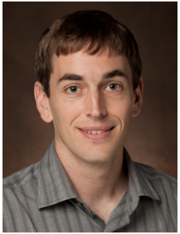
Analyzing Supersaturated Designs

Maria L. Weese

Information Systems & Analytics
Miami University, Oxford, OH

2017 Fall Technical Conference
October 6, 2017

Contributors



Byran Smucker
Miami University

David Edwards
Virginia Commonwealth University

Outline

Supersaturated Designs

Two-level supersaturated designs (SSDs) use $n < k + 1$ runs to examine k factors. For example, the Bayesian D-optimal design, D, uses $n = 6$ runs to examine $k = 9$ factors.

$$D = \begin{pmatrix} -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Supersaturated Designs

Another situation where "supersaturation" can occur is if the total number of *effects* that one wishes to examine, p, is greater than the number runs n. For example, the $n = 12$ and $k = 6$ two-level Bayesian D-optimal design,

$$D_1 = \begin{pmatrix} 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 \end{pmatrix}$$

includes another 15 columns if main effects and two-factor interactions are screened. Making the model matrix, $\mathbf{X}$, $n = 12$ by $p = 21 + 1$.

Notation

k=number of factors
n=number of runs
p=number of effects
D=design matrix
$\mathbf{X}$=model matrix
s_{ij}=off diagonal elements of $\mathbf{X}'\mathbf{X}$
a=number of truly active factors in simulation
S/N=signal to noise ratio for the truly active factors in simulation

$$\text{Power}=\frac{\text{Number of correctly identified active factors}}{a}$$

$$\text{Type I Error}=\frac{\text{Number inactive factors found to be active}}{(k-a)}$$

$$\text{FDR}=\frac{\text{Number of inactive factors found to be active}}{\text{Total Number of Effects found to be active}}$$

Outline

Effect Sparsity

Fitting the standard linear model $y = \mathbf{X}\beta + \epsilon$ is problematic.

Experimenters must operate under the assumption of effect sparsity to use a SSD as a screening experiment.

What is Effect Sparsity?

TECHNOMETRICS, FEBRUARY 1986, VOL. 28, NO. 1

Editor's Note: This article was presented at the *Technometrics* Session of the 29th Annual Fall Technical Conference of the American Society for Quality Control (Chemical and Process Industries Division and Statistics Division) and the American Statistical Association (Section on Physical and Engineering Sciences) in Corning, New York, October 24–25, 1985.

An Analysis for Unreplicated Fractional Factorials

George E. P. Box
Center for Quality and Productivity Improvement
University of Wisconsin
Madison, WI 53706

R. Daniel Meyer
Lubrizol Corporation
Wickliffe, OH 44092

Loss of markets to Japan has recently caused attention to return to the enormous potential that experimental design possesses for the improvement of product design, for the improvement of the manufacturing process, and hence for improvement of overall product quality. In the screening stage of industrial experimentation it is frequently true that the "Pareto Principle" applies; that is, a large proportion of process variation is associated with a small proportion of the process variables. In such circumstances of "factor sparsity," unreplicated fractional designs and other orthogonal arrays have frequently been effective when used as a screen for isolating preponderant factors. A useful graphical analysis due to Daniel (1959) employs normal probability plotting. A more formal analysis is presented here, which may be used to supplement such plots and hence to facilitate the use of these unreplicated experimental arrangements.

1. INTRODUCTION

Alarmed by foreign competition, management at last seems willing to heed those who have long advocated statistical design as a key to improvement of products and processes. The possible importance of fractional factorial designs in industrial applications seems to have been first recognized some 50 years ago (Tippett 1934; also see Fisher 1966, p. 88). Tippett successfully employed a 125th fraction of a 5^5 factorial as a screening design to discover the cause

explained by a small proportion of the process variables. This sparsity hypothesis has implications for both *design* and *analysis*.

Concerning the design aspect, consider, for example, an experimenter who desired to screen eight factors at two levels, believing that not more than three would be active. He might choose to employ a sixteenth replicate of a 2^8 design of resolution four. This 2_{IV}^{8-4} design has the property that every one of its $\binom{8}{3} = 56$ projections into three-space is a duplicated 2^3 factorial. Its use would thus ensure that the design

Sparsity

Factor Sparsity: Most process variation is driven by a few factors (Pareto Principle).

Effect Sparsity: Extends Factor Sparsity to contrasts.

What is Effect Sparsity in an SSD?

Computational Statistics and Data Analysis 54 (2010) 3158–3167

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A comparison of design and model selection methods for supersaturated experiments

Christopher J. Marley, David C. Woods*

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 31 January 2009
Received in revised form 4 February 2010
Accepted 18 February 2010
Available online 2 March 2010

Keywords:
Bayesian *D*-optimal designs
$E(s^2)$-optimal designs
Effect sparsity
Gauss–Dantzig selector
Main effects
Screening
Simulation

ABSTRACT

Various design and model selection methods are available for supersaturated designs having more factors than runs but little research is available on their comparison and evaluation. Simulated experiments are used to evaluate the use of $E(s^2)$-optimal and Bayesian *D*-optimal designs and to compare three analysis strategies representing regression, shrinkage and a novel model-averaging procedure. Suggestions are made for choosing the values of the tuning constants for each approach. Findings include that (i) the preferred analysis is via shrinkage; (ii) designs with similar numbers of runs and factors can be effective for a considerable number of active effects of only moderate size; and (iii) unbalanced designs can perform well. Some comments are made on the performance of the design and analysis methods when effect sparsity does not hold.

