, $A_{4k+2}(D) = 2$, and $A_{4k+4}(D) = 1$. Therefore, the first choice leads to a minimum aberration design. □

Proof of Theorem 7. It follows from Theorem 2. □

Proof of Theorem 8. Without loss of generality, we assume $f_0 = f_3 = 0$ so that $f_1 + f_2 = n$. According to Theorem 3, we need to consider four cases: (i) the branching column is associated with number 1 and $k_1 \geq k_2$, (ii) the branching column is associated with number 1 and $k_1 < k_2$, (iii) the branching column is associated with number 2 and $k_1 - 1 \geq k_2$, (iv) the branching column is associated with number 2 and $k_1 - 1 < k_2$. For each case, we choose f_1 and f_2 to maximize the shortest wordlength and resolution. The resolutions and wordlength patterns of the resulting designs can be calculated by Corollaries 4, 5, 6 and 7.

(a) When $n = 3k-1$, the condition $k_1 \geq k_2$ is equivalent to $f_1 \leq 2k-1$; the condition $k_1 - 1 \geq k_2$ is equivalent to $f_1 \leq 2k - 4/3$. For case (i) we want to maximize k_2 so choose $f_1 = 2k-1$ and $f_2 = k$, which yields $k_1 = 4k$, $k_2 = 4k$, $R(D') = 4k-1$, $A_{4k-1}(D') = 2$ and $A_{4k}(D') = 1$. For case (ii) we want to maximize k_1 so choose $f_1 = 2k$ and $f_2 = k-1$, which yields $k_1 = 4k-1$, $k_2 = 4k+2$, $R(D') = 4k-1-2^{-k}$, and $A_{4k-2}(D') = A_{4k-1}(D') = A_{4k+1}(D') = 1$.

For case (iii) we want to maximize k_2 so choose $f_1 = 2k-2$ and $f_2 = k+1$, which yields $k_1 = 4k+1$, $k_2 = 4k-2$, $R(D') = 4k-2$, $A_{4k-2}(D') = 1$ and $A_{4k}(D') = 2$. For case (iv) we want to maximize k_1 so choose $f_1 = 2k-1$ and $f_2 = k$, which yields $k_1 = 4k$, $k_2 = 4k$, $R(D') = 4k - 2^{-(k-1)}$, $A_{4k-1}(D') = 2$ and $A_{4k}(D') = 1$. Therefore, the design in case (iv) has both maximum resolution and minimum aberration.

(b) When $n = 3k$, the condition $k_1 \geq k_2$ is equivalent to $f_1 \leq 2k - 1/3$; the condition $k_1 - 1 \geq k_2$ is equivalent to $f_1 \leq 2k - 2/3$. For case (i) we shall choose $f_1 = 2k-1$ and $f_2 = k+1$, which yields $k_1 = 4k+2$, $k_2 = 4k$, $R(D') = 4k-1$, and $A_{4k-1}(D') = A_{4k+1}(D') = A_{4k+2}(D') = 1$. For case (ii) we shall choose $f_1 = 2k$ and $f_2 = k$, which yields $k_1 = 4k+1$, $k_2 = 4k+2$, $R(D') = 4k+1-2^{-k}$, $A_{4k}(D') = 1$ and $A_{4k+1}(D') = 2$. For case (iii) we shall choose $f_1 = 2k-1$ and $f_2 = k+1$, which yields $k_1 = 4k+2$, $k_2 = 4k$, $R(D') = 4k$, $A_{4k}(D') = 1$ and $A_{4k+1}(D') = 2$. For case (iv), we shall choose $f_1 = 2k$ and $f_2 = k$, which yields $k_1 = 4k+1$, $k_2 = 4k+2$, $R(D') = 4k+1-2^{-k}$, $A_{4k}(D') = 2$ and $A_{4k+2}(D') = 1$. Therefore, the design in case (ii) has both maximum resolution and minimum aberration.

(c) When $n = 3k+1$, the condition $k_1 \geq k_2$ is equivalent to $f_1 \leq 2k + 1/3$; the condition $k_1 - 1 \geq k_2$ is equivalent to $f_1 \leq 2k$. For case (i) we shall choose $f_1 = 2k$ and $f_2 = k+1$, which yields $k_1 = 4k+3$, $k_2 = 4k+2$, $R(D') = 4k+1$, $A_{4k+1}(D') = A_{4k+2}(D') = A_{4k+3}(D') = 1$. For case (ii) we shall choose $f_1 = 2k+1$ and $f_2 = k$, which yields $k_1 = 4k+2$, $k_2 = 4k+4$, $R(D') = 4k+2-2^{-k}$, $A_{4k+1}(D') = A_{4k+2}(D') = A_{4k+3}(D') = 1$. For case (iii) we shall choose $f_1 = 2k$ and $f_2 = k+1$, which yields $k_1 = 4k+3$, $k_2 = 4k+2$, $R(D') = 4k+2$, and $A_{4k+2}(D') = 3$. For case (iv) we shall choose $f_1 = 2k+1$ and $f_2 = k$, which yields $k_1 = 4k+2$, $k_2 = 4k+4$, $R(D') = 4k+2-2^{-k}$, $A_{4k+1}(D') = 2$ and $A_{4k+4}(D') = 1$.

Therefore, the design in case (iii) has both maximum resolution and minimum aberration. □

Proof of Theorem 9. It follows from Theorem 4. □

Proof of Theorem 10. (a) The quarter-fraction designs given in Theorems 6 and 8 have the same wordlength patterns as the regular minimum aberration designs. Then the result follows from Theorem 2 of Xu (2005), which states that the regular minimum aberration 2^{m-2} design has minimum aberration among all possible designs.

