2021

Topics In Design and Analysis of Experiments: Calibration, Sequential Experimentation, and Model Selection

Christine Miller
Virginia Commonwealth University

DISSERTATION ON TOPICS IN DESIGN AND ANALYSIS OF EXPERIMENTS: CALIBRATION, SEQUENTIAL EXPERIMENTATION, AND MODEL SELECTION

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

CHRISTINE MILLER

Master of Science in Mathematical Sciences with Virginia Commonwealth University

Director: David J. Edwards,

Professor and Chair, Department of Statistical Sciences and Operations Research

Virginia Commonwewalth University

Richmond, Virginia

December, 2021

Acknowledgements

I would like to thank my advisor Dr. David Edwards for his mentorship and guidance. It is truly an honor and an incredible opportunity to work alongside him. I am extremely grateful to Dr. D'Arcy Mays for inspiring me since the beginning of my Masters program, for serving as my Masters program advisor, and for welcoming me back years later when I reached out to him about returning to pursue the Ph.D. program. His passion for statistics is truly contagious. Many thanks to my committee members Dr. David Edwards, Dr. D'Arcy Mays, Dr. Jason Merrick, and Dr. Yanjun Qian for their time and contributions.

I'd like to thank my husband Chris Miller for supporting me in my journey to become another Dr. Chris Miller. I am also extremely grateful to my family and friends for their encouragement and support.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Abstract

DISSERTATION ON TOPICS IN DESIGN AND ANALYSIS OF EXPERIMENTS: CALIBRATION, SEQUENTIAL EXPERIMENTATION, AND MODEL SELECTION

By Christine Miller

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2021.

Director: David J. Edwards,
Professor and Chair, Department of Statistical Sciences and Operations Research

Keywords : Inverse Prediction, Space-Filling Designs, Bayesian D-Optimality, Follow-Up Runs, Screening Designs, Pure Error, Power, False Discovery Rates, Forward Selection, LASSO, Bonferroni

0.1 Dissertation Topics

0.1.1 A Comparison of Experimental Designs for Calibration

The impact of experimental design choice on the performance of statistical calibration is largely unknown. We investigate the performance of several experimental designs with regards to inverse prediction via a comprehensive simulation study.

Specifically, we compare several design types including traditional response surface designs, algorithmically generated variance optimal designs, and space-filling designs.

0.1.2 Optimal Sequential Design Techniques

Uncertainty remains in optimal design techniques regarding the best way to allocate a given set of runs. The focus on maximizing information in optimal design has emphasized the running of a comprehensive large design all at one time, with or without replication. In practice, it may be better to first run a small screening design to identify important factors followed by an additional design building off knowledge gained in the first phase. We use simulations to compare the performance of D-optimal screening designs with follow up runs selected by Bayesian D-optimal augmentation against the performance of a nonsequential D-optimal design. We consider two and three-level screening designs. Simulation scenarios vary the number of total runs, analysis method, number of active main effects, two-factor interactions and quadratic terms.

0.1.3 Model Selection with Pure Error

Identifying important factors from a screening design is often challenging due to the aliasing of factorial effects from small run sizes and limited degrees of freedom to estimate all parameters. Studies on model selection for screening designs find higher power rates when statistical inference during model selection is based on a pure error estimate versus the residual mean square; however false discovery and Type I error rates can be problematically inflated. Currently there is not a suitable method available to incorporate pure error into model selection procedures when analyzing screening designs that achieves high power without the trade-off of high false discovery rates. To counteract the lack of noncentrality in the partial F-test denominator

contributing to larger partial F-tests with pure error, we consider early stopping methods including Bonferroni adjusted p-values and our proposed forward selection method incorporating lack of fit tests after each model selection step. Additionally, we develop a method of incorporating pure error with LASSO penalized regression by computing lack of fit tests along the LASSO solution path. We examine various model selection techniques using a simulation study and propose a strategy for incorporating pure error in model selection procedures that keeps FDR in check.

CHAPTER 1

INTRODUCTION

Experiments are widely used across multiple disciplines to uncover information about a system or processes. For instance, a goal may be to discover which variables (or factors) influence a response and to measure their impact. To achieve the goal, the experiment must be set up so the appropriate data points are captured. Experimental design is a statistical technique devoted to the methodology of selecting the appropriate samples to aid in the subsequent analysis. An effectively designed experiment helps uncover information needed to optimize a process spurring innovation. Incorporating design of experiments allows one to innovate either with new discoveries or in an incremental approach.

The statistical field of Response Surface Methodology focuses on the development and optimizing of processes using statistical techniques (Myers et al., 2016). Optimization seeks to find the minimum or maximum value, visually represented by dips or peaks if we were to map out response levels by independent variables. Myers et al. (2016) define response surface as the graphical perspective of the problem environment. For instance, consider the following response surface application: a chemical process has a goal to increase yield by altering reagent concentration and temperature. In this scenario, the independent variables (reagent concentration and temperature) are known as factors and yield is known as the response. Factors are variables that the experiment varies to measure their impact on a response, and they are commonly represented as x_j for $j = 1, ..., h$ factors with response y_i for $i = 1, .., n$ runs. Say either the reagent concentration, x_1, can either be low or high and the temperature, x_2 can be either low or high. In a design, these levels may be coded as 1 for high and -1 for the low value placing the variables on the same scale. Running both factors at their low value results in a run with each factor labeled -1 and the corresponding yield, y_i is recorded. A common experimental design is a fractional factorial, 2^h, that includes all combinations of each factor at 2 levels, their high and low value. This allows the researcher to determine the impact a change in each factor has on the response. For example, the 2^h factorial design for 2 factors each with two levels -1 and 1 has the 4 points (-1,-1), (-1,1), (1,-1), and (1,1).

Statistical models fitted on the collected experiment data provide information about the underlying system. Continuing with the chemical process example, x_1 and x_2 are main effect terms measuring how each factor independently impacts the response. A simple linear relationship between factors and response can be represented as the main effect model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \tag{1.1}$$

where β_1, β_2 are unknown parameters for each coefficient, β_0 is the intercept, and ϵ_i is an error term distributed normally with mean $E(\epsilon) = 0$ and variance $Var(\epsilon) = \sigma^2$. The model can also be represented in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.2}$$

where $\boldsymbol{y}$, $\boldsymbol{\epsilon}$ are $n \times 1$ vectors, $\boldsymbol{\beta}$ is a $(h+1) \times 1$ column vector and $\boldsymbol{X}$ is the $n \times (h+1)$ model matrix for n experiment runs.

The subsequent analysis of the data tests if terms have an impact on the response and estimates the $\boldsymbol{\beta}$ coefficients. If the practitioner expects that the reagent concentration interacts with temperature, adding two-factor interaction terms to the model will allow them to test for this behavior. Quadratic terms can be added if curvature is expected in the response surface. The full quadratic model takes the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \epsilon_i. \tag{1.3}$$

Three levels of each factor are needed to fit a quadratic model and detect curvature. Using our chemical process example, a center value of 0 can be added for each factor. The fractional factorial is now a 3^h design and requires additional runs and hence resources.

Effect sparsity and heredity are two key assumptions of experimental design. Effect sparsity assumes only a fraction of main effects and low order interactions are truly active. For instance, consider our quadratic model in equation 1.3, effect sparsity assumes only a few of the 5 terms are actually active, say just x_1 and x_1x_2. Notice in this example the interaction term x_1x_2 is active but only one of its parent main effects x_1 and x_2 is also active. This brings up our next key assumption of heredity. Heredity requires parent main effects to also be active if interactions and quadratic terms are present in the data (Edwards et al., 2014). Strong heredity requires all parent main effects x_1 and x_2 to be active if their interaction x_1x_2 is also active. Weak heredity requires at least one of the parent main effects to be active as shown in the sparsity example with active effects x_1 and x_1x_2.

We use simulation in each chapter of this dissertation to compare designs and techniques. This enables us to control which factors are active or inactive in the underling data. We can then appropriately determine which designs and techniques most adequately describe the true model. We assume either a full quadratic model or a model with main effects and two-factor interactions, and we randomly select which terms are active as well as their true coefficients. Various metrics are employed to evaluate performance. An Analysis of Variance test (ANOVA) determines differences among means. The mean Euclidean distance between actual and predicted input value is compared to evaluate calibration performance in Chapter 2. To measure how accurately methods classify significant variables, Chapters 3 and 4 use power, the proportion of active terms that are correctly identified, false discovery rate (FDR), the proportion of terms incorrectly identified as active, and Type I error, the proportion of inactive terms declared active.

1.1 Research Objectives

The primary goal of this dissertation is to provide practitioners practical recommendations regarding experimental designs. Topics include design selection in calibration, optimal sequential design techniques, and model selection with pure error. Simulation results in Chapters 2, 3 and 4 allow us to compare various techniques in correctly identifying inputs in the underlying functional form. In Chapter 2, we evaluate the impact of experimental designs on calibration by measuring how accurately designs can infer input values. Our goal is to provide specific design recommendations for calibration as well as detail the influence the number of responses, the number of inputs and other design choices have on the final results. In Chapter 3, we explore how a sequential design followed up with additional runs compares to a nonsequential design. We focus on determining the best way to apportion a given set of runs. In Chapter 4, we research the impact of statistical inference based a pure error estimate versus the residual mean square during model selection procedures. We aim to propose a strategy to incorporate pure error into model selection procedures when analyzing screening designs that achieves high power without the trade-off of high FDR.

1.2 Dissertation Outline

The three main topics of this dissertation are presented in Chapters 2, 3 and 4. Each chapter contains background information and a literature review for each topic. The literature review is followed by a research methodology section and simulation results.

CHAPTER 2

A COMPARISON OF EXPERIMENTAL DESIGNS FOR CALIBRATION

2.1 Background and Literature Review

Calibration, also referred to as inverse prediction, is a statistical technique that uses experimental data to model the relationship between input and response variables in order to ultimately infer inputs based on newly observed response values (Brown, 1993; Sundberg, 1999; Thomas et al., 2017). For example, Anderson-Cook et al. (2015) describes a calibration study originating in nuclear forensics. When a prohibited nuclear material such as plutonium is found, substantial effort is placed into investigating its origin. Deducing the processing parameters (i.e. inputs) such as reaction temperature and reagent concentration from the observed material provides indicators of the facility used to create the nuclear material. Reagents used during production can leave behind specific impurities that may then be linked to the reagent.

There are two primary techniques of conducting statistical calibration: forward modeling and direct inverse modeling; see Lewis et al. (2018) for a comparison of these two methods. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_h)$ represent h input variables and $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)$ represent k responses. For the forward modeling approach, models such as $\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon$, where $f(.)$ is some specified functional form and ϵ is a random error term, are established based on training/experimental data. These forward models are then "inverted", often using nonlinear optimization techniques, to determine a set of inputs for a given set of new response values.

Direct methods, on the other hand, build a model for $\boldsymbol{x}$ as a function of $\boldsymbol{y}$ and often employ dimension reduction techniques such as principal component regression and partial least squares (Haaland and Thomas, 1988). Xiang et al. (2009) notes the potential for *chance correlation* (correlations occurring in a sample due to variability that do not actually exist in the population (Rönkkö, 2014)) when using partial least squares during calibration and demonstrates this via a pharmaceutical application. Lewis et al. (2018) reports forward modeling consistently outperforms direct inverse modeling with respect to prediction accuracy whereas direct inverse modeling proves most beneficial when there are many more responses than input variables and dimension reduction is needed.

Returning to the nuclear forensics example, Anderson-Cook et al. (2015) exemplify the forward approach of calibration using particle morphology measurements of plutonium to infer processing reaction temperature and reagent concentration. The forward model parameter estimates are obtained from a previous plutonium oxide (PuO_2) experiment with three factors (nitric acid (HNO_3) in moles, Pu in g/L, and temperature in °C) and five response variables describing the resulting particle (agglomeration index in weight percent, mode, length, width, and thickness in μm). For each response ($i = 1, 2, ..., 5$), the maximal forward model is specified by

$$\begin{aligned} y_i = \beta_{i0} + \beta_{i1}x_1 + \beta_{i2}x_2 + \beta_{i3}x_3 + \beta_{i12}x_1x_2 + \beta_{i13}x_1x_3 + \\ \beta_{i23}x_2x_3 + \beta_{i123}x_1x_2x_3 + \beta_{i11}x_1^2 + \beta_{i22}x_2^2 + \beta_{i33}x_3^2 + \epsilon_i \end{aligned} \tag{2.1}$$

with $\epsilon_i \sim Normal(0, \sigma_i^2)$. When a new response $\boldsymbol{y}^* = (y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)$ is observed following an experiment, we seek the input $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3)$ that minimizes some distance metric (e.g. Euclidean distance) between $\boldsymbol{y}^*$ and the predicted response

$\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5)$ where

$$\begin{aligned}\hat{y}_i = \hat{\beta}_{i0} + \hat{\beta}_{i1}x_1 + \hat{\beta}_{i2}x_2 + \hat{\beta}_{i3}x_3 + \hat{\beta}_{i12}x_1x_2 + \hat{\beta}_{i13}x_1x_3 + \\ \hat{\beta}_{i23}x_2x_3 + \hat{\beta}_{i123}x_1x_2x_3 + \hat{\beta}_{i11}x_1^2 + \hat{\beta}_{i22}x_2^2 + \hat{\beta}_{i33}x_3^2.\end{aligned} \tag{2.2}$$

Training data for the forward model is often based on an experimental design; however, there is limited literature on appropriate experimental designs for inverse prediction. It is natural, then, to question how and/or why the choice of experimental design may impact the accuracy of inverse predictions. When inverse prediction is to be based on forward models, it is a sensible strategy to seek designs that provide precise estimation of the forward model parameters. In fact, Brown (1993) and Sundberg (1996, 1999) show that confidence regions for $\hat{\boldsymbol{x}}$ depend on the standard errors of the forward model's parameter estimates. This provides some evidence that the judicious choice of experimental design should have an impact on inverse prediction performance. Furthermore, it is intuitive that designs focused on precise prediction of new observations over the experimental region would result in a rich training data set that is useful for inverse prediction.

François et al. (2004) compare design performance for univariate calibration models. These authors propose two new optimization criteria for nonlinear models based on G-optimality (i.e. minimize the maximum prediction variance) and I-optimality (i.e. minimize the average prediction variance over the design region). Comparing these two criteria, their simulation results support the use of I-optimality as the G-criterion placed higher weights on areas where inverse prediction variance was high. Bondi et al. (2012) compare Central composite designs (CCDs), D-optimal, and I-optimal designs in a calibration study using the direct method of partial least squares. Results favor the use of both I-optimal designs and CCDs. Anderson-Cook et al. (2015) also suggest the use of I-optimal designs but caution that research regarding the development of a “direct metric for assessing the quality of a design for the inverse problem” is needed.

To the best of our knowledge, no study currently exists that comprehensively compares experimental designs for multivariate calibration over various scenarios. In this chapter, we investigate the performance of select experimental designs by evaluating inverse prediction accuracy. Specifically, we consider traditional response surface designs as well as algorithmically generated variance optimal designs and space-filling designs. Our simulation study varies seven different attributes (totaling 19,200 different scenarios) including design type, number of inputs (factors), number of response variables, size of error variance, correlation among response variables, percent of model terms that are active, and whether or not the forward model is reduced in some way before performing inverse prediction. Note that some degree of correlation among response variables is expected as the responses may consist of various measurements from the same final product. For instance, with the chemometrics application of particle morphology, multiple measurements of the same output are taken to describe irregular forms. Lewis et al. (2018) caution that substantial relationships between the responses may hinder inference on inputs.

2.2 Experimental Designs

The following subsections provide brief descriptions of the experimental designs that we consider for comparison in this study.

2.2.1 Central Composite Designs

Central composite designs (CCDs) are easily the most commonly used design for response surface exploration and optimization (Myers et al., 2016) and are quite popular in chemometrics (for example, see Panagiotou et al. (2009), Loveday et al.

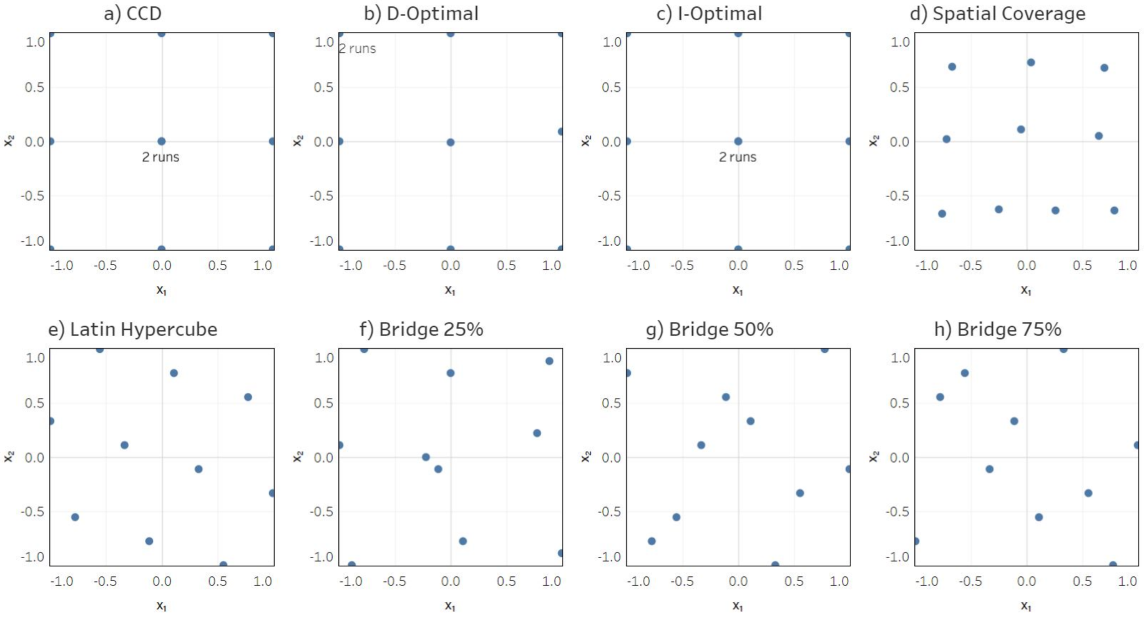


Fig. 1.: Design point locations for 2 factors and 10 runs

(2011), Asadollahzadeh et al. (2014), and Rai et al. (2016) among many others). This popularity arises from their ability to i) be run sequentially and ii) fit a second-order model without the run size requirements of a 3^h factorial design. In general, for h factors, CCDs consists of a 2^h full factorial or Resolution V (or higher) $2^{h-\ell}$ fractional factorial design, $2h$ axial points, and some number (n_0) of center point runs. Figure 1(a) illustrates a 10-run CCD for 2 factors. For straightforward comparison with other designs, we specify an "on face" (i.e. $\alpha = 1$) axial distance. CCD design points fall along the boundary of the experimental region with the exception of the center point runs.

2.2.2 *D*-optimal Designs

D-optimal designs seek to minimize the variance of the model parameter estimates (Myers et al., 2016). This criterion has proven to be particularly useful in screening scenarios where the objective is to identify the few important factors out of the many factors of interest. As the determinant can be used as a measure of overall variance, a D-optimal design locates a set of design points that minimizes the determinant of the inverse information matrix (i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ where $\boldsymbol{X}$ is the model matrix). It is computationally more tractable, however, to maximize the determinant of the information matrix. Thus, a D-optimal design is such that

$$|\boldsymbol{X}'\boldsymbol{X}| \tag{2.3}$$

is maximized. Assuming specification of a full second-order model, Figure 1(b) shows the design locations for a D-optimal design in 2 factors and 10 runs. This design consists of one center point run with all other points located on the boundary of the design region. In this example, the D-optimal design is similar to the CCD except that it contains only one center run and two replicate runs at the point [-1,1]. Like the CCD, D-optimal designs appear prominently in the chemometrics literature (e.g. François et al. (2007), Mannarswamy et al. (2009), Ebrahimi-Najafabadi et al. (2012), Bella et al. (2018)).