1. Introduction

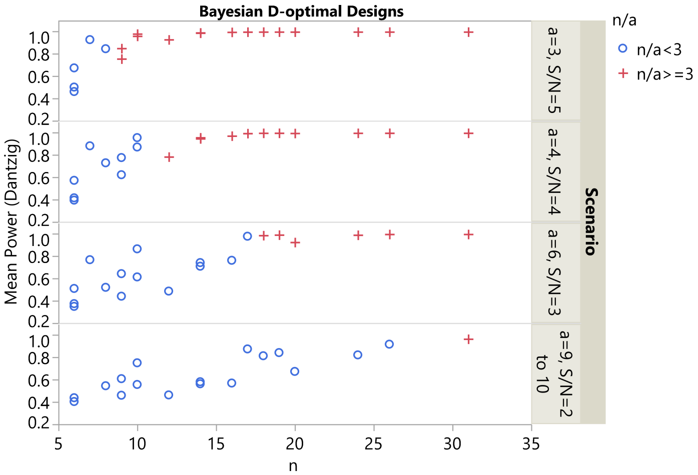
A screening experiment investigates a large number of factors to find those with a substantial effect on the response of interest, that is, the active factors. If a large experiment is infeasible, then using a supersaturated design in which the number of factors exceeds the number of runs may be considered. This paper investigates the performance of a variety of design and model selection methods for supersaturated experiments through simulation studies.

Supersaturated designs were first suggested by Box (1959) in the discussion of Satterthwaite (1959). Booth and Cox (1962) provided the first systematic construction method and made the columns of the design matrix as near orthogonal as possible through the $E(s^2)$ design selection criterion (see Section 2.1.1). Interest in design construction was revived by Lin (1993) and Wu (1993), who developed methods based on Hadamard matrices. Recent theoretical results for $E(s^2)$-optimal and highly efficient designs include those of Nguyen and Cheng (2008). The most flexible design construction methods are algorithmic: Lin (1995), Nguyen (1996) and Li and Wu (1997) constructed efficient designs for the $E(s^2)$ criterion. More recently, Ryan and Bulutoglu (2007) provided a wide selection of designs that achieved lower bounds on $E(s^2)$, and Jones et al. (2008) constructed designs using Bayesian *D*-optimality. For a review of supersaturated designs, see Gilmour (2006).

Sparsity

"The number of runs should be at least three times the number of active factors."

Number Active to n Ratio



How "Supersaturated" can a Design be?

Computational Statistics and Data Analysis 54 (2010) 3158–3167

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

ELSEVIER

A comparison of design and model selection methods for supersaturated experiments

Christopher J. Marley, David C. Woods*

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 31 January 2009
Received in revised form 4 February 2010
Accepted 18 February 2010
Available online 2 March 2010

Keywords:
Bayesian *D*-optimal designs
$E(s^2)$-optimal designs
Effect sparsity
Gauss–Dantzig selector
Main effects
Screening
Simulation

ABSTRACT

Various design and model selection methods are available for supersaturated designs having more factors than runs but little research is available on their comparison and evaluation. Simulated experiments are used to evaluate the use of $E(s^2)$-optimal and Bayesian *D*-optimal designs and to compare three analysis strategies representing regression, shrinkage and a novel model-averaging procedure. Suggestions are made for choosing the values of the tuning constants for each approach. Findings include that (i) the preferred analysis is via shrinkage; (ii) designs with similar numbers of runs and factors can be effective for a considerable number of active effects of only moderate size; and (iii) unbalanced designs can perform well. Some comments are made on the performance of the design and analysis methods when effect sparsity does not hold.

1. Introduction

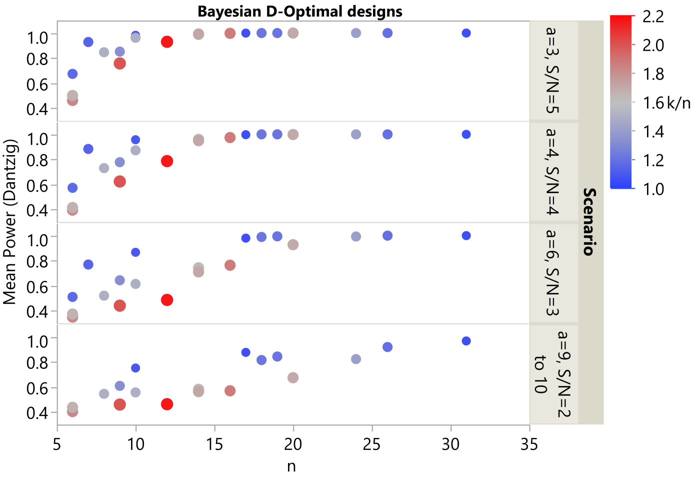
A screening experiment investigates a large number of factors to find those with a substantial effect on the response of interest, that is, the active factors. If a large experiment is infeasible, then using a supersaturated design in which the number of factors exceeds the number of runs may be considered. This paper investigates the performance of a variety of design and model selection methods for supersaturated experiments through simulation studies.

Supersaturated designs were first suggested by Box (1959) in the discussion of Satterthwaite (1959). Booth and Cox (1962) provided the first systematic construction method and made the columns of the design matrix as near orthogonal as possible through the $E(s^2)$ design selection criterion (see Section 2.1.1). Interest in design construction was revived by Lin (1993) and Wu (1993), who developed methods based on Hadamard matrices. Recent theoretical results for $E(s^2)$-optimal and highly efficient designs include those of Nguyen and Cheng (2008). The most flexible design construction methods are algorithmic: Lin (1995), Nguyen (1996) and Li and Wu (1997) constructed efficient designs for the $E(s^2)$ criterion. More recently, Ryan and Bulutoglu (2007) provided a wide selection of designs that achieved lower bounds on $E(s^2)$, and Jones et al. (2008) constructed designs using Bayesian *D*-optimality. For a review of supersaturated designs, see Gilmour (2006).