(b) The quarter-fraction designs given in Theorems 7 and 9 have 2^{m-2} runs and projectivity $m-3$. It is sufficient to prove that the projectivity of any $2^k \times m$ two-level design D is at most $k-1$ for $m \geq k+2$. Assume that D has projectivity k. Then the projection onto any k factors is an unreplicated 2^k full factorial because D has exactly 2^k runs. Therefore, D is an orthogonal array of strength k. Theorem 2.19 of Hedayat et. al. (1999, p. 24) implies that $m < k+1$. This contradicts the condition $m \geq k+2$. □

Table 5.2: Best Quarter-Fraction Designs

Design	Quaternary-code Designs					Regular	
	Criterion	v^T	WLP	R	pr	R	pr
2^{6-2}	r, a, p	[12]	$A_4 = 3$	4.0	3	4	3
2^{7-2}	r, a, p	$[112]f$	$A_4 = 1,\ A_5 = 2$	4.5	4	4	3
2^{8-2}	r, a, p	[112]	$A_5 = 2,\ A_6 = 1$	5.5	5	5	4
2^{9-2}	r, a	$[1122]l$	$A_6 = 3$	6.0	5	6	5
	p	$[1112]f$	$A_5 = 1,\ A_6 = 2$	5.5	6		
2^{10-2}	r, p	[1112]	$A_6 = 2,\ A_8 = 1$	6.5	7	6	5
	a	[1122]	$A_6 = 1,\ A_7 = 2$	6.0	5		
2^{11-2}	r, a	$[11122]l$	$A_7 = 2,\ A_8 = 1$	7.5	7	7	6
	p	$[11112]f$	$A_6 = A_7 = A_9 = 1$	6.75	8		
2^{12-2}	r, a	[11122]	$A_8 = 3$	8.0	7	8	7
	p	[11112]	$A_7 = 2,\ A_{10} = 1$	7.75	9		
2^{13-2}	r, a	$[111122]f$	$A_8 = 1,\ A_9 = 2$	8.75	8	8	7
	p	$[111112]f$	$A_7 = A_8 = A_{11} = 1$	7.75	10		
2^{14-2}	r, a	[111122]	$A_9 = 2,\ A_{10} = 1$	9.75	9	9	8
	p	[111112]	$A_8 = 2,\ A_{12} = 1$	8.75	11		
2^{15-2}	r, a	$[1111222]l$	$A_{10} = 3$	10.0	9	10	9
	p	$[1111112]f$	$A_8 = A_9 = A_{13} = 1$	8.875	12		
2^{16-2}	r	[1111122]	$A_{10} = 2,\ A_{12} = 1$	10.75	11	10	9
	a	[1111222]	$A_{10} = 1,\ A_{11} = 2$	10.0	9		
	p	[1111112]	$A_9 = 2,\ A_{14} = 1$	9.875	13		

CHAPTER 6

Construction of Two-level Nonregular $(1/16)^{th}$-FFDs

Inspired by those good properties of designs in the quarter-fraction case, it is interesting to extend the application of quaternary code construction method to other classes of designs. In this chapter, we extend our results to $(1/16)^{th}$-fraction designs, which are the closest and simplest extensions from the quarter-fraction designs.

6.1 Properties of $(1/16)^{th}$-FFDs via Quaternary Codes

6.1.1 $(1/16)^{th}$-FFDs with Even Number of Factors

To construct $1/16^{th}$-fraction designs, consider an $n \times (n+2)$ generator matrix $G = (v_1, v_2, I_n)$, where v_i is an $n \times 1$ column vector over Z_4 for $i = 1, 2$ and I_n is an $n \times n$ identity matrix. Let D be the $2^{2n} \times (2n+4)$ two-level design generated by G. It is easy to verify that the identity matrix I_n generates a full $2^{2n} \times 2n$ design; therefore, the property of D depends on the columns v_1 and v_2 only. Throughout this chapter, for $i, j = \{0, 1, 2, 3\}$, let f_{ij} be the number of times that numbers i and j appear together in columns v_1 and v_2 respectively. Theorem 11 characterizes the number of words of D, their lengths and aliasing indexes in terms of the frequency f_{ij}.

Theorem 11. *Suppose that D is the two-level $2^{2n} \times (2n+4)$ design generated by $G = (v_1, v_2, I_n)$. Further suppose $f_{11} + f_{13} + f_{31} + f_{33} > 0$. Table 6.1 provides the definitions of wordlengths $k_1, \ldots, k_9$ and the aliasing index $\rho_I, \rho_{II}, \rho_{III}$, and summarizes the numbers, wordlengths and aliasing indexes of words.*

Example 9. Consider a generator matrix

$$G = \begin{pmatrix} v_1 & v_2 & I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

All linear combinations of the three rows of G form a 64×5 linear code C over Z_4. Applying the Gray map, a 64×10 binary image $D = \phi(C)$ is obtained. According to Theorem 1, design D has 1 complete word of length $k_1 = 8$, 1 complete word of length $k_2 = 6$, 1 complete word of length $k_3 = 6$, 8 partial words of length $k_4 = 5$ with aliasing index $\rho_I = 0.5$, 8 partial words of length $k_5 = 5$ with aliasing index $\rho_{II} = 0.5$, 8 partial words of length $k_6 = 5$ with aliasing index $\rho_I = 0.5$, 8 partial words of length $k_7 = 5$ with aliasing index $\rho_{II} = 0.5$, 8 partial words of length $k_8 = 6$ with aliasing index $\rho_{III} = 0.5$ and 8 partial words of length $k_9 = 4$ with aliasing index $\rho_{III} = 0.5$. Therefore, by definitions (2.2) and (2.3), the resolution of D is 4.5, and the wordlength pattern of D is $A_4(D) = 2$, $A_5(D) = 8$, $A_6(D) = 4$, $A_8(D) = 1$, and $A_i(D) = 0$ for $i \neq 4, 5, 6, 8$.