2.2.3 I-optimal Designs

The prediction variance at some design location $\boldsymbol{x}$ is given by

$$v(\boldsymbol{x}) = n\boldsymbol{f}(\boldsymbol{x})'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{f}(\boldsymbol{x}) \tag{2.4}$$

where $\boldsymbol{f}(\boldsymbol{x})$ takes a factor setting $\mathbf{x}$ and expands it to its corresponding model terms (Anderson-Cook et al., 2009). For example, for a second-order model,

$$\boldsymbol{f}(\boldsymbol{x})' = [1, x_1, x_2, ..., x_h, x_1x_2, ..., x_{h-1}x_h, x_1^2, ..., x_h^2]. \tag{2.5}$$

An I-optimal design minimizes the average prediction variance over the design region (denoted as χ) given by

$$\frac{1}{K}\int_{\chi} \upsilon(\boldsymbol{x})d\boldsymbol{x} \tag{2.6}$$

where $K = \int_{\chi} d\boldsymbol{x}$ is the volume of the design region. Thus, the I-criterion can be written as

$$\min\left\{\frac{1}{K}\int_{\chi} \boldsymbol{x}^{(m)\prime}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^{(m)}d\boldsymbol{x}\right\}. \tag{2.7}$$

Assuming specification of a second-order model, the 10-run I-optimal design in Figure 1(c) is identical to the CCD. For examples of usage of I-optimal designs in the chemometrics literature, see Jeirani et al. (2012) and Akkermans et al. (2017).

2.2.4 Space-Filling Designs

We expand our calibration study to include space-filling designs as they allocate points throughout the design space (see Nunes et al. (2013), Hathurusingha and Davey (2016), and Xu et al. (2018) for examples from chemometrics). Variance-based optimal designs, such as I- and D-optimal designs require *a priori* specification of the underlying model form (i.e. before data collection). This is often cited as a downside to the use of such criteria. Space-filling designs are an alternative to traditional designs and can allow for a more effective exploration of the interior of the experimental region as well as more flexibility in model fitting. Space-filling designs, popularly used for deterministic or near-deterministic models (such as with computer experiments), select n points from a discrete grid of candidate points over the design region, χ, based on some specified distance function d.

One approach for constructing a space-filling design is to select design points, $\boldsymbol{x}_i$, throughout the design region such that the maximum distance from any location in the design region to its nearest point in the design is minimized. These designs are known as minimax designs since the maximum distance from a candidate design point ($\boldsymbol{x}_i$) to its nearest neighbor in the experimental design is minimized. For any point $\boldsymbol{u} \in \chi$, the distance to the closest design point is given by $\min_i d(\boldsymbol{u}, \boldsymbol{x}_i)$ where $d(\boldsymbol{u}, \boldsymbol{x}_i)$ represents some distance measure between $\boldsymbol{u}$ and $\boldsymbol{x}_i$. The maximum distance is therefore $\max_{\boldsymbol{u}} \min_i d(\boldsymbol{u}, \boldsymbol{x}_i)$ and the minimax objective can be written as

$$\min_{\zeta} \max_{\boldsymbol{u} \in \chi} \min_{i} d(\boldsymbol{u}, \boldsymbol{x}_i) \tag{2.8}$$

for some design ζ (Joseph, 2016).

Considering that two neighboring points may provide similar information, another space-filling approach is to spread the sampling points as far away from each other as possible. Sphere packing designs, also known as maximin designs, maximize the minimum distance between design points, $\min_{i,j} d(\boldsymbol{x}_i, \boldsymbol{x_j})$. The maximin-distance criterion maximizes

$$\max_{\zeta} \min_{i,j} d(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{2.9}$$

the minimum distance between two design points (Johnson et al., 1990). See Joseph (2016) for a more thorough review of space-filling designs. The present study considers two specific space-filling designs, spatial coverage and Latin hypercube, that incorporate the minimax and maximin space-filling criteria, respectively.

2.2.5 Spatial Coverage Designs

Spatial coverage designs are constructed as the subset of points that minimizes

$$C = \left[\sum_{\boldsymbol{u} \in \chi} \left[\sum_{\boldsymbol{x} \in \zeta} d(\boldsymbol{u}, \boldsymbol{x})^a \right]^{\frac{b}{a}} \right]^{\frac{1}{b}} \tag{2.10}$$

with parameters $a < 0$ and $b > 0$. As $a \to -\infty$ and $b \to \infty$, C converges to a minimax space-filling criterion. Royle and Nychka (1998) discuss these designs further and provide an iterative point swapping algorithm for design construction. We generate these designs in R (R Core Team, 2018) using the package `fields` and the `cover.design` function (Nychka et al., 2019).

A visual of a spatial coverage design for 2 factors and 10 runs is included in Figure 1d. This design clearly contains points distributed throughout the design region (unlike the CCD, D-optimal, and I-optimal designs). Although we consider other space-filling designs as shown in Figure 1(e) through Figure 1(h), spatial coverage designs are unique in that they have no points along the design boundary.

2.2.6 Latin Hypercube Designs

Similar to the sphere packing design, Latin hypercube designs maximize the minimum distance between points but include an additional constraint to enhance one-dimensional projection properties. This constraint reduces redundant points in the presence of unimportant factors (Butler, 2001). To illustrate, Figure 2(a) shows a sphere packing design with two factors and 10 runs. If x_2 is actually superfluous, then the design points will be condensed along the x_1 axis as shown in Figure 2(c). These overlapping points illustrate poor one-dimensional projections. To enhance projection properties, the Latin hypercube design seeks to space factor levels evenly between their lower and upper bounds, creating uniform one-dimensional projections, see Figure 2(d). Increasing the number of levels per factor increases degrees of freedom and allows for fitting more complex models (but this comes with the trade-off of not being as variance efficient as D- or I-optimal designs). This also leads to Latin hypercube designs having fewer points along the boundary of the experimental region than the sphere packing design. Figure 1(e) shows how the Latin hypercube distributes points throughout the design area.

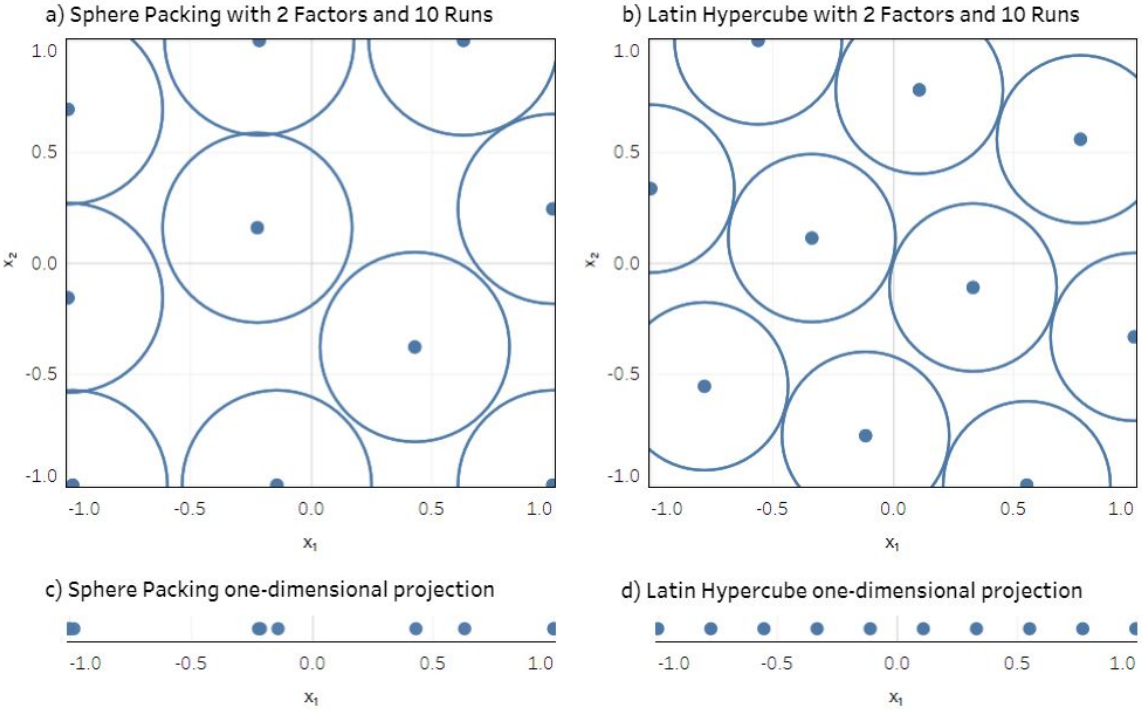


Fig. 2.: Comparison of Space-Filling Sphere Packing and Latin Hypercube designs and their one-dimensional projections.

2.2.7 Bridge Designs

Bridge designs "bridge the gap" between variance optimal and space-filling designs. Although bridge designs are a more recent development (Jones et al., 2015), they have been suggested for further research in calibration studies (Davis et al., 2017). As detailed in Jones et al. (2015), a bridge design seeks to maximize the D-criterion but with a constraint on the minimum distance between points. That is,

$$\max|\boldsymbol{X}'\boldsymbol{X}| \text{ subject to } |x_{ij}-x_{il}| \geq \delta \text{ for } i=1,\dots,h$$
$$\text{and } j,l = 1,\ldots,n,\ j \neq l$$

where δ is the specified minimum univariate interpoint distance. For two-level factors coded as ±1, the tuning parameter ranges from 0, which generates a standard D-optimal design, to $2/(n-1)$, a D-optimal Latin hypercube design. Jones et al. (2015)

suggests the midpoint of this range with a δ of $0.5 \times 2/(n-1)$. The hollow blue circles in Figure 3 illustrate the placement of points using a δ of 0.25 and 0.5 for a design with 21 runs. To compare, points for a D-optimal design are overlaid as orange squares. For $\delta = 0.25$, the tuning parameter is $0.25(2/20) = 0.025$. Figures 3(a) and 3(c) show the points located more closely to D-optimal design locations. As delta increases to 0.5, the tuning parameter increases to $0.5(2/20) = 0.05$ and Figures 3(b) and 3(d) more closely resemble a Latin hypercube design. Figure 1(f) through Figure 1(h) shows the bridge designs for δ of 0.25, 0.5 and 0.75 for 2 factors and 10 runs.

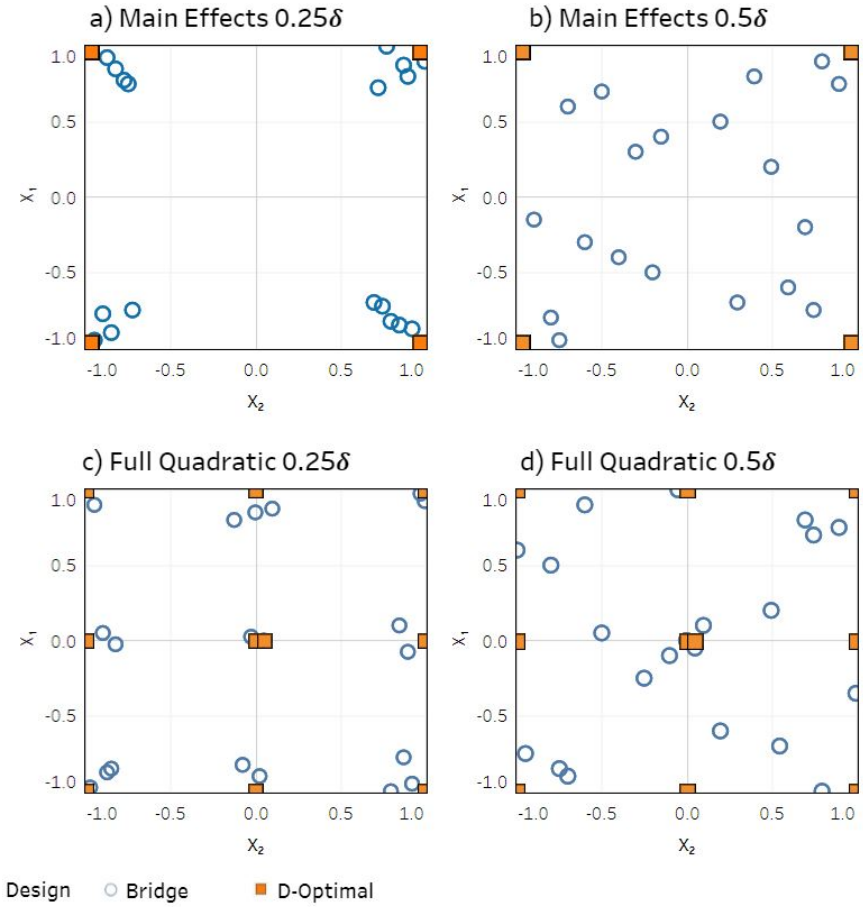


Fig. 3.: Bridge design placement for 2 factors and 21 runs (blue circles); *D*-optimal design is overlaid for reference (orange squares). a) Main effects only with tuning parameter 0.25δ, b) Main effects with tuning parameter 0.5δ, c) Second-order model with tuning parameter 0.25δ, d) Second-order model with tuning parameter 0.5δ

2.3 Simulation Study

This section describes the simulation study to analyze the impact of experimental design on calibration. We first detail the analysis approach utilized followed by an outline of the simulation protocol.

2.3.1 Analysis Approach

In the simulation, we make use of the forward model inverse calibration approach. During data collection, responses are observed for each run using inputs prescribed by the experimental design. After data is collected, a regression model is fit for each response. For the simulation, we assume a full second-order regression model. That is, for each response, the model takes the general form

$$y_i = \beta_{i0} + \sum_{j=1}^{h} \beta_{ij} x_{ij} + \sum \sum_{j<l=2}^{h} \beta_{ijl} x_{ij} x_{il} + \sum_{j=1}^{h} \beta_{ijj} x_{ij}^2 + \epsilon_i \tag{2.11}$$

with $\epsilon_i \sim Normal(0, \sigma^2)$. The number of active terms are varied to mimic the model selection process and we compare design performance for both full and reduced models.

After the forward model is developed, test data is generated to determine how well each design type infers input values from an observed response. That is, we randomly select a number of points from the experimental region to represent the true inputs for the test dataset, $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_h^*)$. For each of these new input points, we simulate true responses $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_k^*)$ to represent new observations after the experiment has taken place. These responses are a function of $\boldsymbol{x}^*$ and true $\boldsymbol{\beta} = (\beta_{i0}, \ldots, \beta_{ijj})$ from the forward model. For each $\boldsymbol{y}^*$, we estimate input values via constrained minimization to determine the set of values $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_h)$ that minimizes the Euclidean distance between $\boldsymbol{y}^*$ and the predicted response $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k)$ where

$$\hat{y}_i = \hat{\beta}_{i0} + \sum_{j=1}^{h} \hat{\beta}_{ij} x_{ij} + \sum \sum_{j<l=2}^{h} \hat{\beta}_{ijl} x_{ij} x_{il} + \sum_{j=1}^{h} \hat{\beta}_{ijj} x_{ij}^2 \tag{2.12}$$

with $\hat{\boldsymbol{\beta}}$ from the fitted forward model. More specifically, we select $\hat{\boldsymbol{x}}$ by minimizing

$$\sqrt{\sum_{i=1}^{k}(y_i^* - \hat{y}_i)^2}. \tag{2.13}$$

The minimization is performed via `nlminb`, an R function that implements a quasi-Newton optimization method that allows us to constrain the estimated input values, $\hat{\boldsymbol{x}}$, to be within the design region (see Gay (1990) for optimizer details). As the optimization method is sensitive to starting values, we utilize the `multiStart-optim` function from R's `mcGlobaloptim` package (Moudiki, 2013). This function allows one to specify a number of optimization trials with randomly selected starting values; we use five random starts in our simulation study.

Finally, design performance is evaluated by comparing the mean/median Euclidean distance (across all test points) between actual input $\boldsymbol{x}^*$ and estimated input $\hat{\boldsymbol{x}}$,

$$\sqrt{\sum_{i=1}^{h}(x_i^* - \hat{x}_i)^2}. \tag{2.14}$$

2.3.2 Simulation Protocol

The simulation study varies i) design type, ii) number of input factors, iii) number of responses, iv) size of error variance, v) correlation among response variables, vi) percent model terms truly active, and vii) a full verses reduced model, to investigate how the various experimental designs perform across different conditions. For a specified number of factors, all designs are generated with the same number of runs as required by a CCD. Input factors are constrained between -1 and 1. All designs, with the exception of the spatial coverage designs, are generated using JMP statistical software (JMP® Pro 14, SAS Institute Inc., Cary, NC). The JMP scripting language code provided by Jones et al. (2015) is utilized to generate a bridge design for a full quadratic model and $\delta = 0.25, 0.5$ and 0.75. These three designs are labeled as Bridge25, Bridge50, and Bridge75, respectively. Spatial coverage designs are generated using the `cover.design` function in R's `fields` package, as previously mentioned. Table 1 summarizes the simulation variables and their levels.

Table 1.: Calibration Summary of Simulation Variables

Simulation Variable (Label)	Levels
Design Type (Design)	CCD, *D*-optimal, *I*-optimal, Latin Hypercube, Spatial Coverage, Bridge
Number of Factors (Factor)	2, 3, 4, 5
Number of Responses (Response)	2, 3, 4, 5, 6, 7
Error Standard Deviation (Standard Deviation)	1, 2, 3
Response Correlation (Rho)	0, ±0.1, ±0.25, ±0.5, ±0.75
Percent Terms Active (Active Percent)	10%, 25%, 50%, 100%
Full or Reduced Model (Reduced)	Full Model, Reduced Model

For each unique combination of attributes, 1,000 iterations of the simulation are run, and the overall mean/median Euclidean distance between actual and predicted inputs is compared to evaluate design performance.

Each iteration begins by generating response values from input values specified by the design matrix. Model terms are randomly selected to be active depending on the specified percent active. Strong effect heredity is assumed when selecting interaction and quadratic terms. Effect heredity assumes interactions and quadratic terms are present only if parent main effects are active (Ockuly et al., 2017). Under strong heredity, interaction terms are allowed in the model only if both main effects are active, and a quadratic term is included only if the associated main effect is active. The true model coefficients for the active terms are randomly assigned from the set $\{\pm 0.5, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \ \pm 7, \pm 8, \pm 9, \pm 10\}$. The intercept, β_0, is fixed at 10. We use the `mvrnorm` function in R's `MASS` package (Venables and Ripley, 2002) to generate training data responses from $\boldsymbol{y} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$

and

$$\mu_i = 10 + \sum_{j=1}^{h} \beta_{ij} x_{ij} + \sum \sum_{j<l=2}^{h} \beta_{ijl} x_{ij} x_{il} + \sum_{j=1}^{h} \beta_{ijj} x_i^2 \tag{2.15}$$

for $i = 1, \ldots, k$ responses. $\mathbf{\Sigma}$ is generated by first constructing a symmetric correlation matrix with off-diagonal values selected from $\rho \in \{0, \pm 0.1, \pm 0.25, \pm 0.5, \pm 0.75\}$. This matrix is then transformed into a positive semidefinite covariance matrix using the eigenvalue decomposition method (Rousseeuw and Molenberghs, 1993) and the `cor2cov` function in R's `MBESS` package (Kelley, 2020).

Recall that a positive semidefinite matrix must have eigenvalues greater than or equal to zero. As we are randomly selecting $\pm\rho$ values during each iteration of the simulation, there is a chance that the generated variance-covariance matrix will not be positive semidefinite. To prevent this from occurring, the generated variance-covariance matrix is decomposed into its eigenvalues and eigenvectors and all eigenvalues are set to greater than or equal to zero. The matrix is then reassembled using the original eigenvectors and adjusted eigenvalues and scaled so diagonal elements are 1. This updated correlation matrix along with the prescribed error standard deviation is transformed into a covariance matrix, $\mathbf{\Sigma}$.