Design Size

"The ratio of factors to runs should be less than 2."

k to n Ratio


Bayesian D-Optimal designs
Mean Power (Dantzig)
1.0
0.8
0.6
0.4
n
5
10
15
20
25
30
35
a=3, S/N=5
a=4, S/N=4
a=6, S/N=3
a=9, S/N=2 to 10
Scenario
2.2
2.0
1.8
1.6 k/n
1.4
1.2
1.0

Outline

Construction Criteria

- $E(s^2)$-optimality [1]

$$E(s^2) = \frac{2}{k(k-1)} \sum_{2 \leq i < j} s_{ij}^2$$

- Bayesian $\mathcal{D}$-optimality [2]

$$\phi_D = |X'X + K/\tau^2|^{1/(1+k)}$$

where

$$K = \begin{pmatrix} 0 & \mathbf{0_{1\times k}} \\ \mathbf{0_{k\times 1}} & \mathbf{I_{k\times k}} \end{pmatrix}.$$

Construction Criteria, cont.

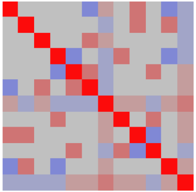
- Constrained Positive Var(s)-optimality [3]

$$Var(s) = E(s^2) - E(s)^2 = \frac{2}{k(k+1)} \sum_{1 \leq i < j} s_{ij}^2 - \left(\frac{2}{k(k+1)} \sum_{1 \leq i < j} s_{ij} \right)^2$$

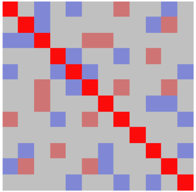
subject to

$$E_{E(s^2)} = \frac{E(s^2)(D^*)}{E(s^2)(D)} > c$$
$$E(s) > 0$$

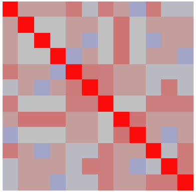
Why does construction matter for analysis?



Bayes D

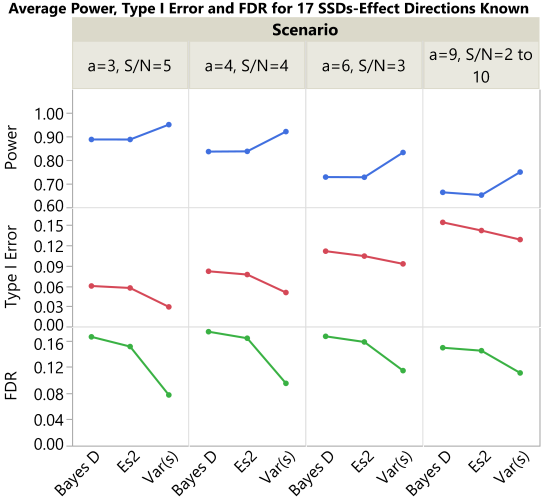


$E(s^2)$

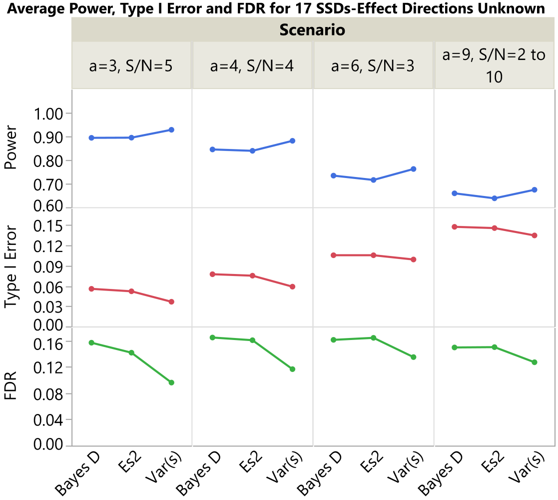


Var(s)

Structure Influences Analysis



Structure Influences Analysis



Outline

Interactions

Recall the $n = 12$ and $k = 6$ two-level Bayesian D-optimal design,

$$D_1 = \begin{pmatrix} 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 \end{pmatrix}$$

which is not a supersaturated design until you consider the addition of

Interactions

the 15 interaction columns:

$$X_2 = \begin{pmatrix}
1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\
-1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
-1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\
-1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\
-1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1
\end{pmatrix}$$

Interactions

Or we might have the case where the design matrix is already saturated, such as the n=12, k=16 design below, and interactions are to be considered in the analysis.

$$D_2 = \begin{pmatrix}
1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\
-1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\
-1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \\
-1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}$$

This would add 120 columns.

Construction of SSDs for Interactions

- Consider the entire all of the p columns in the model matrix, $\mathbf{X}$, simultaneously.
- Consider the k main effects and $\binom{k}{2}$ interaction columns in the model matrix, $\mathbf{X}$, separately.

Proposed Construction Criteria: Considering all Columns in **X**

- Bayesian D-optimality including 2fi's

$$\phi_D = |X'X + K/\tau^2|^{1/(1+p)}$$

where

$$K = \begin{pmatrix} 0 & \mathbf{0_{1\times p}} \\ \mathbf{0_{p\times 1}} & \mathbf{I_{p\times p}} \end{pmatrix}.$$

- Unbalanced $E(s^2)$-optimality including 2fi's

$$E(s^2) = \frac{2}{p(p+1)} \sum_{1\leq i<j} s_{ij}^2$$

Proposed Construction Criteria: Considering all Columns in **X**

- Constrained Positive Var(s)-optimality including 2fi's

$$Var(s) = E(s^2) - E(s)^2 = \frac{2}{p(p+1)} \sum_{1 \leq i < j} s_{ij}^2 - \left(\frac{2}{p(p+1)} \sum_{1 \leq i < j} s_{ij} \right)^2$$

subject to

$$E_{E(s^2)} = \frac{E(s^2)(D^*)}{E(s^2)(D)} > c$$

$$E(s) > 0$$

where D* is the unbalanced $E(s^2)$-optimal including 2fi design, see slide 25.