For ease of presentation, we say that the ith identity column of I_n in $G = (v_1, v_2, I_n)$ is "associated with" number $\{z_1, z_2\}$ if the ith elements of v_1 and v_2 are z_1 and z_2 respectively, where $z = 0, 1, 2$, or 3.

Recall that a regular design has only complete words. Corollary 8 provides a sufficient and necessary condition for D to be a regular design.

Corollary 8. *Design D is regular if and only if* $\sum_{j=0}^{3}(f_{1j}+f_{3j})+\sum_{i=0}^{3}(f_{i1}+f_{i3}) = 0$.

It is obvious that under this condition, ρ_I and ρ_{II} can only be 1. For ρ_{III}, when both $\sum_{j=0}^{3}(f_{1j}+f_{3j}) = 0$ and $\sum_{i=0}^{3}(f_{i1}+f_{i3}) = 0$, $\sum_{i+j=odd} f_{ij}$ can only be 0.

It is straightforward to complete the resolution of D according to the definition (2.2) and Theorem 11.

Corollary 9. *The resolution of D is $r+1-\rho$ where r is the minimum among $k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8, k_9$ and*

(a) $\rho = 1$ *if* $r = min(k_1, k_2, k_3) < min(k_4, k_5, k_6, k_7, k_8, k_9)$,

(b) $\rho = \rho_I$ *if* $r = min(k_4, k_6) \leq min(k_1, k_2, k_3, k_5, k_7, k_8, k_9)$ *and* $\rho_I \geq max(\rho_{II}, \rho_{III})$,

(c) $\rho = \rho_{II}$ *if* $r = min(k_5, k_7) \leq min(k_1, k_2, k_3, k_4, k_6, k_8, k_9)$ *and* $\rho_{II} \geq max(\rho_I, \rho_{III})$,

(d) $\rho = \rho_{III}$ *if* $r = min(k_8, k_9) \leq min(k_1, k_2, k_3, k_4, k_5, k_6, k_7)$ *and* $\rho_{III} \geq max(\rho_I, \rho_{II})$,

According to the definition (2.3), when there are n words of length k with aliasing index ρ, then A_k increases by $n\rho^2$. Therefore, to obtain the wordlength pattern, we define A'_{k_i} as the addition to the element of wordlength pattern from the words of length k_i. For example, if there are $2/\rho_I^2$ partials words of length k_4 with aliasing index ρ_I, then $A'_{k_4} = 2$. Once all nine A'_{k_i} are obtained, then the elements of wordlength pattern are $A_k = \sum_{i=1}^{9} A'_{k_i} \cdot 1_{\{k=k_i\}}$ where $1_{\{A\}}$ is 1 if A is true, or 0 otherwise. The number of words, lengths of words and the corresponding aliasing indexes can be found in Theorem 11.

6.1.2 $(1/16)^{th}$-FFDs with Odd Number of Factors

Suppose a binary design D generated by $G = (v_1, v_2, I_n)$ has 2^{2n} runs and $2n+4$ factors. To construct $1/16^{th}$-fraction designs with 2^{2n-1} runs, we use the half fraction method, which works as follows. Choose any column of D as a branching column, which divides D into two half-fractions according to the symbols of the branching column. Deleting the branching column yields two $2^{2n-1} \times (2n+3)$ designs. It is easy to verify that the two half-fractions of D are equivalent. However, the properties of the half-fraction design depend on the branching column, which are characterized in Theorem 12.

Theorem 12. *Suppose that D is the two-level $2^{2n} \times (2n+4)$ design generated by $G = (v_1, v_2, I_n)$ and that D' is a half-fraction design of D from branching a column. Define $k_1, \ldots, k_9$ and $\rho_I, \rho_{II}, \rho_{III}$ as in Theorem 11. Table 6.2 summarizes the numbers, wordlengths and aliasing indexes of words when we choose certain branching column.*

The result of choosing the branching column associated with either vector v_1 or vector v_2 is too complicated and it is still under investigation. If the branching column is associated with $\{0, 0\}$, D' and D share the same words because the branching column does not appear in any word of D.

According to the definition, the resolution of D' is in the form of $r+1-\rho$. r is the minimum values among a subset of all k_i and $k_i - 1$ such that the number of words of length either k_i or $k_i - 1$ is nonzero. Then ρ can be found in the similar way as in Corollary 9. The wordlength pattern of D' can be found in the similar way as in D. The only difference is that there are nine additional A'_{k_i-1} in D' but not D. Once all nine A'_{k_i} and nine A'_{k_i-1} are obtained, then the elements of wordlength pattern are $A_k = \sum_{i=1}^{9}(A'_{k_i} \cdot 1_{k=k_i} + A'_{k_i-1} \cdot 1_{k=k_i-1})$. The number of

words, lengths of words and the corresponding aliasing indexes can be found in Theorem 12.

Example 10. Consider two choices of half-fractions of D of Example 9.

Choice 1: Branching the fifth or sixth columns of D. We obtain a 32×9 design D' with resolution 4.5 and wordlength pattern $A_4(D') = 6$, $A_5(D') = 8$, $A_8(D') = 1$ and $A_i(D') = 0$ for $i \neq 4, 5, 8$. Design D' has 2 complete words of length 5, 1 complete word of length 8, 24 partial words of length 4 with aliasing index 0.5 and 24 partial words of length 5 with aliasing index 0.5.

Choice 2: Branching any one of the last four columns of D. We obtain a 32×9 design D' with resolution 3.5 and wordlength pattern $A_3(D') = 1$, $A_4(D') = 5$, $A_5(D') = 6$, $A_6(D') = 2$, $A_7(D') = 1$ and $A_i(D') = 0$ for $i \neq 3, 4, 5, 6, 7$. Design D' has 1 complete word of length 5, 1 complete word of length 6, 1 complete word of length 7, 4 partial words of length 3, 20 partial words of length 4 with aliasing index 0.5, 20 partial words of length 5 with aliasing index 0.5 and 4 partial words of length 6 with aliasing index 0.5.