To illustrate, consider the following correlation matrix for 3 responses and correlation levels of ±0.75,

$$\boldsymbol{S} = \begin{bmatrix} 1.00 & 0.75 & 0.75 \\ 0.75 & 1.00 & -0.75 \\ 0.75 & -0.75 & 1.00 \end{bmatrix}. \tag{2.16}$$

$\boldsymbol{S}$ has eigenvalues of

$$\left\{ 1.75 \quad 1.75 \quad -0.5 \right\},$$

one of which is less than zero. To make $\boldsymbol{S}$ positive semidefinite, the negative values are set to zero

$$\left\{1.75 \quad 1.75 \quad 0\right\}$$

and the correlation matrix is adjusted using the original eigenvectors and the adjusted eigenvalues. The resulting matrix

$$\begin{bmatrix} 1.0 & -0.5 & 0.5 \\ -0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{bmatrix}$$

has eigenvalues greater than or equal to zero. Using `cor2cov` and assuming an error standard deviation of 2, $\mathbf{\Sigma}$ becomes

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

with variance 2^2 along the diagonal. We acknowledge that this adjustment will, at times, change the correlation structure from that assigned by the simulation. However, this procedure will allow us to consider a wide variety of possibilities for variance-covariance structure. After generating the true response values, the forward model is fit to the data. The model parameter estimates are determined and, in the case of the reduced model scenario, set to 0 if their p-value is greater than a standard significance level of 0.05.

Fifty new locations in the design region are selected at random to use as actual input values in the test data set. We assume the underlying relationship connecting inputs to responses is the same in both our training and test data. The responses in the test data are generated the same way in our training set. The test data responses and test data model estimates are then used to find the optimal set of input values that minimize the mean Euclidean distance between $\boldsymbol{y}^*$ and $\hat{\boldsymbol{y}}$ for each point.

The designs are evaluated based on their average mean and median Euclidean distance between actual and predicted input values after 1,000 iterations. Results for each of the designs are compared with those having a lower mean Euclidean distance, our measure of prediction error, favored.

The entire simulation process is outlined in Algorithm 1.

Algorithm 1 Simulation Steps

```
1: Assuming strong effect heredity, randomly assign model terms to be active.
2: Create the underlying relationship between inputs and responses by assigning model coefficients from the set $\{\pm 0.5, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8, \pm 9, \pm 10\}$ to the active terms. Inactive terms are assigned a value of zero. Then, the expected value of a response is given by
   $$\mu_i = 10 + \sum_{j=1}^{h} \beta_{ij} x_{ij} + \sum \sum_{j<l=2}^{h} \beta_{ijl} x_{ij} x_{il} + \sum_{j=1}^{h} \beta_{ijj} x_i^2 \qquad (2.17)$$
   for $i = 1, \dots, k$ responses and $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$.
3: Generate a variance-covariance matrix $\boldsymbol{\Sigma}$.
4: Generate training data responses $\boldsymbol{y} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using mvrnorm in R for each experimental run.
5: For each response, fit the assumed forward model using the inputs from the experiment design and training data responses and obtain the model parameter estimates $(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\hat{\beta}_{i0}, \dots, \hat{\beta}_{ijj}))$.
6: Create test data by randomly selecting 50 new points, $\boldsymbol{x}^* = (x_1^*, \dots, x_h^*)$, in the design region bound by -1 and 1.
7: Generate test data responses $\boldsymbol{y}^* \sim Normal(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ using mvrnorm and
   $$\mu_i^* = 10 + \sum_{j=1}^{h} \beta_{ij} x_{ij}^* + \sum \sum_{j<l=2}^{h} \beta_{ijl} x_{ij}^* x_{il}^* + \sum_{j=1}^{h} \beta_{ijj} x_{ij}^{*2} \qquad (2.18)$$
   for $\boldsymbol{\mu^*} = (\mu_1^*, \mu_2^*, \dots, \mu_k^*)$ and each point $\boldsymbol{x}^*$.
8: Using the $\hat{\boldsymbol{\beta}}_i$'s from the forward model, use nonlinear optimization to find $\hat{\boldsymbol{x}}$ values that minimize
   $$\sqrt{\sum_{i=1}^{k} (y_i^* - (\hat{\beta_{i0}} + \sum_{j=1}^{h} \hat{\beta}_{ij} \hat{x}_{ij} + \sum \sum_{j<l=2}^{h} \hat{\beta}_{ijl} \hat{x}_{ij} \hat{x}_{il} + \sum_{j=1}^{h} \hat{\beta}_{ijj} \hat{x}_{ij}^2))^2} \qquad (2.19)$$
   for each point $\boldsymbol{x}^*$.
9: Record the mean Euclidean distance between $\boldsymbol{x}^*$ and $\hat{\boldsymbol{x}}$ for the 50 new points in the test data. Repeat steps 1,000 times and take the average mean and median of the Euclidean distances between $\boldsymbol{x}^*$ and $\hat{\boldsymbol{x}}$ per scenario.
```

2.4 Simulation Results

Figure 4(a) and 4(b) shows the overall mean and median Euclidean distance, respectively, per design. Across all scenarios, I-optimal designs exhibit the best performance followed closely by the CCD. Box plots of the mean and median Euclidean distance are shown in Figure 4(c) and 4(d) and also indicate the I-optimal and CCD designs as being the best performers. Spatial coverage designs, on the other hand, have the largest mean/median Euclidean distance as well as the largest range of mean/median values.

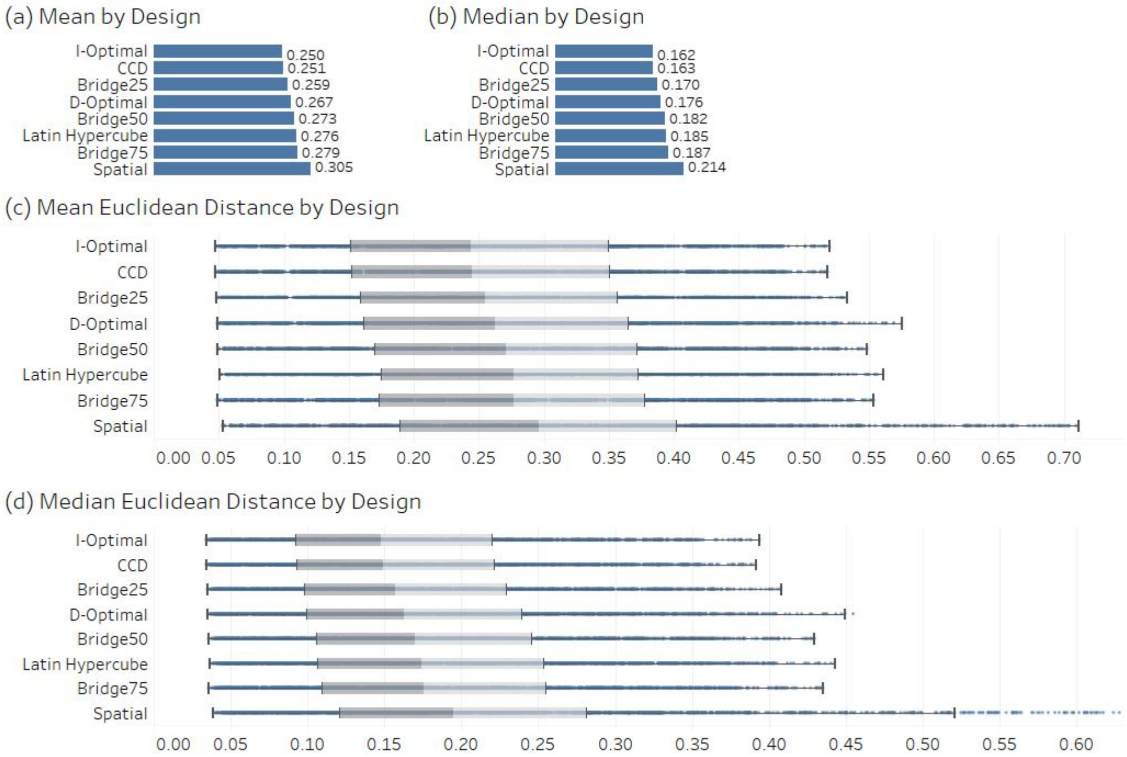


Fig. 4.: Euclidean Distance Metrics by Design

Figure 5 indicates each design's rank according to its average mean Euclidean distance for each specified scenario. With one exception, an I-optimal design was consistently the best performer. Comparing across the attributes varied in the simulation, we note that Bridge25 was among the top 3 performers. In fact, Bridge25 outperformed both D-optimal and Latin hypercube designs (of which Bridge25 is a hybrid of). However as the bridge design tuning parameter, δ, increased to 0.5, it lagged behind D-optimal designs for most of the scenarios. A decrease in performance is clearly seen as the tuning parameter increases and the bridge designs became more akin to a Latin hypercube. Jones et al. (2015) recommend using a small tuning value when there is more certainty that the correct regression model is being used.

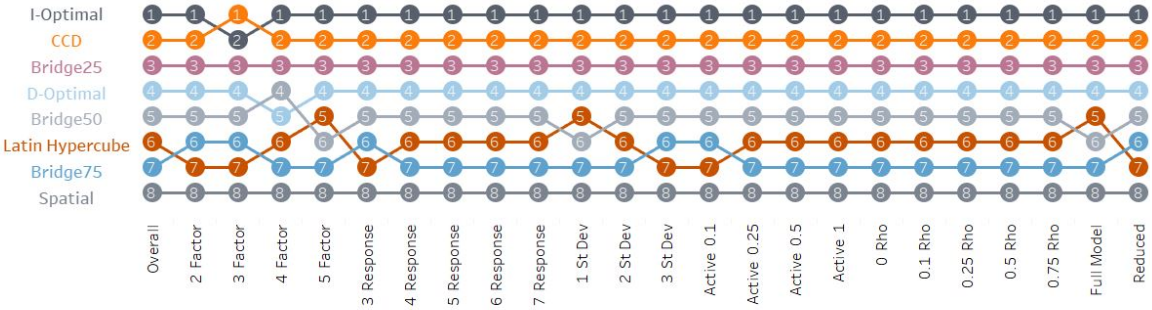


Fig. 5.: Designs Ranked by Mean Euclidean Distance (Lowest to Highest)

To formalize the simulation results, we fit a regression model with the mean Euclidean distance as the dependent variable and simulation scenario attributes as predictors. A model that includes main effects and two-factor interactions with design type (Design) exhibits an adjusted R-square of 0.938; an ANOVA for this model is shown in Table 2. Most notably, the ANOVA results indicate clear differences by experimental design (p-value $<$ 2e-16). All two-factor interactions with Design are important with the exception of Design×Rho. See Figure 6 for interaction plots involving Design.

From Figure 6(a) and 6(c), we see that the mean Euclidean distance increases as the number of input factors and the error standard deviation increase. Thus, it is recommended that an experimenter select a conservative number of input factors when inverse prediction will be performed. Inverse prediction performance also appears to be more favorable when using the full model versus a reduced model (Figure 6(f)). Conversely, prediction error decreases as the number of responses increased (Figure 6(b)). Correlation among the responses, however, does not appear to have a large impact on inverse prediction performance in this study (Figure 6(d)). Finally, prediction error decreases as the percent of active terms increases (Figure 6(e)). All of the interaction plots in Figure 6 clearly show the spatial coverage design lagging behind the other design choices.

Term	Df	F value	P-value
Design	7	744.679	<2e-16
Factor	1	97545.65	<2e-16
Response	1	84660.14	<2e-16
Active Percent	1	6445.327	<2e-16
Standard Deviation	1	90214.61	<2e-16
Rho	1	212.894	<2e-16
Reduced	1	2423.289	<2e-16
Design:Factor	7	165.843	<2e-16
Design:Response	7	14.643	<2e-16
Design:Active	7	5.997	5.34e-07
Design:Standard Deviation	7	103.595	<2e-16
Design:Rho	7	0.325	0.943
Design:Reduced	7	185.057	<2e-16
Residuals	19144		

Table 2.: Calibration ANOVA Results for Main Effect and Design Interactions, degrees of freedom (Df), F value and p-value

A post-hoc Tukey test indicates that *I*-optimal designs are not significantly better than CCD and Bridge25 designs, but are significantly better than the remaining design choices. Considering the worst performer, spatial coverage designs are significantly worse than all other designs. Full results of the Tukey test are included in the Appendix.

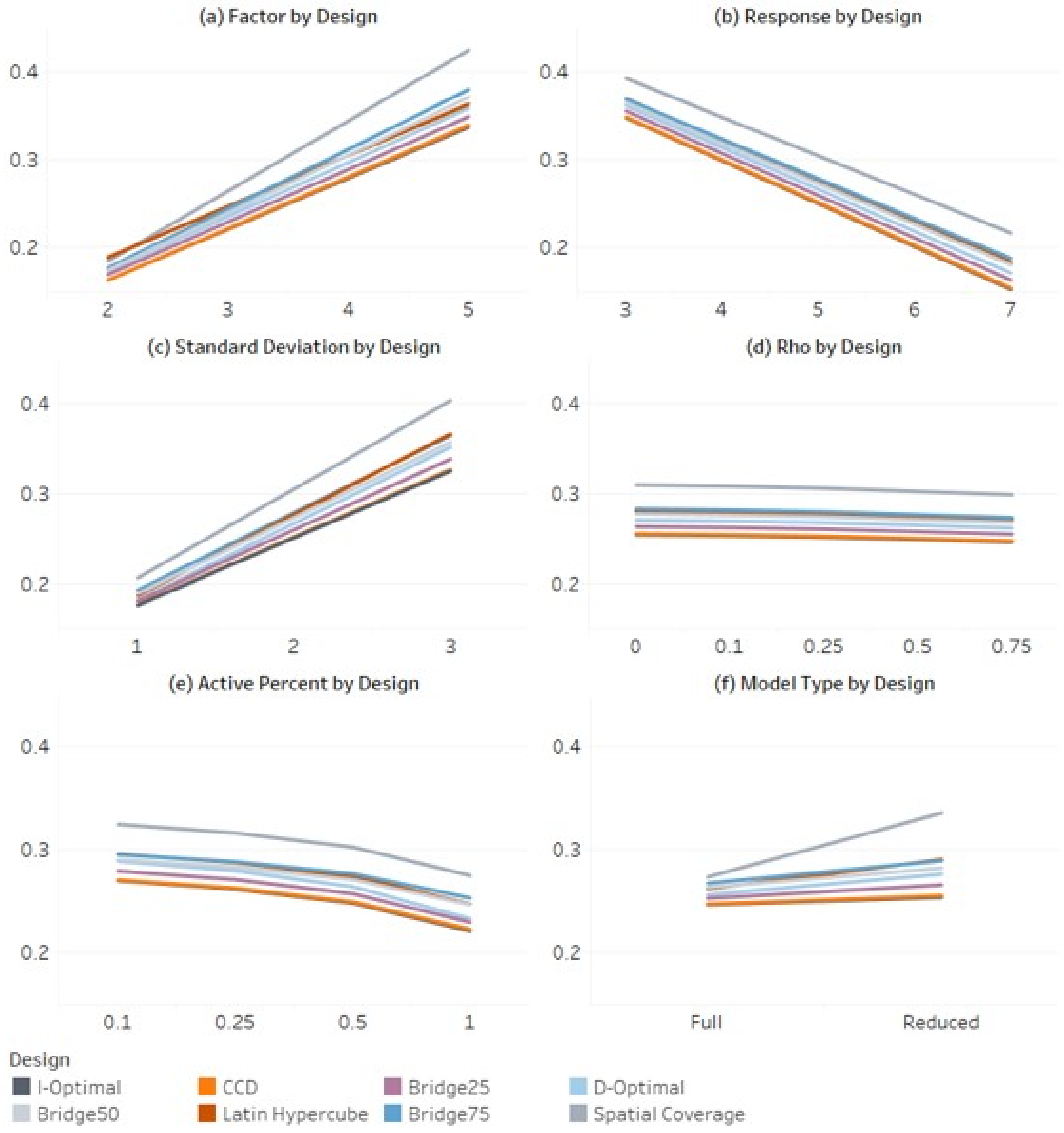


Fig. 6.: Predicted Mean Euclidean Distance by Design (Mean Euclidean distance on y-axis and attribute on x-axis, y-axis starts at 0.15)

2.5 Summary and Conclusion

This chapter compares experimental designs with regards to inverse prediction performance. Our simulation study varied design type, number of input factors, number of response variables, error standard deviation, correlations among responses, number of active terms, and whether or not predictor subset selection is helpful. The designs were evaluated based on the mean/median Euclidean distance between actual and predicted input values. It is our hope that knowledge regarding design performance across various scenarios/conditions may help influence design choice and setup of future calibration studies.

The designs investigated for the study have appeared in the chemometrics literature and include both traditional and non-traditional choices. Results from the simulation indicate a relationship between design choice and inverse prediction performance. Overall, an analysis of the simulation suggests the use of I-optimal, CCD, and bridge designs with a small tuning parameter (Bridge25). These three design types had the lowest Euclidean distance between actual and predicted input overall as well as across all of the simulation scenarios. Additional recommendations from the study include limiting the number of input factors and making use of the full hypothesized model (i.e. no model reduction). Correlation among response variables did not appear to have an overall negative impact on inverse prediction (but further research is recommended).

As noted by Anderson-Cook et al. (2015), research focused on formally evaluating the extent to which design optimality criteria translate into favorable inverse prediction performance is needed. Despite this lack of research, we conjecture that precise prediction via forward models is indeed a good criterion for inverse prediction and, thus, led to the favorable performance of I-optimal designs in the simulation study. Recall that prediction, $\hat{\boldsymbol{y}}$, at an input, $\boldsymbol{x}$, is a key component of the constrained optimization (Equation (2.13)).

CCDs are well known for both their good estimation and prediction properties when fitting a second-order model; this likely explains their inclusion in the top three performers in the simulation study. Finally, the hybrid Bridge25 design is most similar to a D-optimal design vs. the other bridge designs considered. Thus, the performance of the Bridge25 design is likely attributed to its ability to efficiently estimate the second-order model while also allowing for some coverage of points across the design region.

Initially, we were surprised by the poorer performance of the space-filling designs in the simulation study. However, Johnson et al. (2010) compare space-filling designs with optimal designs based on their prediction variance properties and conclude that space-filling designs do not perform as well as optimal designs with regards to this metric when fitting polynomial regression models.

The goal of this chapter has been to contribute to the understanding of the role of design of experiments in inverse problems via an extensive (although admittedly idealistic) simulation study. To name a couple limitations, we did not consider errors in the input factors (also known as random calibration) and we also assumed that a second-order regression model was the true model. As a consequence, model selection was not performed with the exception of model reduction via predictor subset selection. Both considerations are certainly worthy of future research. Another avenue for further research includes comparing design types for inverse prediction when the true model is a higher-order polynomials or more complex nonlinear models but the user locally approximates the response surface by lower-order polynomial models. Investigation of other calibration methods (e.g. direct methods or Bayesian approaches) would also be interesting for future work.

CHAPTER 3

OPTIMAL SEQUENTIAL DESIGN TECHNIQUES

3.1 Background and Literature Review

The process of learning and adapting over time is the foundation of sequential design techniques. Sequential designs, also known as follow-up designs, is a method that involves running a series of experiments with each one building off the previous iteration (see Edwards et al. (2014), and Nelson et al. (2000)). Sequential designs often begin with a screening design aiming to discard negligible factors. Once the results are analyzed, a follow-up design is conducted to hone in on the active factors. With finite resources, experimenters must decide whether to allocate runs for a sequence of experiments or to use all runs in a design without first screening. Yet the performance of sequential designs compared to a single nonsequential design without screening is unknown. To narrow the scope, we focus on algorithmically generated D-optimal designs and a design space fixed to a constrained design region. We aim to provide practitioners guidance on the most efficient way to apportion a given set of runs.

Jensen (2018) cautions the focus optimal designs place on extracting the maximum information at a point in time has shadowed the importance of sequential experimentation and that the best way to apportion a set number of runs is an open problem of optimal design. This chapter investigates the best way to allocate a fixed number of runs within a fixed design space. For example, given that the experimenter has a total of 20 possible runs, is it best to run a single 20 run design in all the factors of interest all at once or use a sequential approach? The sequential approach allows the experimenter to use findings from the initial screening runs to focus on important factors in the next iteration. However the smaller run size of screening designs comes with a drawback of aliasing of factorial effects and potential ambiguity in analysis.