Proposed Construction Criteria: Main Effects and Interactions Separate

The $\mathbf{X}'\mathbf{X}$ matrix can be divided as such:

$$\mathbf{X}'\mathbf{X} = \left(\begin{array}{c|c} \text{Main Effects} & \text{Main Effects X Interactions} \\ \hline \text{Main Effects X Interactions} & \text{Interactions} \end{array}\right)$$

Proposed Construction Criteria: Main Effects and Interactions Separate

We use a $Var(s)$ minimization for the main effects and an $E(s^2)$ minimization for the other elements (Main Effects by Interactions and Interactions). The two criteria are weighted accordingly:

$$min: \quad \alpha Var(s) + (1-\alpha)E(s^2)$$

Proposed Construction Criteria: Main Effects and Interactions Separate

We use a $Var(s)$ minimization for the main effects and an $E(s^2)$ minimization for the other elements (Main Effects by Interactions and Interactions). The two criteria are weighted accordingly:

$$min : \quad \alpha Var(s) + (1-\alpha)E(s^2)$$

When α is based on the number of columns:

- Constrained Positive $Var(s)_{columns}$

Proposed Construction Criteria: Main Effects and Interactions Separate

We use a $Var(s)$ minimization for the main effects and an $E(s^2)$ minimization for the other elements (Main Effects by Interactions and Interactions). The two criteria are weighted accordingly:

$$min: \quad \alpha Var(s) + (1-\alpha)E(s^2)$$

When α is based on the number of columns:

- Constrained Positive $Var(s)_{columns}$

And when α=0.5:

- Constrained Positive $Var(s)_{equal}$

Proposed Construction Criteria: Main Effects and Interactions Separate

We use a $Var(s)$ minimization for the main effects and an $E(s^2)$ minimization for the other elements (Main Effects by Interactions and Interactions). The two criteria are weighted accordingly:

$$min : \quad \alpha Var(s) + (1 - \alpha) E(s^2)$$

When α is based on the number of columns:

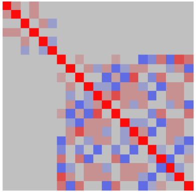
- Constrained Positive $Var(s)_{columns}$

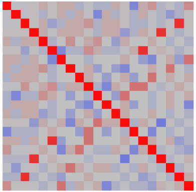
And when α=0.5:

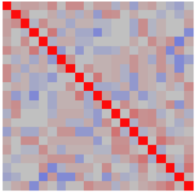
- Constrained Positive $Var(s)_{equal}$

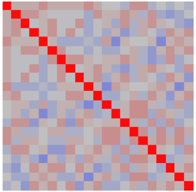
Both require a specified $E(s^2)$ efficiency, c and $E(s) > 0$.

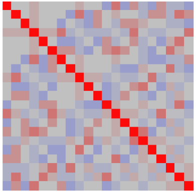
Comparing Construction Criteria


Bayes D 2fi

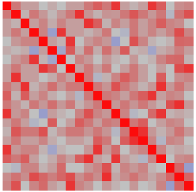

Unbal $E(s^2)$ 2fi


Var(s) 2fi(90)

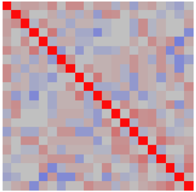

Var(s) Col Wts(80)


Var(s) Eq Wts(80)

Comparing Construction: Impact of c



Var(s) 2fi (20)



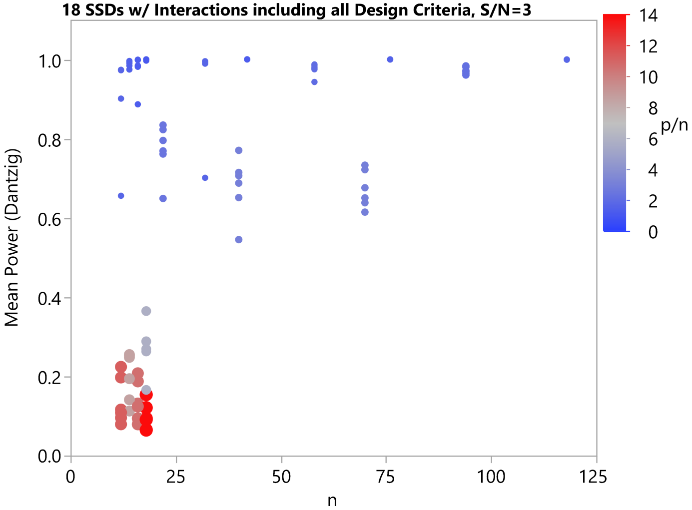
Var(s) 2fi (90)

Some Simulation Details

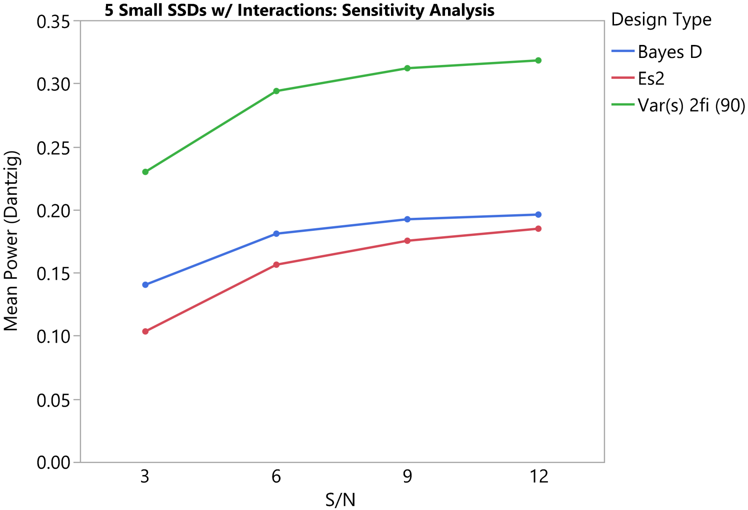
We based the number of active main effect and interaction columns on results of Li et al. [4]. For each of 10,000 iterations:

1. Randomly select m active main effect columns where $m = 0.41 * k$
2. Active interactions are chosen based on the following:
    - Strong heredity, P(AB active | A and B)=0.33
    - Weak heredity, P(AB active | A or B)=0.045
    - No heredity, P(AB active | Neither A or B)=0.0048
3. Assign a true effect size to the active columns where main effects are positive and the signs of interaction coefficients are negative with a probability of 0.17.
4. Inactive effect sizes were sampled from abs($N(0, 0.2)$) with signs flipped as described in number 2.

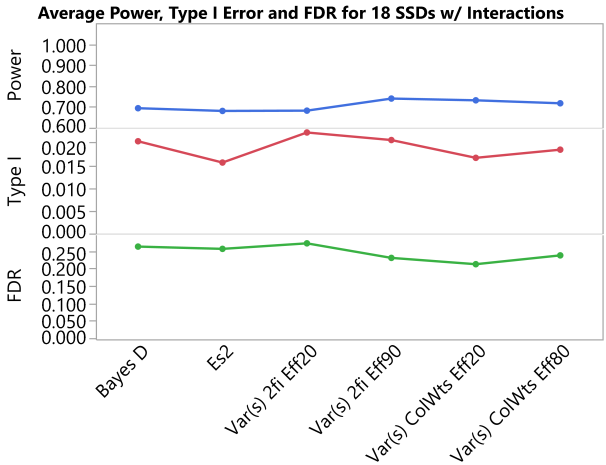
p to n Ratio


18 SSDs w/ Interactions including all Design Criteria, S/N=3
Mean Power (Dantzig)
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
n
14
12
10
8
6
4
2
0
p/n

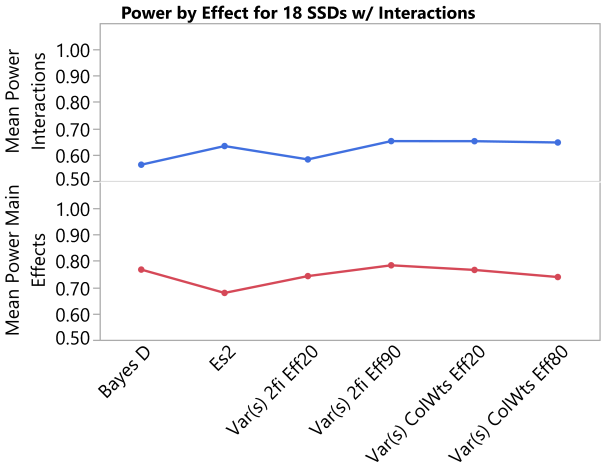
How big does the signal have to be?



Simulation Results (S/N=3)



Simulation Results (S/N=3)



Outline

Methods used to Analyze SSDs

Regression Methods

- Forward Selection [5]
- Stepwise Selection [6]
- All Subsets Regression [6]
- Singular value decomposition principal regression (SVDPR) [7]

Shrinkage Methods

- Dantzig Selector [8]
- LASSO [9]
- Smoothly Clipped Absolute Deviation (SCAD) [10]
- Sure Independence Screening (SIS) [11]

Other Methods

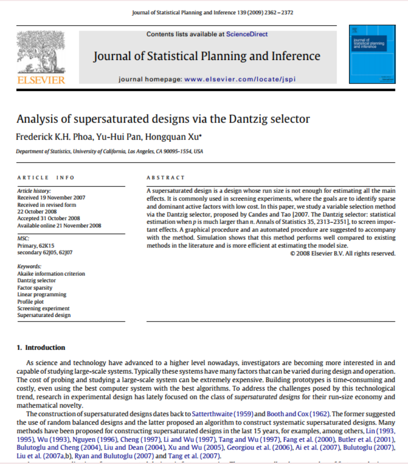
- Simulated Annealing (SA) [9]
- Model Averaging (MA) [12]
- Bayesian Methods (SVSS, CGS, SVSS/IBF) [13], [14]
- Partial Least Squares Variable Selection (PLSVS) [15]
- Stepwise Response Refinement Screener (SRRS) [16]

Results of Simulation Studies

A comparison of methods: "x" indicates the method was included in the study. "1" indicates best performer, "2" indicates the method out performed "1" under certain conditions.

Study	Forward Selection	Dantzig	Bayesian	LASSO	SCAD	SA	PLSVS	SVDPR	MA	SRRS
Marley and Woods (2010)	x	1							2	
Draguljić et al. (2014)		1		x	x	x				
Chen et al. (2013)		1	2 (CGS)		x		x			
Phoa (2014)		1	x		x		x	x		2
Weese et al. (2015)	x	1								
Weese et al. (2017)	x	1								

The Dantzig Selector



Journal of Statistical Planning and Inference 139 (2009) 2362–2372

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Analysis of supersaturated designs via the Dantzig selector

Frederick K.H. Phoa, Yu-Hui Pan, Hongquan Xu*

Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form
22 October 2008
Accepted 31 October 2008
Available online 21 November 2008

MSC:
Primary, 62K15
secondary 62J05, 62J07

Keywords:
Akaike information criterion
Dantzig selector
Factor sparsity
Linear programming
Profile plot
Screening experiment
Supersaturated design

ABSTRACT

A supersaturated design is a design whose run size is not enough for estimating all the main effects. It is commonly used in screening experiments, where the goals are to identify sparse and dominant active factors with low cost. In this paper, we study a variable selection method via the Dantzig selector, proposed by Candes and Tao [2007. The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics 35, 2313–2351], to screen important effects. A graphical procedure and an automated procedure are suggested to accompany with the method. Simulation shows that this method performs well compared to existing methods in the literature and is more efficient at estimating the model size.