6.2 Structure of Some Good $(1/16)^{th}$-FFDs

In this subsection we apply Theorem 11 developed to construct good designs in terms of resolutions.

Theorem 13. *Consider all $2^{2n} \times (2n + 4)$ designs generated by $G = (v_1, v_2, I_n)$., where $n = 15t + b$ for any $t \geq 0$ and $b = 1, \ldots, 15$,*

(a) *If $n = 15t + 6s - 2$ for $t \geq 0$ and $s = \{1, 2, 3, 4\}$, then a two-level design*

D generated by G with $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t+s-1 & t & 0 \\ 2t+s-1 & 2t+s & 2t+s & 2t+s \\ t & 2t+s & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t + 6s + 1 - 2^{-\rho}$*, where* $\rho = 4t + 2s - 1$*.*

(b) *If* $n = 15t + 6s$ *for* $t \geq 0$ *and* $s = \{1, 2, 3, 4\}$*, then a two-level design D generated by G with* $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t+s-1 & t+1 & 0 \\ 2t+s-1 & 2t+s & 2t+s & 2t+s \\ t+1 & 2t+s & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t + 6s + 3 - 2^{-\rho}$*, where* $\rho = 0$ *for* $s = 1$ *or* $\rho = 4t + 2s - 1$ *otherwise.*

(c) *If* $n = 15t + 6s + 2$ *for* $t \geq 0$ *and* $s = \{1, 2, 3\}$*, then a two-level design D generated by G with* $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t+s & t & 0 \\ 2t+s & 2t+s & 2t+s+1 & 2t+s \\ t & 2t+s+1 & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t + 6s + 5 - 2^{-\rho}$*, where* $\rho = 4t + 2s$*.*

(d) *If* $n = 15t - 2$ *for* $t > 0$*, then a two-level design D generated by G with*

$\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t-1 & t & 0 \\ 2t-1 & 2t & 2t & 2t \\ t & 2t & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t$.

(e) *If* $n = 15t$ *for* $t > 0$, *then a two-level design* D *generated by* G *with* $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t & t & 0 \\ 2t & 2t & 2t & 2t \\ t & 2t+s & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t + 2 - 2^{-4t}$.

(f) *If* $n = 15t + 2$ *for* $t > 0$, *then a two-level design* D *generated by* G *with* $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t & t & 0 \\ 2t & 2t & 2t+1 & 2t \\ t & 2t+1 & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t + 4$.

(g) *If* $n = 15t + 11$ *for* $t > 0$, *then a two-level design* D *generated by* G *with* $\{v_1, v_2\}$ *characterized by the frequency matrix*

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t+1 & t+1 & 0 \\ 2t+1 & 2t+2 & 2t+2 & 2t+1 \\ t+1 & 2t+2 & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

has resolution $16t - 2 - 2^{-(4t-1)}$.

(h) *If* $n = 3, 5, 7, 9$*, then the frequency matrix for characterization of* $\{v_1, v_2\}$ *in* G *and the resolution of* D *generated by* G *is given in Table 6.3.*

Table 6.4 provides a comparison, in terms of resolution, between our good quaternary-code designs from the previous theorem and the regular minimum aberration 2^{n-4} FFDs given by Chen and Wu (1991). This table only shows part of the results for $n = \{3, \ldots, 12\}$, and we have verified cases up to $n = 45$ via computer. The comparison shows that our good quaternary-code designs have the same or larger resolution than regular minimum aberration designs, except only two cases when $n = 5$ and $n = 9$. In addition, all of these good quaternary-code designs are nonregular.

6.3 Proofs

In this section, we are going to prove some theorems related to $(1/16)^{th}$-FFDs constructed via quaternary codes.

Proof of Theorem 11. We classify all the words into three types.

Type 1: words involve columns associated with v_1 not columns associated with v_2;

Type 2: words involve columns associated with v_2 not columns associated with v_1;

Type 3: words involve columns associated with both v_1 and v_2.

It is clear that the first two type words can be obtained from Theorem 1 directly. Specially, for type 1, there are one complete word of length k_2 and $2/\rho_I^2$ partial

words of length k_4 with aliasing index ρ_I; for type 2, there are one complete word of length k_3 and $2/\rho_{II}^2$ partial words of length k_5 with aliasing index ρ_{II}. Observe that type 3 words are products of type 1 and type 2 words. The product of two complete words is a complete word and the product of a complete word and a partial word forms a partial word. Therefore, for type 3, there are one complete word of length k_1, $2/\rho_I^2$ partial words of length k_6 with aliasing index ρ_I and $2/\rho_{II}^2$ partial words of length k_7 with aliasing index ρ_{II}. These type 3 partial words consist of three out of the four columns associated with either v_1 or v_2. The situation of the product of two partial words is complicated. To figure out this, without loss of generality, assume the first component of both v_1 and v_2 are 1 and let

$$G = \begin{pmatrix} 1 & 1 & 1 & \mathbf{0}^T \\ a & b & \mathbf{0} & I_{n-1} \end{pmatrix}$$

where a, b and $\mathbf{0}$ are column vectors. Consider another matrix

$$G' = \begin{pmatrix} 1 & 1 & 1 & \mathbf{0}^T \\ \mathbf{0} & b-a & -a & I_{n-1}. \end{pmatrix} \pmod 4$$