While there is robust literature on screening and sequential designs, there are only a limited number of studies on optimal sequential design techniques (see Silvestrini (2015), Edwards et al. (2014), Nelson et al. (2000)). Edwards et al. (2014) compare four design follow-up strategies: foldover (see Montgomery, Douglas C., Runger (1996)), semifoldover (see John (2000), Mee and Peralta (2000)), D-optimal (see Goos and Jones (2011)), and Bayesian MD-optimal (see Meyer, R. Daniel et al. (1996)) for several initial design choices using a metal-cutting case study. A single follow-up strategy did not outperform all others across various scenarios. Foldover and semifoldover augmentation require a fixed number of additional runs while D-optimal and MD-optimal augmentation are more flexible given that a design size that can be customized based on available resources. D-optimal augmentation locates the follow up runs that maximize the D-criterion. Recall that D-optimal designs maximize the determinant of the information matrix. The model matrix for the initial screening design, $\boldsymbol{X}_1$, has columns corresponding to each parameter in the user specified model and a row for each run. With $\boldsymbol{X}_2$ as the model matrix for the follow up runs, the initial design is augmented by choosing additional runs with treatment combinations that maximize

$$|\boldsymbol{X}_1'\boldsymbol{X}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2| \tag{3.1}$$

across possible $\boldsymbol{X}_2$ matrices.

Nelson et al. (2000) compare three design augmentation strategies in the context of industrial experiments. The strategies include a foldover of a 12 run Plackett-Burman design (Plackett and Burman, 1946), and a two-level fractional factorial design (2^{8-4} 16 run screening design) augmented by either a three quarter foldover of the 2^{8-4} design or using D-optimal augmentation with the optimality criterion shown in equation 3.1. The authors caution that in the chemical or processing industries two-factor interactions (2FI) are often as important as the main effects, and a partial aliasing of factors can make it difficult to determine the correct model when interactions are present. Results from a Monte Carlo study suggest that the three-quarters-based augmentation strategy has an advantage over the Plackett-Burman approach with both methods outperforming D-optimal augmentation. The three quarters augmentation approach uses semi-folding to de-alias specific alias chains based on terms identified as significant during the screening design.

Silvestrini (2015) investigates the impact of initial design size, number of runs in each design augmentation stage, total run size, additional center points, and initial model specification on sequential D-optimal designs using a simulation study. Designs are evaluated in terms of D-efficiency, a measure comparing the design's determinate of the information matrix to that of an orthogonal design. That is

$$\left(\frac{|\boldsymbol{X}'\boldsymbol{X}|}{n^p}\right)^{1/p} \tag{3.2}$$

with $\boldsymbol{X}$ is the design's model matrix, and p is the number of model parameters.

Simulation results show increasing total sample size improves D-efficiency in addition to starting with a simpler model form in the initial design when the underlying model is unknown as opposed to a more complex model. These findings rely on the assumption that the experimenter gets closer to identifying the correct model at each subsequent stage. We expand on Silvestrini (2015) in our simulation study by measuring the impact of initial design size and total sample size in terms of power, the proportion of active terms that are correctly identified, and false discovery rates (FDR), the proportion of terms incorrectly identified as active. The rest of this chapter proceeds as follows. Section 2 provides details on the experimental designs and augmentation methods used in our study. Section 3 describes LASSO and forward selection model selection methods. Section 4 covers our simulation study protocol and reviews a simulation example. In Section 5 we examine our simulation results and we conclude with a discussion in Section 6.

3.2 Experimental Designs

3.2.1 *D*-optimal Designs

D-optimal designs seek to minimize the variance of the model parameter estimates (Myers et al., 2016). This criterion has proven to be particularly useful in screening scenarios where the objective is to identify few important factors out of the many factors of interest. As the determinant can be used as a measure of overall variance, a *D*-optimal design locates a set of design points that minimizes the determinant of the inverse information matrix (i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ where $\boldsymbol{X}$ is the $n \times p$ model matrix for n runs). It is computationally more tractable, however, to maximize the determinant of the information matrix. Thus, a *D*-optimal design is such that

$$|\boldsymbol{X}'\boldsymbol{X}| \tag{3.3}$$

is maximized.

3.2.2 Bayesian *D*-optimal augmentation

To add additional runs to our screening designs, we leverage the Bayesian *D*-optimal augmentation method provided in Gutman et al. (2014) and Zhang et al. (2019) for main effects, and we broaden it to include 2FI and quadratic terms. The process is as follows. Let $\boldsymbol{X}_1$ be the $n_1 \times p$ model matrix of an initial design with n_1

screening runs, and $\boldsymbol{y_1}$ as the corresponding vector of responses. We wish to select the $n_2 \times p$ model matrix for n_2 follow up runs, $\boldsymbol{X_2}$, such that the optimality criterion of the final $n = (n_1 + n_2) \times p$ model matrix $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X_2} \end{pmatrix}$ is maximized. Recall that D-optimal designs maximize $|\boldsymbol{X'X}|$; for an augmented design this is

$$|\boldsymbol{X'X}| = | \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}' \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} | = |(\boldsymbol{X}_1'\boldsymbol{X}_2') \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} | = |\boldsymbol{X}_1'\boldsymbol{X}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2|. \tag{3.4}$$

Bayesian D-optimal augmentation incorporates findings from an initial screening design through the prior information matrix $\boldsymbol{R}$ to select follow up runs. Consider the linear model $\boldsymbol{X\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{X}$ as the $n \times p$ model matrix, $\boldsymbol{\beta}$ as the $p \times 1$ vector of coefficients, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n)$ as the $n \times 1$ error vector. The prior distribution of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{R}^{-1})$ where $\boldsymbol{\beta}_0$ is the prior mean for $\boldsymbol{\beta}$, and the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\beta}$ is $\boldsymbol{y}|(\boldsymbol{\beta}, \sigma^2) \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$. Then the posterior distribution for $\boldsymbol{\beta}$ given $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{pmatrix}$, where $\boldsymbol{y_2}$ is the vector of responses for the follow up runs is

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N[\boldsymbol{b}, \sigma^2(\boldsymbol{X_1'X_1} + \boldsymbol{X_2'X_2} + \boldsymbol{R})^{-1}] \tag{3.5}$$

with $\boldsymbol{b} = (\boldsymbol{X_1'X_1} + \boldsymbol{X_2'X_2} + \boldsymbol{R})^{-1}(\boldsymbol{X_1'y_1} + \boldsymbol{X_2'y_2} + \boldsymbol{R\beta}_0)$.

After analyzing the initial experiment, terms are classified into two groups. Terms appearing active are labeled *primary*, and terms not appearing active are *potential*. Term coefficients are given a prior variance based on this grouping. Coefficients of primary terms are given a diffuse prior variance tending to infinity. Potential terms are given a prior mean of zero and a finite variance $\sigma^2\tau^2$ where τ is a user specified constant. Multiplying σ^2 by larger values represents stronger confidence that the factor is active. This study follows Gutman et al. (2014) and sets $\tau = 5$ following a sensitivity analysis analyzing various levels of τ. Let p_1 be the number of primary terms, and p_2 the number of potential terms. The intercept term is always labeled primary and we have $p_1+p_2 = p$ terms. The matrix $\boldsymbol{X}$ is reordered with the intercept and other primary terms first, followed by potential terms. The prior information matrix $\boldsymbol{R}$ is

$$\boldsymbol{R} = \begin{pmatrix} \mathbf{0}_{p_1 \times p_1} & \mathbf{0}_{p_1 \times p_2} \\ \mathbf{0}_{p_2 \times p_1} & \boldsymbol{I}_{p_2}/\tau^2 \end{pmatrix} \tag{3.6}$$

where $\mathbf{0}_{s \times t}$ is a matrix of zeros with s rows and t columns. Bayesian D-optimality incorporates the prior information matrix $\boldsymbol{R}$ and maximizes

$$|\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{R}|. \tag{3.7}$$

We use the Coordinate-Exchange Algorithm (Meyer and Nachtsheim, 1995) to find follow up runs, $\boldsymbol{X}_2$, that maximize the Bayesian D-optimal augmentation criteria $|\boldsymbol{X}_\mathbf{1}'\boldsymbol{X}_\mathbf{1} + \boldsymbol{X}_\mathbf{2}'\boldsymbol{X}_\mathbf{2} + \boldsymbol{R}|$. The algorithm begins each iteration by populating an initial $\boldsymbol{X}_2$ matrix with values randomly selected from $[-1, 0, 1]$. Next, the algorithm iterates through each row i and column j and exchanges the existing $\boldsymbol{X}_{2ij}$ value with the $\{0, -1, 1\}$ coordinate that maximizes $|\boldsymbol{X}_\mathbf{1}'\boldsymbol{X}_\mathbf{1} + \boldsymbol{X}_\mathbf{2}'\boldsymbol{X}_\mathbf{2} + \boldsymbol{R}|$. If the objective criteria improves after a coordinate exchange then $\boldsymbol{X}_2$ is updated. The algorithm proceeds to the next element in the row or column such that every coordinate in each row and column is exchanged for $\{0, -1, 1\}$. This process is repeated using 10 random starts and the resulting $\boldsymbol{X}_2$ is the matrix that maximizes the objective function.

3.3 Model Selection Methods

We utilize model selection in the analysis of screening design results as well as in the final analysis incorporating data from all experiment runs. We compare two methods, forward selection and a penalized regression method LASSO (least absolute shrinkage and selection operator) to see if our findings differ based on model selection method. Details on forward selection and LASSO are provided in the following subsections.

3.3.1 Forward Selection

Forward selection is a model selection procedure that sequentially adds terms to a regression model on the basis of partial F-tests until none of the remaining candidate terms are significant or all variables have been added to the model (Myers, 1990). The process starts with an initial model containing just the intercept term and the procedure chooses the next variable j to include that has the largest significant partial F-test out of all the candidate terms. A statistically significant partial F-test has a p-value less than or equal to some prespecified α threshold. At each stage, let the model matrix with the candidate term j be $\boldsymbol{X}_1$ and $\boldsymbol{X}_{1,-j}$ is the model selected by the previous step. The Type III sums of squares of β_j for the fitted model (SS_{β_j}) is

$$SS_{\beta_j} = \boldsymbol{y}'(\boldsymbol{H}_1 - \boldsymbol{H}_{1,-j})\boldsymbol{y} \tag{3.8}$$

with $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{H}_{1,-j} = \boldsymbol{X}_{1,-j}(\boldsymbol{X}_{1,-j}'\boldsymbol{X}_{1,-j})^{-1}\boldsymbol{X}_{1,-j}'$. The sums of square error is

$$SSE = \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{H}_1\right)\boldsymbol{y} \tag{3.9}$$

where $\boldsymbol{I}$ is an identity matrix of size n. The partial F-statistic is

$$\frac{SS_{\beta_j}}{SSE/n-p} \tag{3.10}$$

where p is the number of model parameters in $\boldsymbol{X}_1$.

Other criteria other than partial F-tests may be used with forward selection such as the Akaike information criterion (AIC) or AICc, a modified version of AIC for small sample sizes (see Smucker et al. (2020) for details). Our simulation study uses forward selection with AICc. This stepwise procedure adds a candidate term to the model if the addition of the term results in the model with minimum AICc out of all candidate terms in each step. With

$$AIC = -2log(\hat{L}) + 2p \tag{3.11}$$

where $p < n-1$, and $\hat{L}$ is the maximum likelihood (Akaike, 1974). We have $AICc = AIC + \frac{2p(p+1)}{n-p-1}$. If the candidate term with the minimum AICc has an AICc greater than the existing model, the process terminates.

3.3.2 LASSO

The least absolute shrinkage and selection operator (LASSO) is a model selection method introduced by Tibshirani (1996). It is a regularization technique often used to analyze small screening designs (Smucker et al., 2020). LASSO shrinks coefficient estimates setting some to zero by minimizing

$$(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \sum_{j=0}^{h}(\lambda|\beta_j|) \tag{3.12}$$

where λ is a tuning parameter (Zhang et al., 2019). The tuning parameter λ is often selected using cross-validation, though Smucker et al. (2020) recommend using AICc for small structured designs. We select the value for λ that minimizes $AICc = AIC + \frac{2p(p+1)}{n-p-1}$ where

$$AIC = n \log \frac{(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L^{\lambda})'(Y - \boldsymbol{X}\boldsymbol{\beta}_L^{\lambda})}{n} + 2p \tag{3.13}$$

where $p < n-1$, and $\boldsymbol{\beta}_L^{\lambda}$ are the LASSO estimates for value λ.

3.4 Simulation Study

This section describes the simulation study used to analyze the best use of a given set of total runs. We first detail the analysis approach utilized followed by an outline of the simulation protocol.

3.4.1 Analysis Approach

All designs included in our simulation study are Bayesian *D*-optimal designs. We create all our initial designs in JMP®. For our sequential designs, we consider a 12 run initial design for 20 total runs, 20 run initial design for 32 total runs, a 32 run initial design for 48 runs, and a 40 run initial design for 64 total runs. All designs have input factors constrained between -1 and 1. We specify main effects and quadratic terms as primary, and 2FI as potential for nonsequential and three-level screening designs. Two-level screening designs have main effects specified as primary, 2FI as potential, and include 3 center runs to test for the presence of curvature. To illustrate, the designs for 6 factors and 32 total runs are provided in Table 3, Table 3(a) shows the two-level screening design with 20 runs including 3 center runs, Table 3(b) contains the three-level screening design with 20 runs, and Table 3(c) contains the nonsequential three-level design with the total 32 runs. Grayscale correlation maps illustrating the aliasing of terms in the 6 factor 20 run screening designs and the 32 run nonsequential designs are shown in Figure 7. All quadratic terms are aliased with each other in the two-level screening design with 3 center runs (Figure 7(a)), the three-level screening design has nine pairs of 2FIs having a correlation above 0.76 (Figure 7(b)), and the larger nonsequential design has overall lower correlations among terms (Figure 7(c)).

For the sequential approach, once we have the results from the initial screening

Table 3.: Comparison of Bayesian D-Optimal designs for 6 factors and 32 total runs

(a) Two-level screening with 3 center runs

x_1	x_2	x_3	x_4	x_5	x_6
1	1	1	-1	-1	-1
-1	-1	-1	1	1	-1
1	1	-1	1	1	-1
-1	-1	1	-1	-1	-1
-1	-1	-1	-1	-1	1
-1	1	-1	1	1	1
1	1	-1	-1	-1	1
1	-1	1	1	1	1
1	1	1	1	-1	1
-1	1	1	-1	1	1
-1	-1	1	1	-1	1
1	-1	-1	1	-1	-1
-1	1	1	1	1	-1
1	-1	1	-1	1	-1
1	-1	-1	-1	1	1
-1	1	-1	-1	1	-1
-1	1	-1	1	-1	-1
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

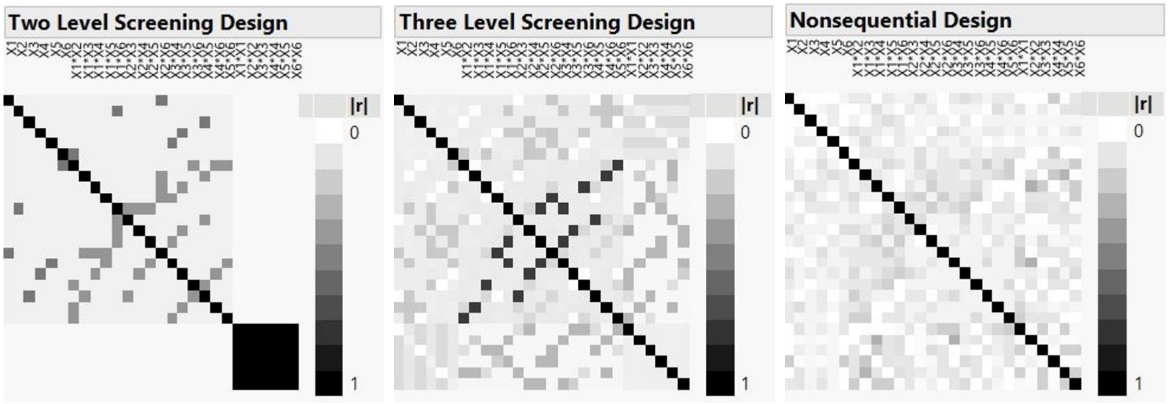
(b) Three-level screening design

x_1	x_2	x_3	x_4	x_5	x_6
1	1	-1	1	0	1
0	-1	1	1	0	0
1	0	0	1	1	0
1	1	1	1	-1	-1
0	1	0	-1	-1	1
-1	1	1	-1	1	-1
-1	-1	0	1	-1	1
1	-1	1	-1	1	1
0	-1	-1	0	1	1
-1	-1	-1	1	1	-1
-1	0	-1	-1	0	1
1	0	1	0	-1	1
1	1	-1	-1	1	-1
0	0	-1	1	-1	-1
-1	1	1	1	1	1
1	-1	-1	-1	-1	0
-1	-1	1	-1	-1	-1
1	-1	0	0	0	-1
-1	1	-1	0	-1	0
1	-1	1	1	1	-1

(c) Nonsequential Three-level design

x_1	x_2	x_3	x_4	x_5	x_6
-1	-1	-1	1	-1	-1
1	-1	0	0	-1	-1
-1	-1	0	-1	-1	1
1	1	-1	1	-1	-1
-1	0	0	-1	1	-1
-1	1	1	-1	-1	0
1	1	1	-1	-1	-1
-1	1	1	1	1	-1
-1	-1	1	1	1	1
1	0	-1	-1	-1	0
-1	-1	1	-1	-1	-1
-1	1	-1	-1	1	1
-1	1	1	0	0	-1
0	1	0	1	1	0
0	1	1	-1	1	1
1	-1	-1	1	1	-1
1	-1	1	-1	-1	1
1	1	0	-1	0	1
0	0	1	1	-1	-1
-1	0	-1	1	0	1
1	-1	-1	-1	1	1
1	-1	1	-1	1	-1
1	1	-1	-1	1	-1
0	1	-1	0	-1	1
1	1	1	1	1	1
-1	1	-1	-1	-1	-1
0	-1	-1	-1	0	-1
-1	1	1	1	-1	1
-1	-1	-1	0	1	0
1	-1	1	1	0	0
1	0	1	0	1	1
1	-1	-1	1	-1	1

runs, we use Bayesian D-optimal augmentation to generate follow up runs for both the two and three-level screening designs. We use model selection, either forward selection or LASSO, to determine which terms are classified as primary or potential for the follow up runs. Both methods select a subset of parameters using the minimum AICc criteria. Main effect and interaction terms in the model chosen are classified as primary while the remaining terms not in the selected model are labeled potential. We include the following additional model selection steps for the two-level screening design to help capture quadratic effects. Forward selection and LASSO



(a) (b) (c)

Fig. 7.: Correlation map of simulation design types for 6 factors

model selection methods include quadratic terms as candidate terms and if at least one quadratic term is selected by model selection, we classify all quadratic terms as primary, and we proceed with the Bayesian D-optimal augmentation. If no quadratic terms are selected by model selection, there is no evidence of curvature and quadratics terms are no longer included in the $\boldsymbol{X}$ model matrix during subsequent D-optimal augmentation and analysis stages.

We then use the process outlined in subsection 3.2.2 to select follow up runs for the remaining trials using Bayesian D-optimal augmentation for both two and three-level screening designs. If no quadratic effects are detected in the two-level screening design, the coordinate exchange algorithm exchanges coordinates for two levels $\{-1, 1\}$ instead of the three-level selection $\{-1, 0, 1\}$. Model selection is conducted on the total number of runs and the terms included in the model with the minimum AICc are identified as active effects.

For the nonsequential approach, we carry out model selection on all the design runs, and the subset of parameters selected by the model selection method using AICc are considered the terms identified as active.