1. Introduction

As science and technology have advanced to a higher level nowadays, investigators are becoming more interested in and capable of studying large-scale systems. Typically these systems have many factors that can be varied during design and operation. The cost of probing and studying a large-scale system can be extremely expensive. Building prototypes is time-consuming and costly, even using the best computer system with the best algorithms. To address the challenges posed by this technological trend, research in experimental design has lately focused on the class of *supersaturated designs* for their run-size economy and mathematical novelty.

The construction of supersaturated designs dates back to Satterthwaite (1959) and Booth and Cox (1962). The former suggested the use of random balanced designs and the latter proposed an algorithm to construct systematic supersaturated designs. Many methods have been proposed for constructing supersaturated designs in the last 15 years, for examples, among others, Lin (1993, 1995), Wu (1993), Nguyen (1996), Cheng (1997), Li and Wu (1997), Tang and Wu (1997), Fang et al. (2000), Butler et al. (2001), Bulutoglu and Cheng (2004), Liu and Dean (2004), Xu and Wu (2005), Georgiou et al. (2006), Ai et al. (2007), Bulutoglu (2007), Liu et al. (2007a,b), Ryan and Bulutoglu (2007) and Tang et al. (2007).

The Dantzig Selector

$\hat{\beta}$ is the solution to the l_1-regularization problem:

$$\min \|\hat{\beta}\|_1 \quad \text{s.t.} \quad \|\mathbf{X}^t(y - \mathbf{X}\hat{\beta})\|_\infty \leq \delta$$

The Automated Gauss-Dantzig Selector

$$\min \|\hat{\beta}\|_1 \quad \text{s.t.} \quad \|\mathbf{X}^t(y - \mathbf{X}\hat{\beta})\|_\infty \leq \delta \tag{1}$$

1. Let δ vary from 0 to $\delta_0 = max|x_i^t y|$ and where x_i is the i^{th} column of $\mathbf{X}$.
2. For each value of δ, solve the linear program in equation (1).
3. Coefficient estimates greater than a user specified threshold value, γ, are retained.
4. Fit a linear model using the factors retained in step (3) and calculate the value of the selection statistic (e.g. AICc, BIC, etc.)
5. The model at the value δ which produces the best value of the selection statistic is chosen.

Using the Dantzig Selector

Phoa et al. (2009) recommend using a Profile Plot of the coefficient estimates vs. δ to find the important factors in a single experiment.

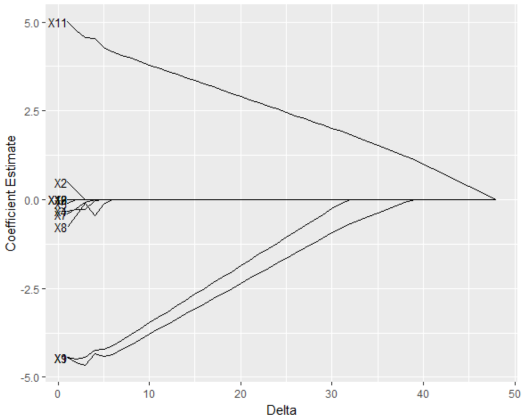
Using the Dantzig Selector

Phoa et al. (2009) recommend using a Profile Plot of the coefficient estimates vs. δ to find the important factors in a single experiment.

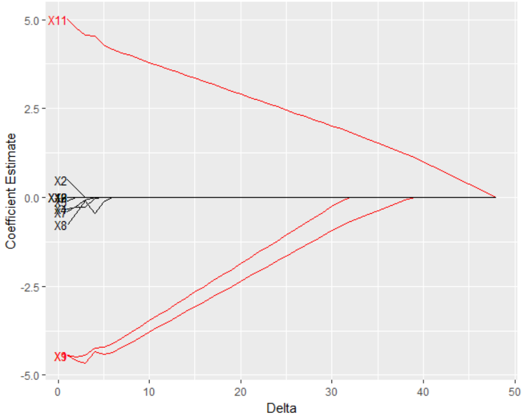
Using the automated procedure on slide 40 is not recommend for use in a single experiment analysis for the following reasons:

1. The specification of γ.
2. The choice of δ.

Example 1: Easy

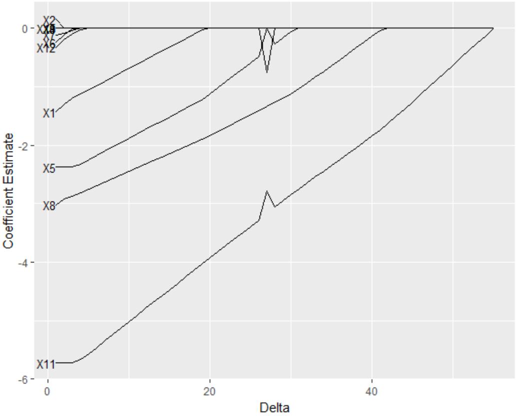


Example 1: Easy

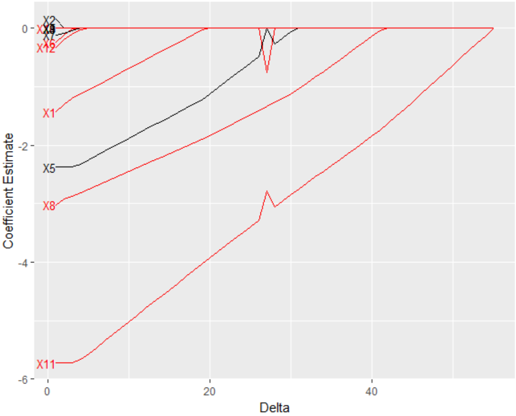


- Design: $n = 8$, $k = 12$ constrained-positive Var(s)-optimal with $c = 0.8$
- $a = 3$, $S/N = 5$ with $\pm$ assigned randomly.
- Inactive coefficients sampled from $N(0, 0.2)$
- $\delta = 0$ to $\delta_0 = max(|x_i' y|)$.