Clearly G and G' generate the same quaternary code. Let f_i be the frequency that i appears in the second column of G'. It is easy to see that

$$f_1 = 1 + f_{01} + f_{12} + f_{23} + f_{30},$$

$$f_2 = f_{02} + f_{13} + f_{20} + f_{31},$$

$$f_3 = f_{10} + f_{21} + f_{32} + f_{03}.$$

Now applying Theorem 1 to the generator matrix G', we know that there are $2/\rho^2$ partial words of length $k_9 = 2f_1 + 2f_3 + 2$ with aliasing index $\rho = 2^{-\lfloor (f_1+f_3)/2 \rfloor} = \rho_{III}$. These are type 3 partial words because they involve one column associated with v_1 and one column associated with v_2. Further they do not involve any

column associated with the third column of G. There are other type 3 partial words that involve both columns associated with the third column of G. Note that the type 1 complete word involves both columns associated with v_1 and both columns associated with the third column of G. The product of a type 3 partial word and the type 1 complete word is also a type 3 partial word. It is easy to verify that the resulting word has length k_8. Therefore, we have another $2/\rho_{III}^2$ partial words of length k_8 with aliasing index ρ_{III} which involve both columns associated with the third column of G. It can be verified that there are no other words. This completes the proof. □

Proof of Theorem 12. The proof is similar to the proof of Theorem 11 and omitted. □

Proof of Theorem 13. We prove the generalized resolution separately in eight cases.

Case (a): $n = 15t + 6s - 2$

Given

$$f = [f_{ij}] = \begin{pmatrix} 0 & 2t+s-1 & t & 0 \\ 2t+s-1 & 2t+s & 2t+s & 2t+s \\ t & 2t+s & t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

then

$$k = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \\ k_8 \\ k_9 \end{pmatrix} = \begin{pmatrix} 16t + 8s \\ 16t + 8s \\ 16t + 8s \\ 16t + 6s \\ 16t + 6s \\ 16t + 6s \\ 16t + 6s \\ 16t + 6s + 1 \\ 16t + 6s + 1 \end{pmatrix}$$

and

$$\rho = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} = \begin{pmatrix} 2^{-(4t+2s-1)} \\ 2^{-(4t+2s-1)} \\ 2^{-(4t+2s-1)} \end{pmatrix}$$

Therefore, when $s = 0$, the generalized resolution is $16t$. When $s = \{1, 2, 3, 4\}$, the generalized resolution is $16t + 6s + 1 - 2^{-(4t+2s-1)}$.

The results of Cases (b)–(h) can be verified using the same idea as above. $\square$

Table 6.1: Number of Words of Different Wordlengths and Aliasing Indexes with their Definitions in Theorem 11

k_i	ρ	No. Words	Definition of k_i
k_1	1	1	$k_1 = 2\sum_{i+j=odd} f_{ij} + 4$
k_2	1	1	$k_2 = 2\sum_{j=0}^{3}(f_{1j} + f_{3j}) + 2$
k_3	1	1	$k_3 = 2\sum_{i=0}^{3}(f_{i1} + f_{i3}) + 2$
k_4	ρ_I	$2/\rho_I^2$	$k_4 = \sum_{j=0}^{3}(f_{1j} + 2f_{2j} + f_{3j}) + 1$
k_5	ρ_{II}	$2/\rho_{II}^2$	$k_5 = \sum_{i=0}^{3}(f_{i1} + 2f_{i2} + f_{i3}) + 1$
k_6	ρ_I	$2/\rho_I^2$	$k_6 = \sum_{j=0}^{3}(f_{1j} + f_{3j}) + 2(f_{20} + f_{01} + f_{22} + f_{03}) + 3$
k_7	ρ_{II}	$2/\rho_{II}^2$	$k_7 = \sum_{i=0}^{3}(f_{i1} + f_{i3}) + 2(f_{02} + f_{10} + f_{22} + f_{30}) + 3$
k_8	ρ_{III}	$2/\rho_{III}^2$	$k_8 = \sum_{i+j=odd} f_{ij} + 2(f_{11} + f_{33} + f_{20} + f_{02}) + 2$
k_9	ρ_{III}	$2/\rho_{III}^2$	$k_9 = \sum_{i+j=odd} f_{ij} + 2(f_{13} + f_{31} + f_{20} + f_{02}) + 2$
Definitions of ρ: $\rho_I = 2^{-\lfloor \frac{1}{2}\sum_{j=0}^{3}(f_{1j}+f_{3j})\rfloor}$ $\rho_{II} = 2^{-\lfloor \frac{1}{2}\sum_{i=0}^{3}(f_{i1}+f_{i3})\rfloor}$ $\rho_{III} = 2^{-\lfloor \frac{1}{2}+\frac{1}{2}\sum_{i+j=odd} f_{ij}\rfloor}$			

Table 6.2: Number of Words of Different Wordlengths and Branching Column in Theorem 12