3.4.2 Simulation Protocol

We consider the following:

1. Number of factors: 6, 7, 8
2. Total runs: 20 runs (with a 12 run initial experiment), 32 runs (with a 20 run initial experiment), 48 runs (with a 32 run initial experiment), 64 runs (with a 40 run initial experiment)
3. Active Main Effects: 2, 3, 4, 5
4. Active 2FI: 0, 1, 3, 5, 7
5. Active Quadratics: 0, 1, 3, 5
6. Initial Bayesian D-optimal Design:
   (a) Two-level screening design (main effects are specified as primary, 2FI as potential, with 3 center point runs)
   (b) Three-level screening (main effects and quadratics are specified as primary, 2FI as potential)
   (c) Three-level nonsequential (main effects and quadratics are specified as primary, 2FI as potential)
7. Number of stages:
   (a) 1 stage (nonsequential design approach)
   (b) 2 stages (sequential approach including a smaller screening design + number of follow-up runs to get to total run size)
8. Analysis method: Forward selection, LASSO

Table 4 summarizes the simulation variables and their levels. There are a total of 3,240 distinct simulation scenarios. Each iteration begins by generating response values from input values specified by the design matrix. Model terms are randomly selected to be active depending on the specified number of active terms. Strong effect heredity is assumed when selecting 2FI and quadratic terms. Effect heredity assumes interactions are present only if parent main effects are active (Ockuly et al., 2017). Under strong heredity, quadratic terms are allowed in the model only if their parent main effect is active, and interaction terms are allowed in the model only if both their parent main effects are active. This limits the number of possible active 2FI and quadratic terms in our scenarios. For example, with 2 active main effect terms, a maximum of one 2FI and a maximum of two quadratic terms are possible. The true model coefficients for the active terms are randomly assigned from the set $\{\pm 0.5, \pm 1, \pm 2, \pm 3\}$. The intercept, β_0, is fixed at 10. We use the following to generate responses,

$$y = 10 + \sum_{j=1}^{h} \beta_j x_j + \sum \sum_{i<j=2}^{h} \beta_{ij} x_i x_j + \sum_{j=1}^{h} \beta_{jj} x_j^2 + \epsilon \tag{3.14}$$

with $\epsilon \sim Normal(0, \sigma^2)$, $\sigma^2 = 1$, and h is the number of factors.

Table 4.: Optimal sequential design summary of simulation variables

Simulation Variable	Levels
Design Type	Two Level Screening, Three Level Screening, Nonsequential Design
Number of Factors	6, 7, 8
Active Main Effects	2, 3, 4, 5
Active Two-Factor Interactions*	0, 1, 3, 5, 7
Active Quadratic Terms*	0, 1, 3, 5
Total Run Size	20, 32, 48, 64
Analysis Method	Forward Selection, LASSO

* Strong effect heredity enforced

For each unique combination of attributes, 1,000 iterations of the simulation are run, and the overall mean power and FDR are compared to evaluate design performance. The entire simulation process is outlined in Algorithm 2.

Algorithm 2 Simulation Steps

```
1: Assuming strong effect heredity for 2FI and quadratic terms, randomly assign model terms to be active based on the number of active main effect terms and possible 2FI and quadratics.
2: Create the underlying relationship between inputs and responses by assigning model coefficients from the set $\{\pm 0.5, \pm 1, \pm 2, \pm 3\}$ to the active terms. Inactive terms are assigned a value of zero. Note some scenarios do not have any active 2FI or quadratic terms as part of our simulation test cases.
3: Generate a response for each run in the initial experimental design using 3.14 and $\sigma^2 = 1$.
4: Two-level screening design: Use forward selection or LASSO on the initial generated responses. Label terms chosen by model selection as primary, else potential. If any quadratic term is in the subset of terms identified by model selection then label all quadratics primary. If no quadratics are selected by model selection then exclude all quadratics from the model matrix during subsequent design augmentation and analysis steps.
5: Three-level screening design: Use forward selection or LASSO on the generated responses. Label terms chosen by model selection as primary, else potential.
6: All screening designs: add remaining runs (total run size - screening design run size) using Bayesian $D$-optimal augmentation and the coordinate exchange algorithm. Generate a response for each new run using 3.14.
7: All designs: Use forward selection or LASSO on total experiment runs. All terms in the selected model are identified as active.
8: Calculate power and FDR.
9: Repeat all previous steps 1,000 times and take the average mean power and FDR per scenario.
```

3.4.3 Simulation Example

This section details the various designs and augmentation strategies considered using an example with 8 specified active terms and coefficients for 32 total runs and 6 factors. Suppose the true model is assigned 3 active main effects $\{x_1, x_3, x_4\}$, 3 active 2FI $\{x_1x_3, x_1x_4, x_3x_4\}$ and 2 active quadratic terms $\{x_1^2, x_3^2\}$. Note that strong heredity is enforced and 2FI and quadratic terms only include the active main effects x_1, x_3, and x_4. There are a total of 8 active terms but the experimenter does not know which of the 6 main effects, 15 2FI and 6 quadratic terms are active, if any. The active terms $\{x_1, x_3, x_4, x_1x_3, x_1x_4, x_3x_4, x_1^2, x_3^2\}$ are assigned coefficients

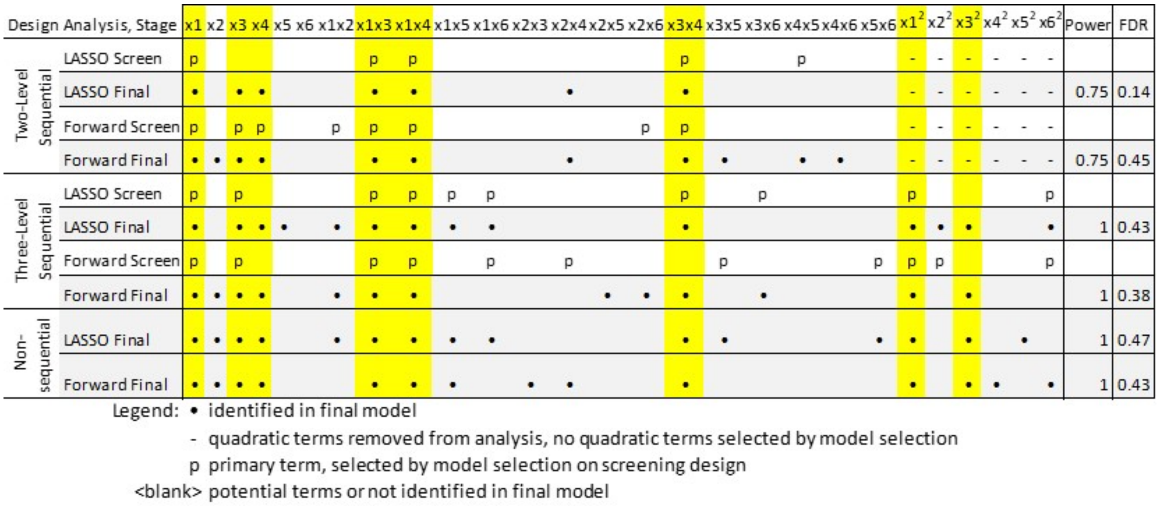
$\{3, -1, 0.5, 1, -3, 2, 2, -2\}$ respectively. The true model is

$$y = 10 + 3x_1 - x_3 + 0.5x_4 + x_1x_3 - 3x_1x_4 + 2x_3x_4 + 2x_1^2 - 2x_3^2. \tag{3.15}$$

Figure 8 shows which terms are selected by model selection in each round for each design and analysis method. Sequential designs have two rounds, the first round contains the results from model selection on the initial screening runs, and the second round displays results from model selection on the full total runs after design augmentation. Nonsequential designs only have one round. Figure 8 highlights the columns corresponding to the active terms in yellow. For the screening design rounds, terms chosen by model selection on the initial screening runs are specified as primary and labeled "p" in the table and the rest are specified potential. If no quadratics are significant in the analysis of the two-level screening design then all quadratic terms are removed from the model and indicated in the table by "-". Rows corresponding to final results are shown in the light gray shaded rows, with a solid circle if the term was included in the final LASSO or forward selection model and identified as active. Power, the number of active terms correctly identified divided by total active terms, and FDR, the number of inactive terms incorrectly identified as active divided by total number of terms labeled active, are calculated for the final round.

The sequential approach starts with either a two-level Bayesian D-optimal screening design (see Table 3(a) for design) or a three-level Bayesian D-optimal screening design (see Table 3(b). The nonsequential approach employs a 32 run Bayesian D-optimal design with main effects and quadratics labeled primary, and 2FI as potential (Table 3(b)).

To start, consider the sequential approach with a two-level screening design and the LASSO analysis method shown in the first two rows of Figure 8. For the first round, we simulate responses to the 20 run screening design locations and include



Design Analysis, Stage		x1	x2	x3	x4	x5	x6	x1x2	x1x3	x1x4	x1x5	x1x6	x2x3	x2x4	x2x5	x2x6	x3x4	x3x5	x3x6	x4x5	x4x6	x5x6	$x1^2$	$x2^2$	$x3^2$	$x4^2$	$x5^2$	$x6^2$	Power	FDR
Two-Level Sequential	LASSO Screen	p							p	p							p			p			-	-	-	-	-	-		
	LASSO Final	●		●	●				●	●				●			●						-	-	-	-	-	-	0.75	0.14
	Forward Screen	p		p	p			p	p	p						p	p						-	-	-	-	-	-		
	Forward Final	●	●	●	●				●	●				●			●	●		●	●		-	-	-	-	-	-	0.75	0.45
Three-Level Sequential	LASSO Screen	p		p					p	p	p	p					p		p				p					p		
	LASSO Final	●		●	●	●		●	●	●	●	●					●						●	●	●			●	1	0.43
	Forward Screen	p		p					p	p		p		p				p				p	p	p				p		
	Forward Final	●	●	●	●			●	●	●					●	●	●		●				●		●				1	0.38
Non-sequential	LASSO Final	●	●	●	●			●	●	●	●	●					●	●				●	●		●		●		1	0.47
	Forward Final	●	●	●	●				●	●	●		●	●			●						●		●	●		●	1	0.43

Legend: ● identified in final model
- quadratic terms removed from analysis, no quadratic terms selected by model selection
p primary term, selected by model selection on screening design
<blank> potential terms or not identified in final model

Fig. 8.: Simulation example for 6 factors, 32 total runs and 8 active effects

quadratic terms into the model selection process to test for curvature. LASSO model selection on the screening design's initial 20 runs identifies 5 terms $\{x_1, x_1x_3, x_1x_4, x_3x_4, x_4x_5\}$ in the model with the minimum AICc. No quadratic terms are selected, and as there is no evidence of curvature, all quadratics are excluded from the subsequent design augmentation and final analysis steps. The identified 5 terms are labeled primary and the remaining main effects and 2FI are potential. We find 12 (32 total runs minus 20 screening design runs) follow up runs using Bayesian D-optimal augmentation and the coordinate exchange algorithm. LASSO analysis on the 32 total runs selects the model with 7 terms $x_1, x_3, x_4, x_1x_3, x_1x_4, x_2x_4, x_3x_4$. Active terms $x_1, x_3, x_4, x_1x_3, x_1x_4, x_3x_4$ are correctly identified, x_2x_4 is incorrectly identified as active, and quadratic terms x_1^2, x_3^2 are incorrectly not identified as active. The final result for the two-level design with LASSO and two stages has a power of $6/8 = 0.75$ and FDR of $1/7 = 0.14$.

This next example walks through the process for a three-level screening design with forward selection with the results for the first round in row 7, and final results in row 8 of Figure 8. For the first round, we simulate responses to the 20 run screening design locations. Forward selection on the screening design's initial

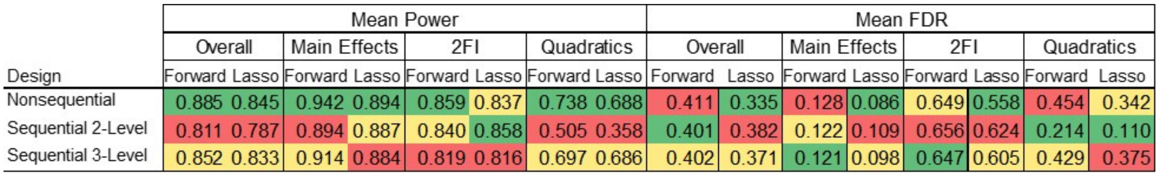
20 runs identifies 11 terms $\{x_1, x_3, x_1x_3, x_1x_4, x_1x_6, x_2x_4, x_3x_5, x_5x_6, x_1^2, x_2^2, x_6^2\}$ in the model with the minimum AICc. The 11 terms are labeled primary and the remaining main effects, 2FI and quadratic terms are labeled potential. We find 12 follow up runs using Bayesian D-optimal augmentation and the coordinate exchange algorithm. Forward selection on the 32 total runs selects the model with 13 terms $x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1x_4, x_2x_5, x_2x_6, x_3x_4, x_3x_6, x_1^2, x_3^2$. All active terms are correctly identified and five extra terms $\{x_2, x_1x_2, x_2x_5, x_2x_6, x_3x_6\}$ are incorrectly identified as active. The final result for the three-level design with forward selection has a power of 1 and FDR of $5/13 = 0.38$.

Lastly, consider the nonsequential design with forward selection with one stage in row 10 of Figure 8. Forward selection on the design's 32 runs selects a model with 14 terms x_1, x_2, x_3, x_4, x_1x_3, x_1x_4, x_1x_5, x_2x_3, x_2x_4, x_3x_4, x_1^2, x_3^2, x_4^2, x_6^2. All active terms are correctly identified and six additional terms, x_2, x_1x_5, x_2x_3, x_2x_4, x_4^2, and x_6^2 are incorrectly identified as active. The final results for the nonsequential design with forward selection has a power of 1 and FDR of $6/14 = 0.43$.

3.5 Simulation Results

Table 5 shows the mean power and FDR by analysis method overall, and for main effects, 2FI and quadratic terms. Overall, the nonsequential design with forward selection achieves the highest mean power (0.885); however this combination also has the highest overall mean FDR (0.411). The sequential three-level design with forward selection has the second highest mean power (0.852) as well as the second highest mean FDR (0.402). The sequential two-level design with forward selection achieves a slightly lower FDR than the three-level design (0.401) and a lower mean power of 0.811. Turning to the LASSO method, out of the three designs, the nonsequential design has the highest overall power (0.845) as well as the lowest overall mean

FDR (0.335). The sequential two-level design has the lowest mean power (0.787) but highest FDR rate with LASSO (0.382). The two-level sequential design also has the lowest mean power for quadratic terms, 0.505 with forward selection and 0.358 with LASSO. The sequential two-level design does better at identifying 2FI having the highest mean 2FI power with LASSO (0.858) and the second highest mean 2FI power with forward selection (0.840).



	Mean Power								Mean FDR							
	Overall		Main Effects		2FI		Quadratics		Overall		Main Effects		2FI		Quadratics	
Design	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso
Nonsequential	0.885	0.845	0.942	0.894	0.859	0.837	0.738	0.688	0.411	0.335	0.128	0.086	0.649	0.558	0.454	0.342
Sequential 2-Level	0.811	0.787	0.894	0.887	0.840	0.858	0.505	0.358	0.401	0.382	0.122	0.109	0.656	0.624	0.214	0.110
Sequential 3-Level	0.852	0.833	0.914	0.884	0.819	0.816	0.697	0.686	0.402	0.371	0.121	0.098	0.647	0.605	0.429	0.375

Table 5.: Overall, main effect, 2FI and quadratic mean power and FDR by analysis method (forward selection and LASSO). Cells are color coded green if they contain the highest mean power or lowest mean FDR per column while red indicates lowest mean power or highest mean FDR per column.

Figure 9 plots mean power and mean FDR by the total number of active terms. The designs perform similarly when the total number of active terms is small. For two active terms, the sequential two-level design has the highest mean power with LASSO but achieves a lower mean power for total active terms greater than two, and maintains the highest mean FDR regardless of the number of active terms. Nonsequential designs sustain the highest mean power with forward selection and mean FDR falls below the other designs when there are more than 9 active terms. Both model selection methods show a slight spike in mean power for the sequential two-level design at 12 active terms due to half the simulation scenarios having 0 active quadratics at that value.

To see how the designs fare when there are no active quadratic terms, Table 6 provides the summary metrics for simulation scenarios that exclude active quadratic terms in the underlying model. The nonsequential design with forward selection

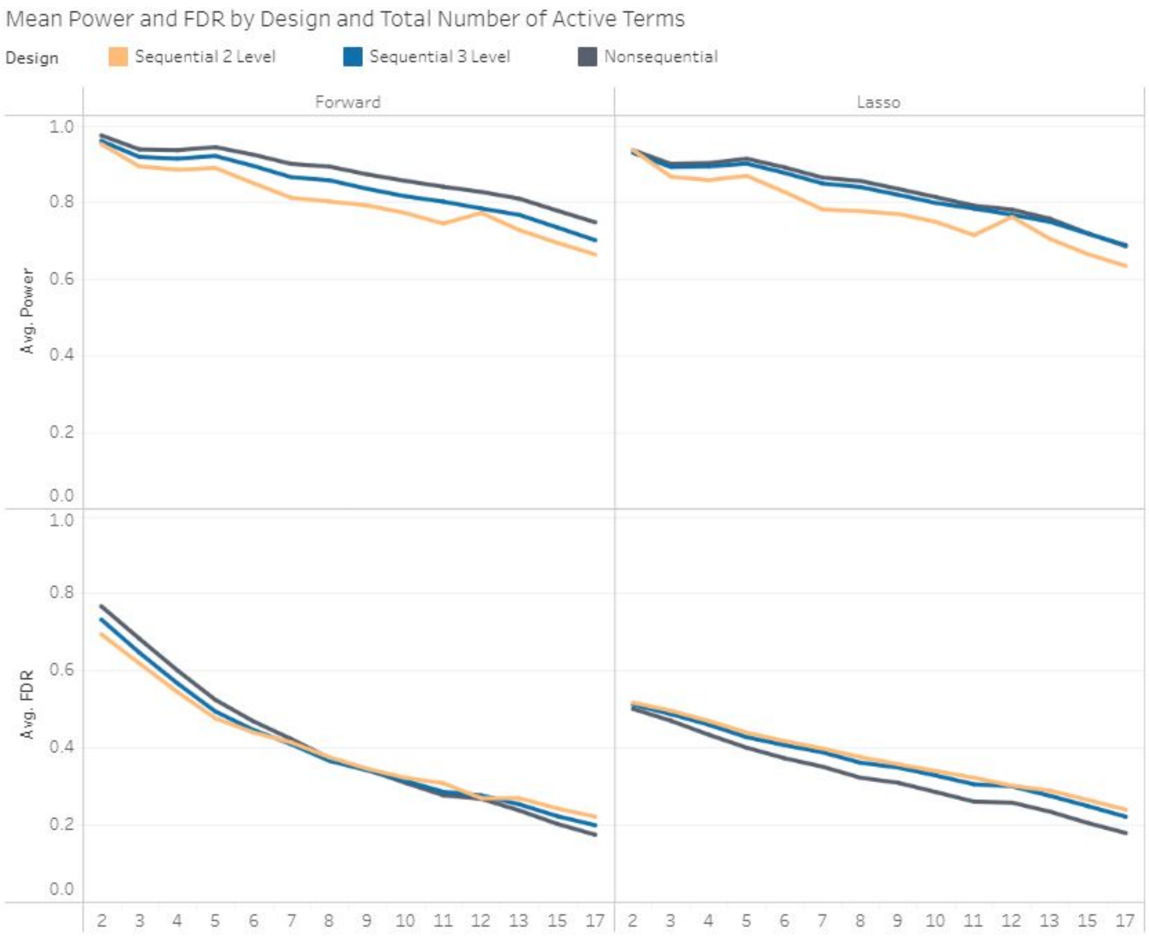
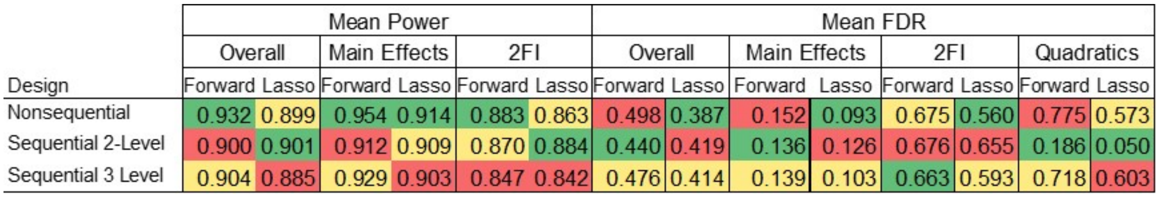


Fig. 9.: Mean Power and FDR by design and total number of active terms (Mean Power and FDR on y-axis, total active terms on x-axis)

maintains the highest mean power overall (0.932) and the sequential two-level design now has the highest mean power overall with LASSO (0.901). Notably, the sequential two-level design exhibits the lowest mean quadratic FDR rate, 0.186 with forward selection and 0.05 with LASSO, while the other two designs have a mean quadratic FDR over 0.5 and as high as 0.775 as seen with the nonsequential design and forward selection. This illustrates the benefit of not using a three-level design when quadratic terms are not expected. Using a three-level design leads to identifying quadratic terms that are not actually important.