Example 2: Effect Directions Unknown

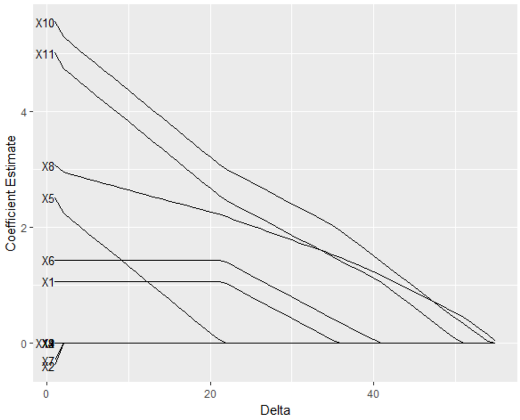


Example 2: Effect Directions Unknown

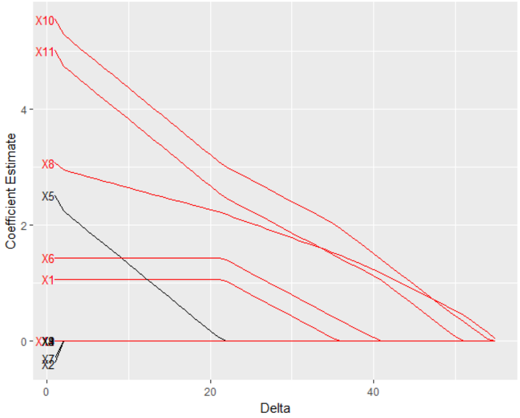


- Design: $n = 8$, $k = 12$ constrained-positive Var(s)-optimal with $c = 0.8$
- $a = 6$, $S/N = 3$ with $\pm$ assigned randomly.
- Inactive coefficients sampled from $N(0, 0.2)$

Example 2: Effect Directions Known

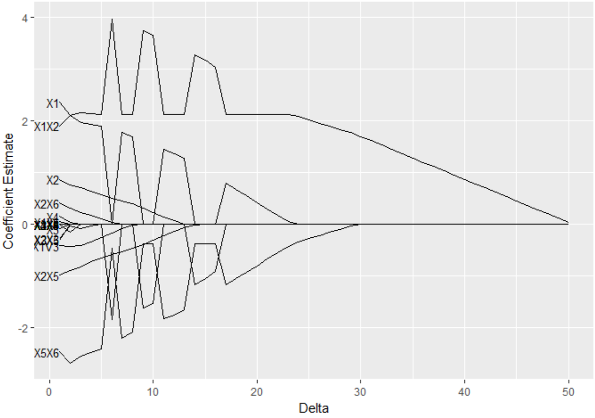


Example 2: Effect Directions Known

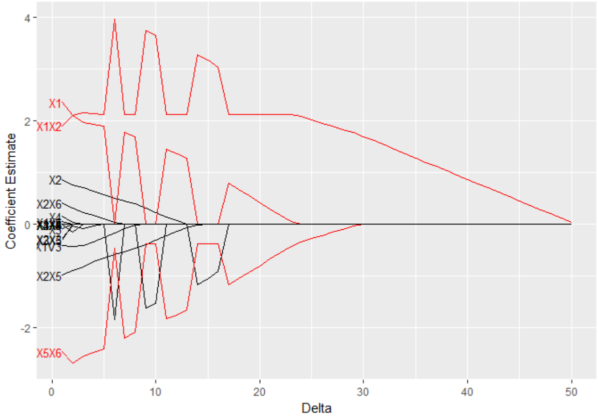


- Design: $n = 8$, $k = 12$ Constrained-positive Var(s)-optimal with $c = 0.8$
- $a = 6$, $S/N = 3$ now with *all positive* signs.
- Inactive coefficients sampled from abs($N(0, 0.2)$)

Example 3: Interactions Included

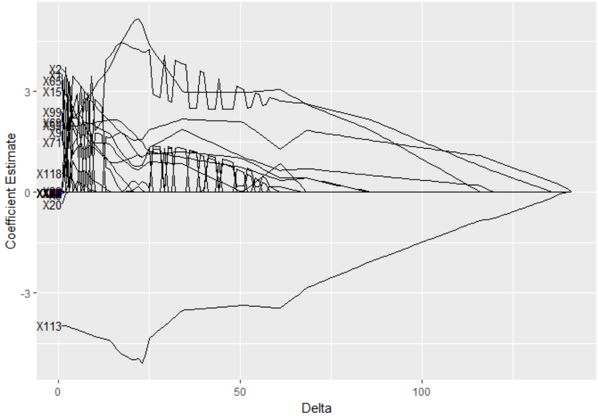


Example 3: Interactions Included

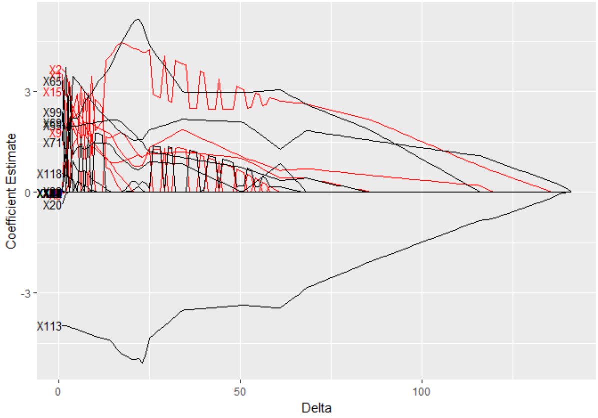


- Design: $n = 12$, $k = 6$ Constrained-positive 2fi Var(s)-optimal with $c = 0.9$
- Active effects generated according to probabilities on slide 31 and $S/N = 3$.
- Assume main effect directions are known.
- Inactive coefficients sampled from abs($N(0, 0.2)$) with signs assigned as described on slide 31.