Word-Length	Branching Column								
	{0,1}	{0,2}	{1,0}	{1,1}	{1,2}	{1,3}	{2,0}	{2,1}	{2,2}
k_1	0	1	0	1	0	1	1	0	1
k_1-1	1	0	1	0	1	0	0	1	0
k_2	1	1	0	0	0	0	1	1	1
k_2-1	0	0	1	1	1	1	0	0	0
k_3	0	1	1	0	1	0	1	0	1
k_3-1	1	0	0	1	0	1	0	1	0
k_4	$2/\rho_I^2$	$2/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	0	0	0
k_4-1	0	0	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$2/\rho_I^2$	$2/\rho_I^2$	$2/\rho_I^2$
k_5	$1/\rho_{II}^2$	0	$2/\rho_{II}^2$	$1/\rho_{II}^2$	0	$1/\rho_{II}^2$	$2/\rho_{II}^2$	$1/\rho_{II}^2$	0
k_5-1	$1/\rho_{II}^2$	$2/\rho_{II}^2$	0	$1/\rho_{II}^2$	$2/\rho_{II}^2$	$1/\rho_{II}^2$	0	$1/\rho_{II}^2$	$2/\rho_{II}^2$
k_6	0	$2/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	0	$2/\rho_I^2$	0
k_6-1	$2/\rho_I^2$	0	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$1/\rho_I^2$	$2/\rho_I^2$	0	$2/\rho_I^2$
k_7	$1/\rho_{II}^2$	0	$2/\rho_{II}^2$	$1/\rho_{II}^2$	$2/\rho_{II}^2$	$1/\rho_{II}^2$	$2/\rho_{II}^2$	$1/\rho_{II}^2$	0
k_7-1	$1/\rho_{II}^2$	$2/\rho_{II}^2$	0	$1/\rho_{II}^2$	0	$1/\rho_{II}^2$	0	$1/\rho_{II}^2$	$2/\rho_{II}^2$
k_8	$1/\rho_{III}^2$	0	$1/\rho_{III}^2$	0	$1/\rho_{III}^2$	$2/\rho_{III}^2$	0	$1/\rho_{III}^2$	$2/\rho_{III}^2$
k_8-1	$1/\rho_{III}^2$	$2/\rho_{III}^2$	$1/\rho_{III}^2$	$2/\rho_{III}^2$	$1/\rho_{III}^2$	0	$2/\rho_{III}^2$	$1/\rho_{III}^2$	0
k_9	$1/\rho_{III}^2$	0	$1/\rho_{III}^2$	$2/\rho_{III}^2$	$1/\rho_{III}^2$	0	0	$1/\rho_{III}^2$	$2/\rho_{III}^2$
k_9-1	$1/\rho_{III}^2$	$2/\rho_{III}^2$	$1/\rho_{III}^2$	0	$1/\rho_{III}^2$	$2/\rho_{III}^2$	$2/\rho_{III}^2$	$1/\rho_{III}^2$	0

Table 6.3: Theorem 13: Frequency Matrix and Generalized Resolution for case (f)

n	Frequency Matrix f	Resolution
3	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	4.5
5	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	6.5
7	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	8.75
9	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	10.875

Table 6.4: Resolution Comparison on $(1/16)^{th}$-Fraction Designs

n	Design	v_1 and v_2	Resolution	
			Z_4 FFDs	Max R Regular FFDs
3	2^{10-4} design	(112) (121)	4.5	4.0
4	2^{12-4} design	(1112) (1231)	6.5	6.0
5	2^{14-4} design	(11112) (11231)	6.5	7.0
6	2^{16-4} design	(011122) (212301)	8.0	8.0
7	2^{18-4} design	(1111112) (0112231)	8.75	8.0
8	2^{20-4} design	(01111122) (10122311)	10.75	10.0
9	2^{22-4} design	(011111122) (101122311)	10.875	11.0
10	2^{24-4} design	(0111111122) (1011223311)	12.875	12.0
11	2^{26-4} design	(00111111222) (12011223011)	13.875	13.0
12	2^{28-4} design	(001111111222) (120112233011)	14.875	14.0

CHAPTER 7

Summary and Conclusions

It has been almost 20 years since Hamada and Wu suggested the use of nonregular FFDs in their 1992 paper. This thesis reviews the development on nonregular FFDs, including the projective properties, the generalized resolution and aberration. However, due to the lack of structure, nonregular FFDs are practically used only in screening experiments nowadays. It is a significant contrast when regular FFDs become the tradition and primary choice for users, both experimenters and industrialists.

It is not trivial to analyze an experiment using nonregular FFDs. This thesis tries to re-illustrate the correct ways to perform this complicated analysis. In particular, disentanglement is an important procedure when main effects are of primary interests in an experiment, but there exist some non-negligible interactions aliased with those main effects. This statement is true no matter one uses regular and nonregular FFDs. Three real-life examples from chemometrics are given as demonstrations, showing the potential pitfalls without the disentanglement of significant interactions. The first example shows that some truly significant main effects may be missed in the model selection because these significant main effects are canceled with the significant interactions through their aliased structure. The second example shows that the factor level assignments may be messed up because of the existence of significant interactions. The third example shows that some insignificant main effects may be falsely treated as

significant because they are aliased with some significant interactions. .

There exists numerous model selection methods for identifying active factors among all factors in an experiment. This thesis introduces the Dantzig selector for selecting active effects in supersaturated designs. We propose a graphical procedure called "profile plot" and an automatic variable selection method to accompany with the Dantzig selector. Simulation shows that the Dantzig selector method performs well compared to existing methods in the literature and is more efficient at estimating the model size. The advantages of the Dantzig selector includes (1) it is a theory-based method; (2) the program is relatively fast, easy and simple to use; and (3) the Dantzig selector is able to handle a large number of factors in two-level, multi-level and mixed-level experiments.

The major challenge in the research of nonregular FFDs does not focus on the analysis procedures, but the construction of a class of nonregular FFDs with good properties. This thesis aims at constructing a class of nonregular FFDs via quaternary codes. It takes on values $\{0, 1, 2, 3\}$ mod 4. The construction procedure is analogue to the regular FFDs, except an additional step of transformation using the Gray map is needed. The properties of quarter-fraction designs constructed via quaternary codes, which include the resolution, the wordlength pattern and the projectivity, are fully explored. Theorems 1 to 4 and their corollaries characterize the properties of a quarter-fraction design D constructed from $G = (v, I_n)$ by using a frequency vector f, which are the frequencies of $\{0, 1, 2, 3\}$ of v. Theorems 5 to 9 suggest the structure of the best quarter-fraction designs constructed via quaternary codes. The exploration of the design properties and structure are extended to $(1/16)^{th}$-fraction designs. In particular, Theorems 11 and 12 characterize the properties of a $(1/16)^{th}$-fraction design D, while Theorem 13 shows the structure of some good $(1/16)^{th}$-fraction designs with even number

of factors. The structure of some good $(1/16)^{th}$-fraction designs with odd number of factors is still under investigation.