	Mean Power						Mean FDR							
	Overall		Main Effects		2FI		Overall		Main Effects		2FI		Quadratics	
Design	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso	Forward	Lasso
Nonsequential	0.932	0.899	0.954	0.914	0.883	0.863	0.498	0.387	0.152	0.093	0.675	0.560	0.775	0.573
Sequential 2-Level	0.900	0.901	0.912	0.909	0.870	0.884	0.440	0.419	0.136	0.126	0.676	0.655	0.186	0.050
Sequential 3 Level	0.904	0.885	0.929	0.903	0.847	0.842	0.476	0.414	0.139	0.103	0.663	0.593	0.718	0.603

Table 6.: No active quadratics: overall, main effect, 2FI and quadratic mean power and FDR by analysis method (forward selection and LASSO). Power for quadratic terms is 0 as there are no active quadratics. Cells are color coded green if they contain the highest mean power or lowest mean FDR per column while red indicates lowest mean power or highest mean FDR per column.

To formalize the simulation results, we fit two ANOVA models each with either power or FDR as the dependent variables and simulation scenario attributes as predictors. The ANOVA results are included in Table 16 in the Appendix. To highlight the ANOVA model results, we also provide plots visualizing the largest F-values for each model in Figure 21(a) for the power model, and Figure 21(b) for the FDR model in the Appendix. We use these charts to guide the selection of variables to include in the plots visualizing our simulation results.

Total runs has the largest F-value in the ANOVA power model and the charts in Figure 10 show power noticeably increases with the number of total runs for the smaller run sizes and levels out from 48 to 64 runs. The nonsequential design tends to have a higher mean overall power and main effect power than the sequential designs with a larger difference between designs seen at smaller run sizes using forward selection (see Figure 10(a) Avg. Power, Avg. Power ME). The sequential three-level design performs similar to the nonsequential design in terms of overall, main effect, 2FI and quadratic power with LASSO. The sequential two-level design edges above the nonsequential design with LASSO for mean 2FI power but mean quadratic power for the sequential two-level design is distinctively lower than the other two designs considering both analysis methods (see Figure 10(a) Avg. Power Quad). Looking at power by total runs and number of factors in Figure 10(b), a decrease in mean power is seen as the number of factors increase for all design types at 20 and 32 total runs. At 20 runs, the sequential three-level design exhibits a steeper drop in main effect power and quadratic mean power, while the nonsequential design shows a sharper decrease with mean 2FI power. Figure 10(c), total runs and the number of active 2FI, reveals mean power decreases as the number of active 2FI increase for the smaller run sizes of 20 and 32 runs. Reviewing power by total runs and the number of active quadratics, 10(d), mean power for all categories decreases for all designs as the number of active quadratics increase for the smaller run sizes of 20 and 32 runs.

Mean power deteriorates in particular as the number of active quadratic terms increase for the sequential two-level design (see Figure 10(d) Avg. Power and Avg. Power Quad). Given the presence of active quadratics, the initial two-level screening design with three center runs fails to identify the presence of quadratic effects overall 42% of the time. Table 7 breaks down the percent of times the two-level screening design fails to identify any quadratic effects by method and screening design size. The smallest two-level screening design of 12 runs fails to capture the presence of quadratic terms 42% of the time using forward selection and 90% with LASSO. For the largest screening design run size of 48 runs, the two-level screening design identifies quadratic terms using forward selection 21% and LASSO 29% of the time.

Table 7.: Sequential two-level design with 3 center runs: percent of times active quadratic terms not identified in initial screening design by analysis method and initial run size

	Screening Design Run Size			
Method	12	20	32	48
Forward	42%	27%	21%	21%
LASSO	90%	67%	40%	29%

Plots of mean FDR overall, for main effects, 2FI and quadratics by design and

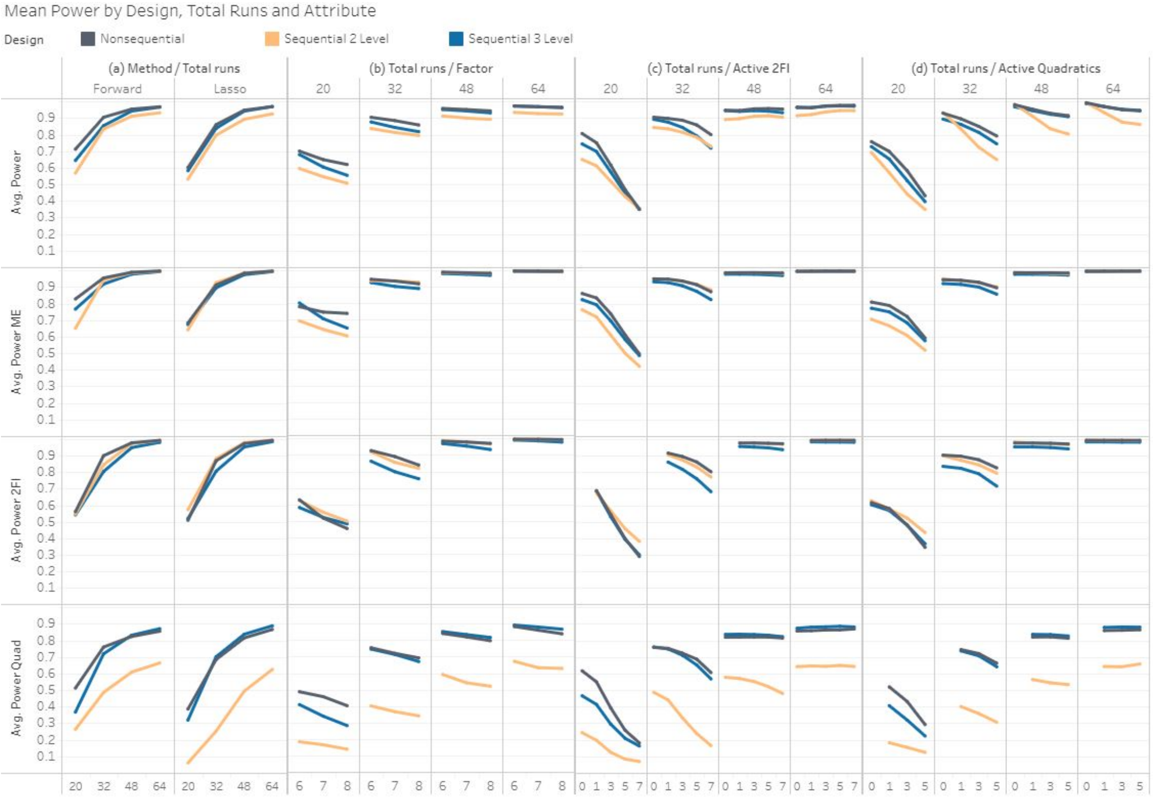


Fig. 10.: Mean power by design, total runs, and attribute (mean power on y-axis, total runs and attribute on x-axis)

analysis method are shown in Figure 11. Figure 11(a) shows with LASSO, the nonsequential design has lower mean overall, main effect, and 2FI FDR than the sequential designs, and mean FDR increases with total run size for all designs. With forward selection, FDR remains level as the number of runs increase, and the nonsequential design has higher mean FDR than the sequential designs for all run sizes except at 20 runs. Looking at the number of factors, mean FDR increases with the number of factors using forward selection for all designs, while with LASSO, a smaller slope is observed 11(b). A decrease in mean FDR is seen for all designs as the number of active 2FI and the number of active quadratics increase (11(c,d,e)). The sequential two-level design maintains significantly lower mean quadratic FDR compared to the

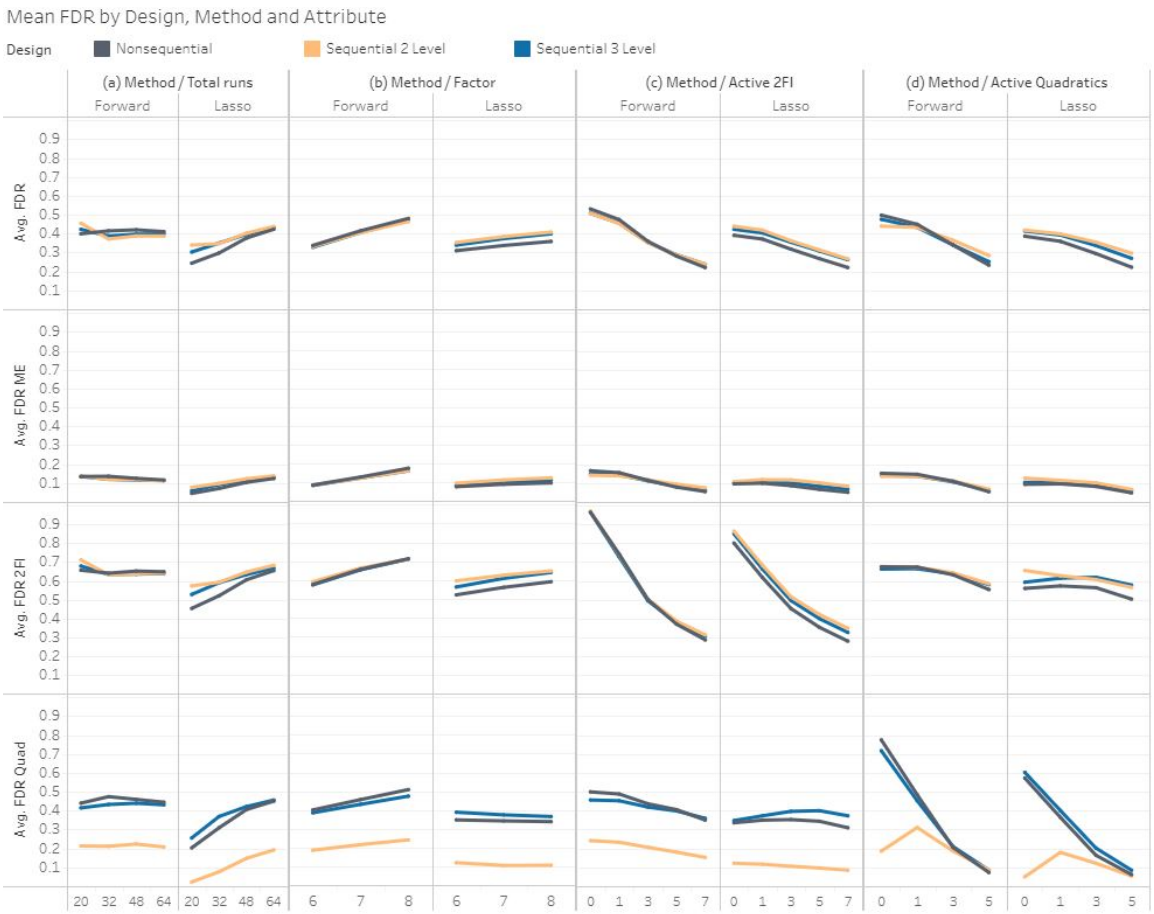
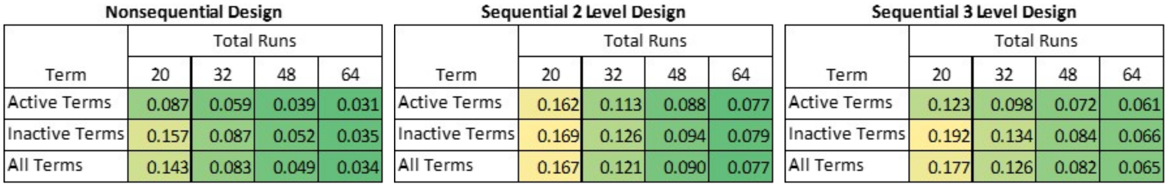


Fig. 11.: Mean FDR by design, method, and attribute (Mean FDR on y-axis, method and attribute on x-axis)

other two designs for less than 3 active quadratic terms however all designs perform similarly with 3 or 5 active quadratic terms (see 11(d) Avg. FDR Quad).

To investigate why nonsequential designs outperform sequential designs in terms of mean power, we compared the mean absolute value of term correlation for active, inactive and all terms in Table 8. For sequential designs, this metric is taken after the design is augmented. The nonsequential design has a lower mean absolute value of term correlation than both sequential designs, helping the identification of active effects. The sequential three-level design has lower mean absolute value of term correlation than the sequential two-level design. The correlation maps of the initial designs, Figure 7, also reveal lower correlation for the nonsequential design compared to both initial screening designs.



Nonsequential Design

Term	Total Runs			
	20	32	48	64
Active Terms	0.087	0.059	0.039	0.031
Inactive Terms	0.157	0.087	0.052	0.035
All Terms	0.143	0.083	0.049	0.034

(a)

Sequential 2 Level Design

Term	Total Runs			
	20	32	48	64
Active Terms	0.162	0.113	0.088	0.077
Inactive Terms	0.169	0.126	0.094	0.079
All Terms	0.167	0.121	0.090	0.077

(b)

Sequential 3 Level Design

Term	Total Runs			
	20	32	48	64
Active Terms	0.123	0.098	0.072	0.061
Inactive Terms	0.192	0.134	0.084	0.066
All Terms	0.177	0.126	0.082	0.065

(c)

Table 8.: Design terms' total run mean absolute value of correlation

We performed sensitivity analysis on the choice of parameter $\tau = \{1, 5, 20, 50, 100\}$, the prior variance given to potential terms during Bayesian D-optimal augmentation. An ANOVA on the results of 10 iterations varying the levels of τ reveals no significant difference in power and FDR between these choices of τ.

As part of our research, we examined various other approaches in comparing sequential to nonsequential designs leading up to our final simulation. We investigated the performance of two-level sequential and nonsequential designs without the presence of any active quadratic terms. Our Bayesian D-optimal screening design and nonsequential designs had main effects specified as primary and 2FI as potential. Our results showed the nonsequential design had higher power overall while a benefit to the sequential design was seen looking at 2FI power. This provides some evidence that sequential design performance may help when little is known about the underlying true model by allowing the experimenter to pivot and add runs to help the identification of terms initially not considered primary.

For the sequential designs, we additionally considered dropping terms from further design augmentation and analysis stages that were not identified by model selection on the screening design results. In this case, terms selected by screening design model selection were considered primary, any non primary terms with a heredity relationship to the primary terms were labeled potential, and all remaining terms were dropped. The results, however, revealed a decrease in mean power for sequential designs. We added a third option of analyzing and augmenting the sequential design twice to see if a more incremental approach was beneficial, but there was no significant difference between mean power and FDR if we augmented the screening design once or twice keeping the number of total runs constant. Furthermore, we originally included A- and I-optimal designs but D-optimal designs performed more favorably in terms of power and error rates for both nonsequential and sequential designs. Lastly, we considered augmenting sequential designs using Bayes screening and model discrimination follow-up designs via the R `BsMD` package (Barrios and Meyer, 2020), but this augmentation method resulted in lower mean power compared to Bayesian D-optimal augmentation.

3.5.1 Desirability Analysis

Our results from the summary table (Table 5) reveal attributes that have a positive impact to power may also unfavorably increase FDR. For instance, a nonsequential design with forward selection generated the highest mean power but also the highest mean FDR for overall, main effect, and quadratic terms. To find a balance between high power and low FDR levels, we utilize multiple response optimization with desirability functions to provide an overall design recommendation considering a sequential two-level, sequential three-level or nonsequential design. Upper and lower thresholds representing the acceptable range are specified for each response. Each objective is translated into an individual desirability function on a scale ranging from 0 to 1, with 1 indicating the objective is at the optimal target level (Myers et al., 2016). Let $f_r(\boldsymbol{X})$ for $r = \{1, ..., R\}$, represent a function to optimize. The desirability function to maximize $f_r(\boldsymbol{X})$ is

$$d_r^{max} = \begin{cases} 0, & f_r(\boldsymbol{X}) < L_r \\ \left(\frac{f_r(\boldsymbol{X})-L_r}{T_r-L_r}\right)^w, & L_r \leq f_r(\boldsymbol{X}) \leq T_r \\ 1, & f_r(\boldsymbol{X}) > T_r \end{cases}$$

where L_r and T_r are the lower and top user specified thresholds, and weight $w = 1$ is linear desirability function. The desirability function to minimize $f_r(\boldsymbol{X})$ is

$$d_r^{min} = \begin{cases} 1, & f_r(\boldsymbol{X}) < L_r \\ \left(\frac{T_r-f_r(\boldsymbol{X})}{T_r-L_r}\right)^w, & L_r \leq f_r(\boldsymbol{X}) \leq T_r \\ 0, & f_r(\boldsymbol{X}) > T_r \end{cases}$$

where L_r is the optimal target level and T_r is a larger but still acceptable level. The functions are combined using the geometric mean

$$D = \left(\prod_{r=1}^{R} d_r\right)^{1/R} \tag{3.16}$$

to produce an overall desirability score and the preferred design has the highest mean desirability score.

Our goal is to find the design that maximizes predicted power and minimizes predicted FDR. The main effect with 2FI regression model is used for each function $f_r(\boldsymbol{X})$, $r = 1, 2$. Targets are established with the desire to have at least above average power and below average FDR. For power, the desirability function maximizes $f_1(\boldsymbol{X})$ where target L_1 is the mean predicted power (0.835) and $T_1 = 1$. For FDR, the desirability function minimizes $f_2(\boldsymbol{X})$ where $L_2 = 0$ and T_2 is the mean predicted FDR (0.384). We generate the overall desirability score in R (R Core Team, 2018) using the package `desirability` and the `dMax,dMin,dOverall` functions (Kuhn, 2016).

Our desirability analysis reveals the design with the highest mean desirability score (0.080) is the nonsequential design followed by the sequential three-level design (0.067). Table 9 shows the mean desirability score for each design as well as the mean predicted power, mean predicted FDR, mean desirability score for power, and mean desirability score for FDR. The nonsequential design has the highest mean predicted power out of the designs (0.865) as well as lowest mean predicted FDR (0.373). We repeat the desirability analysis focusing on main effect power and FDR (see Table 17 in the Appendix), 2FI power and FDR (Table 18 in the Appendix), quadratic term power and FDR (Table 19 in the Appendix) as well as when there are no active quadratic terms in the underlying model (Table 20 in the Appendix). The nonsequential design maintains the highest desirability score for the main effect, 2FI, and quadratic term power and FDR analysis. The sequential two-level design outperforms the sequential three-level design when focusing on 2FI power and FDR. When there are no active quadratic terms, sequential two-level designs outperform both nonsequential and sequential three-level designs in terms of desirability.

Table 9.: Desirability analysis results (mean values)

Design	Predicted Power	Predicted FDR	Desirability Power	Desirability FDR	Overall Desirability
Nonsequential	0.865	0.373	0.413	0.160	0.080
Sequential 3 Level	0.842	0.386	0.366	0.122	0.067
Sequential 2 Level	0.799	0.413	0.273	0.105	0.056

3.6 Summary and Conclusions

The performance of sequential variance optimal designs compared to a single variance optimal design without screening is an open problem of optimal design. This chapter evaluates sequential design techniques and nonsequential designs with regards to power and FDR rates. Our simulation study varies the number of total experimental runs, number of input factors, number of active terms, and considers model selection methods of LASSO and forward selection. The designs are assessed based on mean power and FDR rates. We limit our scope to focus on Bayesian D-optimal designs and Bayesian D-optimal design augmentation. Our study aims to help experimenters decide whether to run a small screening design first or use all runs in a nonsequential approach to identify active terms when faced with a limited number of experimental trials.

Our simulation results indicate a difference in power and FDR between sequential and nonsequential designs as well as between forward selection and LASSO analysis methods. Our simulation results support the use of nonsequential designs whether maximizing power is the main objective or balancing high power with low FDR rates. Prior to this study, we conjectured sequential designs would achieve higher mean power compared to nonsequential designs given the ability to specify follow up runs based on initial screening design findings. A limitation of our study is the design space is constrained to a fixed design region. Our study did not consider the ability to move outside the original design region that may be a key benefit to sequential experiments and we recommend this for future research. We also focus primarily on Bayesian D-optimal designs as well as Bayesian D-optimal augmentation. Other design types as well as other design augmentation methods may yield different results.