Example 4: Interactions Included



Example 4: Interactions Included



- Design: $n = 14$, $k = 15$ Constrained-positive 2fi Var(s)-optimal with $c = 0.9$
- Active effects generated according to probabilities on slide 31 and $S/N = 3$.
- Assume main effect directions are known.
- Inactive coefficients sampled from abs($N(0, 0.2)$) with signs assigned as described on slide 31.

Outline

Conclusion

1. What size SSD is reasonable?
    - $k/n < 2$ is a good rule of thumb.
    - Evidence is in favor of $n/a \geq 3$.
    - For designs including interactions consider the p to n ratio.

Conclusion

1. What size SSD is reasonable?
   - $k/n < 2$ is a good rule of thumb.
   - Evidence is in favor of $n/a \geq 3$.
   - For designs including interactions consider the p to n ratio.

2. How should the design be constructed?
   - For consideration of main effects
     - Using the constrained-positive $Var(s)$-optimality with $c = 0.8$.
     - Attempt to guess your effect directions a priori.
     - Even all effect directions are misspecified, performance will be equivalent to using a Bayesian D-optimal or a balanced $E(s^2)$-optimal design [3].
   - For consideration of main effects and interactions
     - Design size is more important than construction method.

Conclusion

1. What size SSD is reasonable?
   - $k/n < 2$ is a good rule of thumb.
   - Evidence is in favor of $n/a \geq 3$.
   - For designs including interactions consider the p to n ratio.

2. How should the design be constructed?
   - For consideration of main effects
     - Using the constrained-positive $Var(s)$-optimality with $c = 0.8$.
     - Attempt to guess your effect directions a priori.
     - Even all effect directions are misspecified, performance will be equivalent to using a Bayesian D-optimal or a balanced $E(s^2)$-optimal design [3].
   - For consideration of main effects and interactions
     - Design size is more important than construction method.

3. How should the experiment be analyzed?
   - Use the Dantzig selector and a Profile Plot.

Thank you for listening!

weeseml@miamioh.edu

References

[1] K. H. V. Booth and D. R. Cox.
Some systematic supersaturated designs.
Technometrics, 4(4):489–495, 1962.

[2] Bradley Jones, Dennis K. J. Lin, and Christopher J Nachtsheim.
Bayesian D-optimal supersaturated designs.
Journal of Statistical Planning and Inference, 138(1):86–92, 2008.

[3] Maria L Weese, David J Edwards, and Byran J Smucker.
Powerful supersaturated designs when effect directions are known.
Journal of Quality Technology, 49(3):265–277, 2017.

[4] Xiang Li, Nandan Sudarsanam, and Daniel D Frey.
Regularities in data from factorial experiments.
Complexity, 11(5):32–45, 2006.

[5] Peter H Westfall, S Stanley Young, and Dennis KJ Lin.
Forward selection error control in the analysis of supersaturated designs.
Statistica Sinica, pages 101–117, 1998.

[6] B Abraham, H Chipman, and K Vijayan.
Some risks in the construction and analysis of supersaturated designs.
Technometrics, 41(2):135–141, 1999.

[7] Stelios D Georgiou.
Modelling by supersaturated designs.
Computational Statistics & Data Analysis, 53(2):428–435, 2008.

References (cont.)

[8] F.K.H. Phoa, Yu-Hui Pan, and H. Xu.
Analysis of supersaturated designs via the Dantzig Selector.
Journal of Statistical Planning and Inference, 139:2362–22372, 2009.

[9] D. Draguljić, D. C. Woods, A. M. Dean, S. M. Lewis, and A.-J. E. Vine.
Screening strategies in the presence of interactions.
Technometrics, 56(1):1–16, 2014.

[10] Runze Li and Dennis KJ Lin.
Analysis methods for supersaturated design: some comparisons.
Journal of Data Science, 1(3):249–260, 2003.

[11] Lindsey Nicely.
Applications of sure independence screening analysis for supersaturated designs.
2012.

[12] Christopher J Marley and David C Woods.
A Comparison of design and model selection methods for supersaturated experiments.
Computational Statistics and Data Analysis, 54:3158–3167, 2010.

[13] H. Chipman, M. Hamada, and C. F. J. Wu.
Bayesian variable selection for designed experiments with complex aliasing.
Technometrics, 39:372–381, 1997.

[14] Ray-Bing Chen, Jian-Zhong Weng, and Chi-Hsiang Chu.
Screening procedure for supersaturated designs using a bayesian variable selection method.
Quality and Reliability Engineering International, 29(1):89–101, 2013.

References (cont.)

[15] Qiao-Zhen Zhang, Run-Chu Zhang, and Min-Qian Liu.
A method for screening active effects in supersaturated designs.
Journal of Statistical Planning and Inference, 137(6):2068–2079, 2007.

[16] Frederick Kin Hing Phoa.
The stepwise response refinement screener (srrs).
Statistica Sinica, 2013.

[17] G. S. Watson.
A study of the group screening method.
Technometrics, 3(3):371–388, 1961.

[18] B. Bettonvil and J. P. C. Kleijnen.
Searching for important factors in simulation models with many factors: Sequential bifurcation.
European Journal of Operational Research, 96(1):180–194, 1996.