The development of the fast and easy-to-use Dantzig selector method for active factor identification and the simple quaternary code method for nonregular FFD construction will lead to a much wider use of nonregular FFDs in both scientific and industrial fields, which will eventually reduce costs and resources for scientific researches and increase the performance of existing processes for industrial applications.

Bibliography

[1] Abraham, B., Chipman, H., Vijayan, K. (1999). Some risks in the construction and analysis of supersaturated designs. *Technometrics*, **41**, 135–141.

[2] Ai, M.Y., Li, P.F., Zhang, R.C. (2005). Optimal criteria and equivalence for nonregular fractional factorial designs. *Metrika*, **62**, 73–83.

[3] Beattie, S.D., Fong, D.K.H., Lin, D.K.J. (2002). A two-stage Bayesian model selection strategy for supersaturated designs. *Technometrics*, **44**, 55–63.

[4] Box, G.E.P., Hunter, J.S. (1961). The 2^{k-p} fractional factorial designs. *Technometrics*, **3**, 311–351, 449–458.

[5] Box G.E.P., Hunter W.G., Hunter J.S. (2005). Statistics for Experimenters: Design, Innovation, and Discovery (2nd edition). Wiley: New York.

[6] Box, G.E.P., Meyer, R.D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

[7] Box G.E.P. and Meyer R.D. (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology*, **25**, 94–105.

[8] Box, G.E.P., Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika*, **83**, 950–955.

[9] Bulutoglu, D.A., Cheng, C.S. (2003). Hidden projection properties of some nonregular fractional factorial designs and their applications. *Annals of Statistics*, **31**, 1012–1026.

[10] Bulutoglu, D.A., Margot, F. (2008). Classification of orthogonal arrays by integer programming. *Journal of Statistical Planning and Inference*, **138**, 654–666.

[11] Cai, T.T., Lv, J. (2007). Discussion: The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, **35**, 2365–2369.

[12] Candes, E.J., Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, **35**, 2313–2351.

[13] Chen, J., Wu, C.F.J. (1991). Some results on s^{n-k} fractional factorial designs with minimum aberration or optimal moments. *Annals of Statistics*, **19**, 1028–1041.

[14] Cheng, C.S. (1980). Orthogonal arrays with variable numbers of symbols. *Annals of Statistics*, **8**, 447–453.

[15] Cheng, C.S. (1995). Some projection properties of orthogonal arrays. *Annals of Statistics*, **23**, 1223–1233.

[16] Cheng, C.S. (1998). Some hidden projection properties of orthogonal arrays with strength three. *Biometrika*, **85**, 491–495.

[17] Cheng, C.S. (2006). Projection properties of factorial designs for factor screening. In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, Edited by Dean, A. and Lewis, S., 156–168. New York: Springer.

[18] Cheng, C.S., Deng, L.Y., Tang, B. (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Statistica Sinica*, **12**, 991–1000.

[19] Cheng, C.S., Steinberg, D.M., Sun, D.X. (1999). Minimum aberration and model robustness for two-level fractional factorial designs. *Journal of the Royal Statistical Society: Series B*, **61**, 85–93.

[20] Cheng, S.W., Li, W., Ye, K.Q. (2004). Blocked nonregular two-level factorial designs. *Technometrics*, **46**, 269–279.

[21] Cheng, S.W., Wu, C.F.J. (2001). Factor screening and response surface exploration (with discussion). *Statistica Sinica*, **11**, 553–604.

[22] Cheng, S.W., Ye, K.Q. (2004). Geometric isomorphism and minimum aberration for factorial designs with quantitative factors. *Annals of Statistics*, **32**, 2168–2185.

[23] Chipman, H., Hamada, M., Wu, C.F.J. (1997). A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, **39**, 372–381.

[24] Dean, A.M., Voss, D.T. (1999). Design and analysis of experiments. New York: Springer.

[25] Deng, L.Y., Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Statistica Sinica*, **9**, 1071–1082.

[26] Deng L.Y., Tang B. (2002). Design selection and classification for Hadamard matrices using generalized minimum aberration criteria. *Technometrics*, **44**, 173–184.

[27] Dey, A. (2005). Projection properties of some orthogonal arrays. *Statistics and Probability Letters*, **75**, 298–306.

[28] Dopico-Garcia, M.S., Valentao, P., Guerra, L., Andrade, P.B., Seabra, R.M. (2007). Experimental design for extraction and quantification of phenolic compounds and organic acids in white Vinho Verde grapes. *Analytica Chimica Acta*, **583**, 15–22.

[29] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499.

[30] Evangelaras, H., Koukouvinos, C., Dean, A.M., Dingus, C.A. (2005). Projection properties of certain three level orthogonal arrays. *Metrika*, **62**, 241–257.

[31] Evangelaras, H., Koukouvinos, C., Lappas, E. (2007). 18-run nonisomorphic three level orthogonal arrays. *Metrika*, **66**, 31–37.

[32] Fries, A., Hunter, W.G. (1980). Minimum aberration 2^{k-p} designs. *Technometrics*, **22**, 601–608.

[33] George, E.I., McMulloch, R.E. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

[34] Giles, J. (2006). Animal experiments under fire for poor design. *Nature*, **444**, 981.

[35] Hamada M., Wu C.F.J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, **24**, 130–137.

[36] Hedayat, A.S., Sloane, N.J.A., Stufken, J. (1999). Orthogonal Arrays: Theory and Applications. New York: Springer.

[37] Holcomb, D.R., Montgomery, D.C., Carlyle, W.M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, **35**, 13–27.

[38] Hammons, A.R., Jr., Kumar, P.V., Calderbank, A.R., Sloane, N.J.A. and Sole,P. (1994). The Z_4-linearity of Kerdock, Preparata, Goethals, and related codes. *IEEE Transactions on Information Theory*, **40**, 301–319.

[39] Hurvich, C.M., Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76(2)**, 297–307.