It may be worth highlighting the different approaches taken in our study from the motivating example in Silvestrini (2015). Silvestrini (2015) assumes the sequential design after augmentation is able to estimate the true model. Our simulation study found an initial two-level screening design with three center runs failed to identify the presence of active quadratic terms 42% of the time. Without any evidence of curvature, the final augmented sequential design remained a two-level design and was not able to estimate the true quadratic model. Additionally, we compare the use of a two as well as a three-level initial screening design in sequential experimentation to a three-level nonsequential design while the previous study compares initial designs with the same assumed model; for instance, the use of a two-level screening design to a two-level nonsequential design.

CHAPTER 4

MODEL SELECTION WITH PURE ERROR

4.1 Background and Literature Review

The goal of screening designs is to correctly identify a subset of factors that are important so subsequent analysis can focus on key effects and their relationships. If the screening design fails to identify an important factor, that factor is not included in the follow-up analysis and the results will not fully capture the underlying system. Conversely, if an inactive effect is incorrectly identified as important the results may mislead the experimenter as well as cost additional resources to research the factor further. Thus screening designs are evaluated in terms of how well the designs accurately identify truly active effects. Ideal designs have high power, the proportion of active terms that are correctly identified, a low false discovery rate (FDR), the proportion of terms incorrectly identified as active, and Type I error, the proportion of inactive terms declared active.

It is often the case in a screening experiment that there are more effects to be estimated than the number of screening runs. The aliasing of factorial effects from small run sizes and limited degrees of freedom to estimate all parameters poses challenges in identifying active effects during the model selection process. As a way to improve power, studies (see Westfall et al. (1998), Gilmour and Trinca (2012), Leonard and Edwards (2017), Mee et al. (2017)) have based statistical inference on a pure error estimate versus the residual mean square during model selection procedures. Replicate points provide a model independent estimate of the variance known as pure error. Studies on model selection for screening designs (see Westfall et al. (1998), Leonard and Edwards (2017), Mee et al. (2017)) find higher power rates when statistical inference is based on a pure error variance estimate versus the residual mean square; however FDR and Type I error rates can be problematically inflated (Westfall et al. (1998), Leonard and Edwards (2017), Mee et al. (2017)).

Westfall et al. (1998) propose multiple test procedures to control for the inflation of Type I error in forward selection seen in the analysis of supersaturated designs (SSDs). The authors stress the difficulty of identifying important factors in SSDs as many Type I and Type II errors are expected using stepwise forward selection. Forward selection is an iterative model selection approach adding a single term to the model during each step having the largest significant partial F-statistic out of the candidate terms; partial F-statistic p-values less than or equal to some specified significance level (α) are considered significant. The process terminates when none of the remaining candidate terms have a significant partial F-statistic or all terms have been added to the model. Each step calculates a individual partial F-test for each candidate term. However with multiple tests, the experiment-wise error rate (EER) is not ensured to be at most α. The Bonferroni approach decreases the critical level α at the individual test level to be α divided by the number of tests so that the EER$\leq \alpha$ (Rencher and Schaalje, 2008). Alternatively, partial F-test p-values are multiplied by the number of eligible terms in each step for the Bonferroni correction. To help control for the compounding of errors occurring with sequential tests in forward selection, Westfall et al. (1998) propose using adjusted p-values either with the Bonferroni method or via resampling using control variates.

Gilmour and Trinca (2012) introduce new design criteria that supports statistical inference with pure error. Textbooks, such as Myers (1990) on page 19, often recommend using the residual mean square error from the fitted model for observational data. However Gilmour and Trinca (2012) caution that most textbooks on the design and analysis of experiments, for instance Antony (2003) on page 101, recommend using the pure error mean square for inference, stating pure error provides the most reliable estimate of variance available. The residual mean square error is model dependent and contains an unknown level of bias making inferences difficult to interpret (see Gilmour and Trinca (2012) and a motivating example is provided in Section 4.2). Gilmour and Trinca (2012) recommend using a pure error estimate from replicate design points and designing an experiment with sufficient degrees of freedom for pure error.

Leonard and Edwards (2017) extend the work of Gilmour and Trinca (2012) to screening designs. The authors develop partially replicated designs with a modified Bayesian D-optimality criterion. The modified Bayesian D-optimality criterion is a balance between traditional Bayesian D-optimal designs that do not often contain replicated points and DP-optimal designs that often over-replicate. While their findings display higher FDR for partially replicated designs, the results also reveal higher power rates using inference with mean square pure error. The authors note that higher power is often preferred over reductions in the false positive rate during screening to ensure an active effect is not excluded in further analysis.

Mee et al. (2017) compare the model selection performance of forward selection and the Dantzig selector (Candes and Tao, 2007) in two-level fractional factorial screening designs. Results show two-level fractional factorial screening designs achieve superior power when using a model independent estimate with forward selection but also attain higher FDR. The authors cite the need for a method to reliably integrate pure error in model selection procedures without high FDR. They also recommend investigating the performance of penalized regression with pure error for future research.

Currently there is not a suitable method available to incorporate pure error into model selection procedures when analyzing screening designs that achieves high power without the trade-off of high FDR. In this chapter, we detail the behavior of the partial F-test statistic's noncentrality parameter(s) to better understand the impact of including pure error in model selection. We provide a small simulation as a motivating example in Section 4.2 to compare the performance of statistical inference with pure error verses residual mean square during model misspecification. We then detail various model selection strategies with and without pure error, including the use of penalized regression, followed by a large simulation study comparing the performance of 11 different model selection methods across various scenarios. Our goal is to propose a strategy for incorporating pure error in model selection that attains high power and also inhibits FDR.

4.1.1 The F-statistic

We first introduce noncentral chi-square distributions followed by a discussion of noncentral F-distributions. For a random vector $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I})$, the sum of its squares $\sum_{i=1}^{n} b_i^2 = \boldsymbol{b}'\boldsymbol{b}$ is distributed as $\chi^2(n)$, a central chi square random variable with n degrees of freedom. Consider a random vector $\boldsymbol{d}$ with mean μ instead of mean 0, $\boldsymbol{d} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$; the sum of its squares $\sum_{i=1}^{n} d_i^2 = \boldsymbol{d}'\boldsymbol{d}$ is distributed as $\chi^2(n, \lambda)$, a noncentral chi square random variable with n degrees of freedom and a noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}$. A comparison of central and noncentral chi-square densities is shown in Figure 12, notice how the noncentral distribution is right-skewed having more observations in the right tail.

An F-test statistic is a ratio of two independent mean squares and has an F-distribution. A central F-distribution is the ratio of two chi-squared variates $s = \frac{u/p}{v/q} \sim F(p, q)$ where $u \sim \chi^2(p)$ with p degrees of freedom and $v \sim \chi^2(q)$ with q degrees of freedom. Alternatively, if u is distributed as a noncentral chi-square $u \sim \chi^2(p, \lambda)$

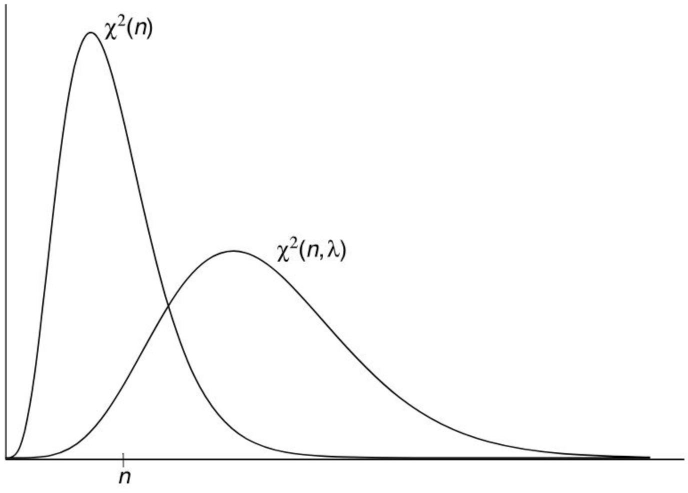


Fig. 12.: Central and noncentral chi-square densities (Rencher and Schaalje, 2008).

and v remains $v \sim \chi^2(q)$, then

$$w = \frac{u/p}{v/q} \sim F(p, q, \lambda) \tag{4.1}$$

is now a noncentral F-distribution with noncentrality parameter λ. During a statistical test of a null hypothesis, an F-distribution typically will be centrally distributed if the null hypothesis is true and noncentrally distributed if the hypothesis is false. Rencher and Schaalje (2008) defines the power of a test as the probability of rejecting the null hypothesis for a given value of λ. As shown by the shaded region in Figure 13, the power of the F-test is defined as $P(p, q, \alpha, \lambda) =$Prob$(w \geq F_\alpha)$, where F_α is the upper α percentage point of the central F-distribution and w is the noncentral random variable defined in Equation 4.1.

To illustrate, consider the model M with h regressors in the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is the $n \times p$ model matrix with n number of runs and p parameters, $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ as the $p \times 1$ vector of coefficients, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$. As shown in Rencher and Schaalje (2008), the total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ can be partitioned into the model sum of squares (SSM) plus the sum of squares error

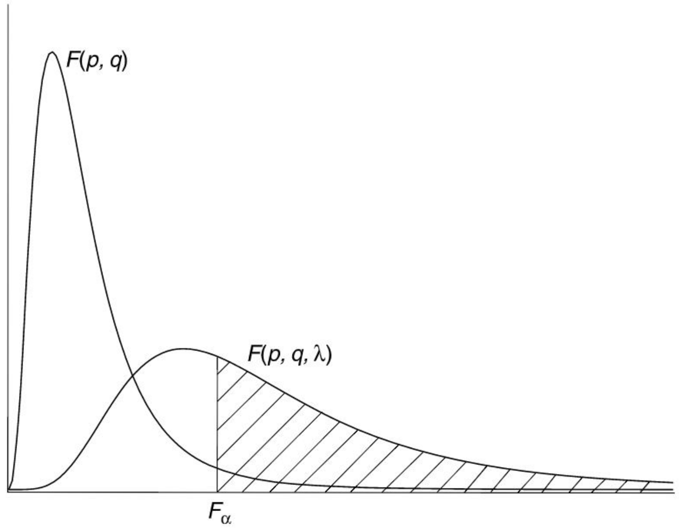


Fig. 13.: Central $F(p,q)$, and noncentral $F(p,q,\lambda)$. The shaded area is the power of the F-test (Rencher and Schaalje, 2008).

(SSE),

$$SST = SSM + SSE. \tag{4.2}$$

This is

$$\begin{aligned} \boldsymbol{y}'\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{y} &= SSM + SSE \\ &= \boldsymbol{y}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{y} + \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} \end{aligned} \tag{4.3}$$

where the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{I}$ is an identity matrix of size n, and $\boldsymbol{J}$ is a $n \times n$ matrix of ones.

Frommlet et al. (2010) detail the behavior of the partial F-test statistic in regards to testing model fit during model selection when the correct model is unknown. Suppose model M, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, is the true model and can be partitioned as $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$. We instead fit model M_1, $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$, leaving out $\boldsymbol{X}_2\boldsymbol{\beta}_2$. We want to test $H_0 : \beta_j = 0$, $\beta_j \in \boldsymbol{\beta}_1$. The hat matrix for the true model M is $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and the hat matrix for the model M_1 is $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$.

To test the significance of β_j we use $\boldsymbol{H}_{1,-j} = \boldsymbol{X}_{1,-j}(\boldsymbol{X}'_{1,-j}\boldsymbol{X}_{1,-j})^{-1}\boldsymbol{X}'_{1,-j}$ where $\boldsymbol{X}_{1,-j}$ is $\boldsymbol{X}_1$ without column j. The Type III sums of squares of β_j for the fitted model M_1 (SS_{β_j}) is

$$SS_{\beta_j} = \boldsymbol{y}'(\boldsymbol{H}_1 - \boldsymbol{H}_{1,-j})\boldsymbol{y} \tag{4.4}$$

and

$$\frac{SS_{\beta_j}}{\sigma^2} \sim \chi^2(1, \frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{H}_1 - \boldsymbol{H}_{1,-j})\boldsymbol{X}\boldsymbol{\beta}). \tag{4.5}$$

Note the noncentrality parameter includes all the terms in the true model $\boldsymbol{X\beta}$. The sums of square error of model M_1 is

$$SSE_1 = \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{H}_1\right)\boldsymbol{y} \tag{4.6}$$

and

$$\frac{SSE_1}{\sigma^2} \sim \chi^2(n - p_1, \frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}\boldsymbol{\beta}) \tag{4.7}$$

where p_1 is the number of terms in $\boldsymbol{X}_1$. This simplifies to

$$\frac{SSE_1}{\sigma^2} \sim \chi^2(n - p_1, \frac{1}{2\sigma^2}\boldsymbol{\beta}_2'\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\boldsymbol{\beta}_2). \tag{4.8}$$

We have

$$\mathbb{E}(MSE) = \sigma^2 + \frac{\boldsymbol{\beta}_2'\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\boldsymbol{\beta}_2}{n - p_1} \tag{4.9}$$

and if the fitted model is adequate, $\boldsymbol{\beta}_2 = 0$, then $\mathbb{E}(MSE) = \sigma^2$. The partial F-statistic, the contribution of β_j given the terms already added to model M_1, is

$$\frac{SS_{\beta_j}}{SSE_1/n - p_1} \sim F(df_{SS_{\beta_j}}, df_{SSE}, \frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{H}_1 - \boldsymbol{H}_{1,-j})\boldsymbol{X}\boldsymbol{\beta}, \frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}\boldsymbol{\beta})), \tag{4.10}$$

a doubly noncentral F-distribution with SS_{β_j} degrees of freedom $df_{SS_{\beta_j}} = 1$ and SSE_1 degrees of freedom $df_{SSE} = n - p_1$.

4.1.2 Pure Error

To compute mean square pure error, MSPE, we first create a replicate matrix $\boldsymbol{Z}$, a $n \times c$ matrix where c is the number of unique rows, to map each run to a unique point. For each column, a run is assigned 1 if it belongs to that unique point, else 0. For example, suppose we have 7 total runs for two factors A and B containing two replicated points (1,-1) and (-1,-1) as shown in Table 10. Point (1,-1) appears in run ID 1, 5 and 6, and point (-1,-1) appears in run ID 3 and 7. The model matrix with runs reordered to show replicate points together and including the intercept, main effects and two-factor interactions (2FI) is

Table 10.: Experimental design for 2 factors and 7 runs with points (1,-1) and (-1,-1) replicated

Run ID	A	B
1	1	-1
2	1	1
3	-1	-1
4	-1	1
5	1	-1
6	1	-1
7	-1	-1

$$\boldsymbol{X} = \begin{array}{cc} & \begin{array}{cc} l & m \end{array} \\ \left[\begin{array}{rrrr} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{array}\right] & \begin{array}{cc} \mathit{1} & \mathit{1} \\ \mathit{1} & \mathit{2} \\ \mathit{1} & \mathit{3} \\ \mathit{2} & \mathit{1} \\ \mathit{3} & \mathit{1} \\ \mathit{3} & \mathit{2} \\ \mathit{4} & \mathit{1} \end{array} \end{array}$$

where column l is the unique row ID and m is the row count for each unique point. Replicate point (1,-1) has $l = 1$ and is in rows 1, 2 and 3, and replicate point (-1,-1) has $l = 3$ and is in rows 5 and 6.

Each unique row corresponds to column l in matrix $\boldsymbol{Z}$,

$$\boldsymbol{Z} = \begin{array}{c} \begin{matrix} \mathit{1} & \mathit{2} & \mathit{3} & \mathit{4} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

where each of the 7 runs are mapped to a unique point indicated by having a 1 in the column corresponding to the unique row ID.

We now use sums of square pure error (SSPE) instead of SSE in the denominator of the partial F-statistic,

$$SSPE = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H}_Z)\boldsymbol{y} \tag{4.11}$$

where $\boldsymbol{H}_Z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. With pure error degrees of freedom $df_{pe} = n - c$, we have

$$MSPE = \frac{SSPE}{n - c} \tag{4.12}$$

and

$$\frac{SSPE}{\sigma^2} \sim \chi^2(n - c, \frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}_Z)\boldsymbol{X}\boldsymbol{\beta}). \tag{4.13}$$

The noncentrality parameter $\frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}_Z)\boldsymbol{X}\boldsymbol{\beta}$ reduces to 0 and the denominator of the partial F-statistic is

$$\frac{SSPE}{\sigma^2} \sim \chi^2(n-c). \tag{4.14}$$

Thus, the partial F-test statistic becomes a noncentral F-distribution with numerator noncentrality rather than a doubly noncentral F-distribution. Without the denominator noncentrality, the partial F-statistic denominator tends to be smaller resulting in a larger partial F-statistic. This leads one more often than not to reject the null hypothesis and as a result, power increases and FDR may be inflated. We investigate incorporating early stopping criteria such as a lack of fit (LOF) test and Bonferroni adjusted p-values in our simulation study to prevent high FDR with pure error.

4.2 Motivating Example

We illustrate how statistical inference with residual mean square is dependent on model fit, and we compare power and FDR performance to inference with pure error using a small simulation. Here we ignore model selection and fit a main effects model as well as a main effects and 2FI model. We use a D-optimal design with partial replication for 7 factors and 40 total runs; a run size large enough to allow us to estimate all main effects and 2FI. Designs contain $\{0, 2, 4, 6, 8, 10\}$ replicate runs as part of the 40 total runs.

We create the underlying relationship between the inputs and response by randomly selecting terms to be active based on the number of specified active terms. We assume weak heredity when selecting active 2FI, that is 2FI are present only if at least one parent main effects is active (Ockuly et al., 2017). We randomly assign model coefficients from the set $\{\pm1, \pm2, \pm3\}$ to the active terms. We use the following to generate responses,

$$y = 10 + \sum_{j=1}^{h} \beta_j x_j + \sum \sum_{i<j=2}^{h} \beta_{ij} x_i x_j + \epsilon \tag{4.15}$$

for h factors and $\epsilon \sim Normal(0, \sigma^2)$ with $\sigma^2 = 1$.

Given the t-statistic calculation,

$$\hat{\beta}_i / (\widehat{var}(\hat{\beta}_i))^{1/2}, \tag{4.16}$$

when a MSPE estimate is available we compute the variance as

$$\widehat{var}(\hat{\beta}_i) = MSPE \times \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]_{ii} \tag{4.17}$$

where $\boldsymbol{X}$ is the model matrix and the ii subscript indicates the ith diagonal element of the matrix (Leonard and Edwards, 2017). The p-values are based on a t-distribution with df_{pe} degrees of freedom. The t-statistics are calculated using residual mean square for our design with 0 replicate runs. That is,

$$\widehat{var}(\hat{\beta}_i) = MSE \times \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]_{ii} \tag{4.18}$$

with df_{SSE} degrees of freedom. Terms with a p-value≤ 0.05 are identified as active. Our simulation steps are outlined in Algorithm 3.

Algorithm 3 Motivating Example Simulation Steps

```
1: Assuming weak effect heredity for 2FI, randomly assign model terms to be active based
   on the number of active main effect terms and 2FI.
2: Create the underlying relationship between inputs and responses by assigning model
   coefficients from the set $\{\pm1, \pm2, \pm3\}$ to the active terms. Inactive terms are assigned
   a value of zero.
3: Generate a response for each run in the experimental design using 4.15.
4: Fit a model with all main effect terms, as well as a model with all main effect terms
   and 2FI. MSPE is used for statistical inference for our designs with replicate runs, else
   MSE is used. Terms with a $p$-value $\leq 0.05$ are identified as active.
5: Calculate power and FDR.
6: Repeat all previous steps 1,000 times and take the mean power and mean FDR per
   scenario.
```

Higher mean power and higher mean FDR is observed for the main effects model using statistical inference based on pure error compared to SSE as seen in Table 11. Across all replicate run sizes inference with pure error achieves a mean power of 0.973 and mean FDR of 0.190 in contrast to a mean power of 0.668 and mean FDR of 0.001 with SSE. Here MSPE is close to 1, however MSE is notably larger (23.697) when statistical inference is based on SSE as active 2FI are not included in the fitted model.