[40] Li, R., Lin, D.K.J. (2002). Data analysis in supersaturated designs. *Statistics and Probability Letters*, **59**, 135–144.

[41] Li, R., Lin, D.K.J. (2003). Analysis methods for supersaturated design: some comparisons. *Journal of Data Science*, **1**, 249–260.

[42] Lin, D.K.J. (1993). A new class of supersaturated designs. *Technometrics*, **35**, 28–31.

[43] Lin, D.K.J. (1999). Supersaturated designs. In *Encyclopedia of Statistical Sciences, updated volume 3*. 721–731. New York: Wiley.

[44] Lin, D.K.J., Draper, N.R. (1992). Projection properties of Plackett and Burman designs. *Technometrics*, **34**, 423–428.

[45] Lu, X., Wu, X. (2004). A strategy of searching active factors in supersaturated screening experiments. *Journal of Quality Technology*, **36**, 392–399.

[46] Ma, C.X., Fang, K.T. (2001). A note on generalized aberration in factorial designs. *Metrika*, **53**, 85–93.

[47] Mandal, A., Mukerjee, R. (2005). Design efficiency under model uncertainty for nonregular fractions of general factorials. *Statistica Sinica*, **15**, 697–707.

[48] Montgomery, D.C. (2005). Design and Analysis of Experiments (6th edition). New York: Wiley.

[49] Moreda-Pineiro, J., Alonso-Rodriguez, E., Lopex-Mahia, P., Muniategui-Lorenzo, S., Prada-Rodriguez, D., Moreda-Pineiro, A., Bermejo-Barrera, P. (2007). Development of a new sample pre-treatment procedure based on pressurized liquid extraction for the determination of metals in edible seaweed. *Analytica Chimica Acta*, **598**, 95–102.

[50] Mukerjee, R., Wu, C.F.J. (2006). A Modern Theory of Factorial Designs. New York: Springer.

[51] Paley, R.E.A.C. (1933). On orthogonal matrices. *Journal of Mathematical Physics*, **12**, 311–320.

[52] Phoa, F.K.H., Pan, Y.H., Xu, H. (2008). Analysis of supersaturated designs via the Dantzig selector. Accepted by *Journal of Statistical Planning and Inference*.

[53] Phoa, F.K.H., Xu, H. (2008). Quarter-fraction factorial designs constructed via quaternary codes. Accepted by *Annals of Statistics*.

[54] Plackett, R.L., Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

[55] Rao, C.R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society: Series B*, **9**, 128–139.

[56] Rao, C.R. (1973). Some combinatorial problems of arrays and applications to design of experiments. In *Survey of Combinatorial Theory*, Edited by Srivastava, J.N. pp. 349–359. Amsterdam: North-Holland.

[57] Sun, D.X., Li, W., Ye, K.Q. (2002). An algorithm for sequentially constructing nonisomorphic orthogonal designs and its applications. Technical report SUNYSB-AMS-02-13, Department of Applied Mathematics and Statistics, SUNY at Stony Brook.

[58] Sun, D.X., Wu, C.F.J. (1993). Statistical properties of Hadamard matrices of order 16. In *Quality Through Engineering Design*. Edited by Kui, W. 169–179. New York: Elsevier.

[59] Svensgaard D.J., Hertzberg R.C. (1994) Statistical methods for the toxicological evaluation of the additivity assumption as used in the environmental protection agency chemical mixture risk assessment guideline. In *Toxicology of Chemical Mixtures*. Edited by Yang, R.S.H. 599–640. Academic Press, San Diego, CA.

[60] Tang, B., Deng, L.Y. (1999). Minimum G_2-aberration for non-regular fractional factorial designs. *Annals of Statistics*, **27**, 1914–1926.

[61] Tsai, P.W., Gilmour, S.G., Mead, R. (2000). Projective three-level main effects designs robust to model uncertainty. *Biometrika*, **87**, 467–475.

[62] Tsai, P.W., Gilmour, S.G., Mead, R. (2004). Some new three-level orthogonal main effects plans robust to model uncertainty. *Statistica Sinica*, **14**, 1075–1084.

[63] Vander Heyden, Y., Jimidar, M., Hund, E., Niemeijer, N., Peeters, R., Smeyers-Verbeke, J., Massart, D.L., Hoogmartens, J. (1999). Determination of system suitability limits with a robustness test. *Journal of Chromatography A*, **845**, 145–154.

[64] Wang, J.C., Wu, C.F.J. (1995). A hidden projection property of Plackett-Burman and related designs. *Statistica Sinica*, **5**, 235–250.

[65] Westfall, P.H., Young, S.S., Lin, D.K.J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, **8**, 101–117.

[66] Wu, C.F.J., Hamada, M. (2000). Experiments: Planning, Analysis and Parameter Design Optimization New York: Wiley.

[67] Xu, H. (2005). Some nonregular designs from the Nordstrom and Robinson code and their statistical properties. *Biometrika*, **92**, 385–397.

[68] Xu, H., Cheng, S.W., Wu, C.F.J. (2004). Optimal projective three-level designs for factor screening and interaction detection. *Technometrics*, **46**, 280–292.

[69] Xu H., Wong A. (2007). Two-level nonregular designs from quaternary codes. *Statistica Sinica*, **17**, 1191–1213.

[70] Xu, H., Wu, C.F.J. (2001). Generalized minimum aberration for asymmetrical fractional factorial designs. *Annals of Statistics*, **29**, 1066–1077.

[71] Ye, K.Q. (2004). A note on regular fractional factorial designs. *Statistica Sinica*, **14**, 1069–1074.

[72] Yuan, M., Joseph, V.R., Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, **49**, 430–439.

[73] Zhang, Q.Z., Zhang, R.C., Liu, M.Q. (2007). A method for screening active effects in supersaturated designs. *Journal of Statistical Planning and Inference*, **137(9)**, 235–248.