Error Type	Replicate Runs	Mean MSPE or MSE	Mean Power	Mean FDR
PE	2	1.011	0.940	0.084
	4	1.002	0.982	0.176
	6	0.991	0.989	0.165
	8	1.005	0.981	0.239
	10	0.990	0.975	0.285
SSE	0	23.697	0.668	0.001

Table 11.: Motivating example simulation results for fitted main effects model

To compare the performance when the fitted model matches that of the underlying model, Table 12 shows results for the main effects and 2FI model. MSPE is once again near 1 and now MSE is also closer to 1, as expected. SSE has a mean power of 1 and FDR of 0.097. In this case, using inference with pure error has a slightly lower mean power (0.987) and lower FDR (0.087) across replicate run sizes. As the degrees of freedom pure error increases to 10, mean power and mean FDR with pure error increase to values close to that of SSE; our SSE example has $df_{SSE} = 11$.

Error	Replicate	Mean MSPE	Mean	Mean	Mean Power	Mean FDR	Mean Power	Mean FDR
Type	Runs	or MSE	Power	FDR	ME	ME	2FI	2FI
PE	2	0.996	0.942	0.069	0.944	0.036	0.939	0.100
	4	1.005	0.995	0.086	0.997	0.040	0.994	0.133
	6	0.993	0.999	0.091	0.999	0.041	0.998	0.147
	8	1.004	0.999	0.096	0.999	0.042	0.998	0.157
	10	1.001	0.999	0.096	0.999	0.041	0.999	0.161
SSE	0	1.006	1.000	0.097	1.000	0.041	1.000	0.162

Table 12.: Motivating example simulation results for fitted main effects and 2FI model

Our motivating example simulation results show with model misspecification, that is the underlying model has active main effects and 2FIs and we fit a main effects only model, statistical inference with SSE resulted in lower mean power and a much higher MSE compared to inference with pure error. Statistical inference with pure error resulted in higher mean power as well as higher mean FDR rates. When there is no model misspecification, statistical inference with pure error produced similar mean power rates compared to SSE and lower mean FDR.

4.3 *D*-optimal Designs

We use *D*-optimal designs in this chapter. *D*-optimal designs seek to minimize the variance of the model parameter estimates (Myers et al., 2016). This criterion has proven to be particularly useful in screening scenarios where the objective is to identify the few important factors out of the many factors of interest. As the determinant can be used as a measure of overall variance, a D-optimal design locates a set of design points that minimizes the determinant of the inverse information matrix (i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ where $\boldsymbol{X}$ is the model matrix). It is computationally more tractable, however, to maximize the determinant of the information matrix. Thus, a D-optimal design is such that

$$|\boldsymbol{X}'\boldsymbol{X}| \tag{4.19}$$

is maximized.

4.4 Model Selection Methods

We compare various model selection methods with and without pure error to analyze methods of incorporating pure error that achieve higher power without inflated FDR. We test incorporating early stopping criteria of LOF as well as Bonferroni adjusted p-values to account for the unbalanced noncentrality in the numerator of the partial F-statistic when using pure error.

4.4.1 Forward Selection with SSE

Forward selection is a model selection procedure that sequentially adds terms to a regression model on the basis of partial F-tests until none of the remaining candidate terms are significant or all variables have been added to the model (Myers, 1990). The process starts with an initial model containing just the intercept term and the procedure chooses the next variable j to include that has the largest significant partial F-test out of all the candidate terms. A statistically significant partial F-test has a p-value less than or equal to some prespecified α threshold. At each stage, let the model matrix with the candidate term j be $\boldsymbol{X}_1$ and $\boldsymbol{X}_{1,-j}$ is the model selected by the previous step. The partial F-test using SSE is given by Equation 4.10 and we refer to the p-value as p_{sse} to distinguish from the p-value pure error, p_{pe}, described in the following subsection.

We enforce weak heredity during the forward selection process. For the first stage, candidate terms only include main effects. Candidate terms for the subsequent stages include remaining main effect terms and all possible 2FI under weak heredity given the main effect terms already included in the model. For example, suppose there are 7 factors, the first round of model selection contains candidate terms x_1, x_2, x_3, x_4, x_5, x_6, x_7, all 7 main effect terms. Let x_1 be the candidate term with the minimum p_{sse}, having a $p_{sse} \leq \alpha$, and is added to the model. The next step has candidate terms x_2, x_3, x_4, x_5, x_6, x_7, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_1x_6, x_1x_7, all remaining main effect terms and all 2FI including x_1. Suppose x_3 has the minimum p_{sse}, with a $p_{sse} \leq \alpha$, of the candidate terms. Term x_3 is added to the model and the following step has all remaining main effect terms and all 2FI containing either x_1 or x_3 as candidate terms. The process continues until no candidate term has a $p_{sse} \leq \alpha$ or all terms have been added to the model.

4.4.2 Forward Selection with Pure Error

This forward selection method enforces weak heredity and uses partial F-tests based on a pure error estimate. We first create a replicate matrix and calculate MSPE (see Equation 4.12). The partial F-test pure error is $F_{pe} = SS_{\beta_j}/MSPE$ with a p_{pe} equal to Prob($F_{df_{SS_{\beta_j}}, df_{pe}} \geq F_{pe}$). We enforce weak heredity during the forward selection process. The process starts with an initial model containing just the intercept term and the procedure chooses the next candidate term j to include that has the minimum p_{pe}, given $p_{pe} \leq \alpha$. The process continues until no candidate term has a $p_{pe} \leq \alpha$ or all terms have been added to the model.

4.4.3 Forward Selection with AICc

This stepwise procedure adds a candidate term to the model if the addition of the term results in the model with minimum AICc, Akaike information criterion with a correction for small sample sizes (see Smucker et al. (2020)), out of all candidate terms in each step. With

$$AIC = -2log(\hat{L}) + 2p \tag{4.20}$$

where p is the number of parameters, $p < n - 1$, and $\hat{L}$ is the maximum likelihood (Akaike, 1974). We have $AICc = AIC + \frac{2p(p+1)}{n-p-1}$. If the candidate model with the minimum AICc has an AICc greater than the existing model, the process terminates. We enforce weak heredity in the forward selection process. For example, suppose in the first step x_2 along with the intercept is the model with the lowest AICc compared to the intercept with any of the other main effect terms. Term x_2 is added to the model. The second step compares the AICc of the model with x_2 with any of the remaining main effects and 2FIs containing x_2. Suppose the model with x_2 and x_4 has the lowest AICc out of the candidate terms but has a higher AICc compared to the model with just x_2. There is no improvement to AICc by adding a term in the second step and the process terminates with the model containing just x_2 as the final model.

4.4.4 Bonferroni Adjustment with Forward Selection

We investigate the performance of forward selection with Bonferroni adjusted p-values. As discussed in Westfall et al. (1998), Benjamini and Hochberg (1995) and Mee (2013) the experiment wise error rate (EER), probability of making at least one Type I error, can be quite high with multiple statistical tests including forward selection and these studies recommend the use of Bonferroni adjusted p-values to control FDR.

Using forward selection enforcing weak heredity, we apply a Bonferroni adjustment to the p-value of the partial F-test, that is we multiply the p-value by the number of candidate terms in the forward selection step. For instance, suppose there are 7 factors, the first round of model selection includes candidate terms x_1, x_2, x_3, x_4, x_5, x_6, x_7, all 7 main effect terms. The Bonferroni adjusted p-value is the candidate's partial F-test p-value multiplied by 7. The candidate term with the minimum Bonferroni adjusted p-value, given the Bonferroni adjusted p-value $\leq \alpha$, is added to the model otherwise the model selection process terminates. Say x_1 is added to the model during the first step. The next step has 12 candidate terms x_2, x_3, x_4, x_5, x_6, x_7, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_1x_6, x_1x_7, and the p-value of the partial F-test for each candidate term is multiplied by 12 for the Bonferroni adjustment. If the candidate term with the minimum Bonferroni adjusted p-value, given the Bonferroni adjusted p-value $\leq \alpha$, then the term is added to the model. The process continues until no candidate term has a Bonferroni adjusted p-value $\leq \alpha$ or all terms have been added to the model. Forward selection methods without the Bonferroni adjustment are referred to in our study as "Unadjusted".

4.4.5 Additional Lack of Fit Test

We propose adding a LOF test after each step in a model selection process as an early stopping criterion to help inhibit FDR. When a pure error estimate is available, the additional LOF method incorporates a LOF test after each term is entered into the model during model selection. As described in Khuri and Cornell (1996), a model lacking in fit indicates the fitted model lacks a sufficient number of terms, either missing factors or higher order terms. The residual variation in sums of square error excluding the estimate of pure error is the variation due to LOF. Therefore, sum of squares LOF SS_{LOF} is

$$SS_{LOF} = SSE - SSPE \tag{4.21}$$

with SSE from Equation 4.6 and SSPE from Equation 4.11. Similarly, the degrees of freedom for LOF is found by subtracting the degrees of freedom pure error from the degrees of freedom SSE, $df_{LOF} = (n - p_1) - (n - c) = c - p_1$ where n is the number of runs, p_1 is the number of terms in fitted model and c is the number of unique runs.

The following F-ratio tests the null hypothesis that the fitted model is adequate

$$F_{LOF} = \frac{SS_{LOF}/df_{LOF}}{SSPE/df_{pe}}. \tag{4.22}$$

with a p-value LOF, p_{LOF}, equal to Prob$(F_{df_{LOF}, df_{pe}} \geq F_{LOF})$. Depending on the level of significance α, if the $p_{LOF} \leq \alpha$, then we reject the null hypothesis; there is evidence of LOF in the fitted model and we continue the model selection process. When a LOF test determines that the model is adequate having a $p_{LOF} > \alpha$, there is no more evidence of LOF and we do not add any additional terms.

We add a LOF test if a pure error estimate is available. The following example details the process for forward selection with weak heredity. The p-value type depends on the specified model selection method, our Unadjusted PE LOF method uses p_{pe}, our Unadjusted SSE LOF method uses p_{sse}, and the Bonferroni PE LOF uses Bonferroni adjusted p_{pe}. Given 7 factors, the first round of model selection includes candidate terms x_1, x_2, x_3, x_4, x_5, x_6, x_7, all 7 main effect terms. The candidate term with the minimum p-value, given the p-value $\leq \alpha$, is added to the model, say this is x_1. A LOF test is done on the model including the intercept term and x_1. If $p_{LOF} > \alpha$ the procedure stops. Suppose in the first step $p_{LOF} \leq \alpha$, we continue to the next step with candidate terms x_2, x_3, x_4, x_5, x_6, x_7, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_1x_6, x_1x_7, all remaining main effect terms and all 2FI including x_1. Let x_3 have the minimum p-value out of the candidate terms with a value $\leq \alpha$; x_3 is added to the model. A LOF test is done on the model including the intercept term, x_1 and x_3. Suppose $p_{LOF} > \alpha$, then there is no more evidence of LOF and we stop adding terms to the model. The final model contains x_1 and x_3.

Our AICc LOF method incorporates the use of pure error through an additional LOF test after each step in stepwise forward selection with the AICc criteria. During each forward selection step, a candidate term is selected to be included in the model if the addition of the term results in the model with minimum AICc. If the addition of any candidate term to the model does not lower AICc the process terminates. After each forward selection step, a LOF test is added to determine if the selected model is adequate. If the LOF p-value is greater than the α significance level, then there is no longer evidence of LOF and the model selection process terminates if it has not already terminated by the model selection procedure.

4.4.6 LASSO

The least absolute shrinkage and selection operator (LASSO) is a model selection method introduced by Tibshirani (1996). It is a regularization technique often used to analyze small screening designs such as Definitive Screening Designs and supersaturated designs (Smucker et al., 2020). LASSO shrinks coefficient estimates setting some to zero by minimizing

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=0}^{h}(\lambda|\beta_j|) \tag{4.23}$$

where λ is a tuning parameter (Zhang et al., 2019). The tuning parameter λ is often selected using cross-validation, though Smucker et al. (2020) recommend using AICc for small structured designs. For our LASSO AICc model selection method, we select the value for λ that minimizes $AICc = AIC + \frac{2p(p+1)}{n-p-1}$ where

$$AIC = n \log \frac{(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L^{\lambda})'(Y - \boldsymbol{X}\boldsymbol{\beta}_L^{\lambda})}{n} + 2p \tag{4.24}$$

and p is the number of parameters, $p < n - 1$, and $\boldsymbol{\beta}_L^{\lambda}$ are the LASSO estimates for value λ.

We propose a new model selection method, LASSO LOF, that incorporates inference with pure error with the LASSO technique. This method computes lack of fit tests moving along the LASSO solution path. We first compute the LASSO regularization path for the regularization parameter λ covering the entire range of possible solutions. Starting with the largest λ value corresponding to the null model, a LOF test in Equation 4.22 is done to assess if the selected model is adequate. IF $p_{LOF} \leq \alpha$ there is evidence of lack of fit. The process continues for the next largest λ and continues assessing candidate models until the p_{LOF} is greater than the α significance level. There is no longer evidence of LOF and the model selection process terminates selecting the model corresponding to that λ value.

4.5 Simulation Study

This section describes the simulation study analyzing model selection with pure error. We first detail the analysis approach utilized followed by an outline of the simulation protocol.

4.5.1 Analysis Approach

We consider the following:

1. Number of factors: 7, 11
2. Total runs: 16, 24, 32, 40
3. Run type: Added or Total Runs

4. Number of Replicate Points: 0, 2, 4, 6
5. Active Main Effects: 2, 3, 4, 5
6. Active Two Factor Interactions: 0, 1, 3, 5, 7
7. Alpha Significance Level: 0.01, 0.05, 0.10, 0.25, 0.50
8. Model Selection Methods with Pure Error:
   (a) Unadjusted PE: Forward selection with p_{pe}
   (b) Unadjusted PE LOF: Forward selection with p_{pe} and an additional LOF test
   (c) Bonferroni PE: Forward selection using Bonferroni adjusted p_{pe}
   (d) Bonferroni PE LOF: Forward selection using Bonferroni adjusted p_{pe} and an additional LOF test
   (e) AICc LOF: Forward selection using AICc with an additional LOF test
   (f) Unadjusted SSE LOF: Forward selection using unadjusted p_{sse} and an additional LOF test
   (g) LASSO LOF: LASSO model selection using LOF as model selection criteria.
9. Model Selection Methods with SSE:
   (a) Unadjusted SSE: Forward selection using unadjusted p_{sse}
   (b) AICc: Forward selection with AICc
   (c) Bonferroni SSE: Forward selection using Bonferroni adjusted p_{sse}
   (d) LASSO AICc: LASSO model selection using AICc criteria

All designs included in our simulation study are Bayesian D-optimal designs. We create all our initial designs in JMP (JMP® Pro 14, SAS Institute Inc., Cary, NC) employing the Bayesian framework for either 7 or 11 factors. The framework leverages prior information to allow for precise estimation of terms classified as primary, and the estimation of some of the terms labeled potential when allowable (SAS Institute Inc., 2019). We specify main effect terms as primary, and 2FI as potential. For the "added runs" run type, replicate points are selected randomly and are added in addition to the total run size. For instance given a total run size of 16 and two replicate points, a design with "added runs" is a 16 run D-optimal design with two additional replicate runs resulting in 18 runs. For the "Total Runs" run type, D-optimal designs with partial replication are created in JMP for the total run size. Given the 16 total run example, this design type will have 16 runs regardless of the number of replicate points added. Model selection methods without pure error use designs with 0 replicate points. Model selection methods with pure error use designs with 2, 4, or 6 replicate points. The same significance level α specified by the simulation scenario is used in model selection criteria as well as in any additional LOF tests.

4.5.2 Simulation Protocol

Our simulation study varies the following to compare power and FDR performance across various model selection methods with pure error and SSE: i) number of factors, ii) number active main effects and 2FI, iii) total run size, iv) design run type, v) number of replicate points, vi) α significance level, and vii) 11 different model selection methods. All designs have input factors constrained between -1 and 1. We investigate model selection with pure error and SSE across forward selection and penalized regression methods (LASSO). For forward selection, we evaluate the usage of p_{pe}, p_{sse}, and AICc in the selection criteria. Additionally, we consider unadjusted p-values as well as Bonferroni adjusted p-values. For LASSO, we compare the AICc criterion to our new approach incorporating pure error by using LOF tests to select the model. Lastly, we try integrating an additional LOF test to selection methods in an effort to curb FDR. Table 13 summarizes the simulation variables and their levels. There are a total of 42,000 distinct simulation scenarios.

Each iteration begins by generating response values from input values specified by the design matrix. Model terms are randomly selected to be active depending on the specified number of active terms. Weak effect heredity is assumed when selecting 2FI, that is 2FI are present only if at least one parent main effects are active (Ockuly

Table 13.: Model selection summary of simulation variables

Simulation Variable	Levels
Number of Factors	7, 11
Active Main Effects	2, 3, 4, 5
Active Two-Factor Interactions[*]	0, 1, 3, 5, 7
Total Run Size[**]	16, 24, 32, 40
Design Run Type	Added runs or Total runs
Number of Replicate Points	0, 2, 4, 6
α significance levels	0.01, 0.05, 0.10, 0.25, 0.50
Model Selection Method incorporating Pure Error	AICc LOF, Bonferroni PE, Bonferroni PE LOF, LASSO LOF, Unadjusted PE, Unadjusted PE LOF, Unadjusted SSE LOF
Model Selection Method without Pure Error	AICc, Bonferroni SSE, LASSO AICc, Unadjusted SSE

[*] Weak effect heredity enforced
[**] Design run type of added runs has replicate points added to a D-optimal design of Total Run Size

et al., 2017). The true model coefficients for the active terms are randomly assigned from the set $\{\pm 1, \pm 2, \pm 3\}$. The intercept, β_0, is fixed at 10. We use the following to generate responses,

$$y = 10 + \sum_{j=1}^{h} \beta_j x_j + \sum \sum_{i<j=2}^{h} \beta_{ij} x_i x_j + \epsilon \tag{4.25}$$

with $\epsilon \sim Normal(0, \sigma^2)$, $\sigma^2 = 1$, and h is number of factors. We apply a model selection method and all terms in the selected model are identified as active.

For each unique combination of attributes, 1,000 iterations of the simulation are run, and the overall mean power and FDR are compared to evaluate design performance. The entire simulation process is outlined in Algorithm 4.

Algorithm 4 Model selection with pure error simulation steps

```
1: Assuming weak effect heredity for 2FI, randomly assign model terms to be active based on the number of active main effect terms and 2FI.
2: Create the underlying relationship between inputs and responses by assigning model coefficients from the set $\{\pm 1, \pm 2, \pm 3\}$ to the active terms. Inactive terms are assigned a value of zero. Note some scenarios do not have any active 2FI as part of our simulation test cases.
3: Generate a response for each run in the experimental design using Equation 4.25 and $\sigma^2 = 1$.
4: Apply model selection method using specified $\alpha$ level, experimental design data as regressors, and the generated response as the predictor. All terms in the selected model are identified as active.
5: Calculate power and FDR.
6: Repeat all previous steps 1,000 times and take the mean power and mean FDR per scenario.
```

4.6 Simulation Results

Figure 14 plots each model selection method's mean power and FDR by number of factors, error type and α significance level. As α levels increase, we see a rise in power as well as FDR for all methods except for AICc and LASSO AICc. Note that AICc and LASSO AICc do not use an α significance level during model selection and thus show a flat line. AICc and LASSO AICc are the two methods using AICc without an additional LOF test and have high mean power rates but also some of the highest FDR rates. Unadjusted PE achieves a similar mean power to the LASSO AICc method when $\alpha \geq 0.05$ and lower mean FDR rates when $\alpha < 0.25$. Unadjusted PE and Unadjusted SSE have the highest mean FDR rates at $\alpha = 0.5$. On the other hand, Bonferroni SSE has the lowest mean FDR but also the lowest mean power across all α levels. The gap between Unadjusted PE and Unadjusted PE LOF mean power narrows while the difference in mean FDR widens as α levels increase, providing support for the use of a LOF test when p-values are unadjusted.

We fit two ANOVA models including simulation scenario attributes as independent variables and each with either power or FDR as the dependent variable. The

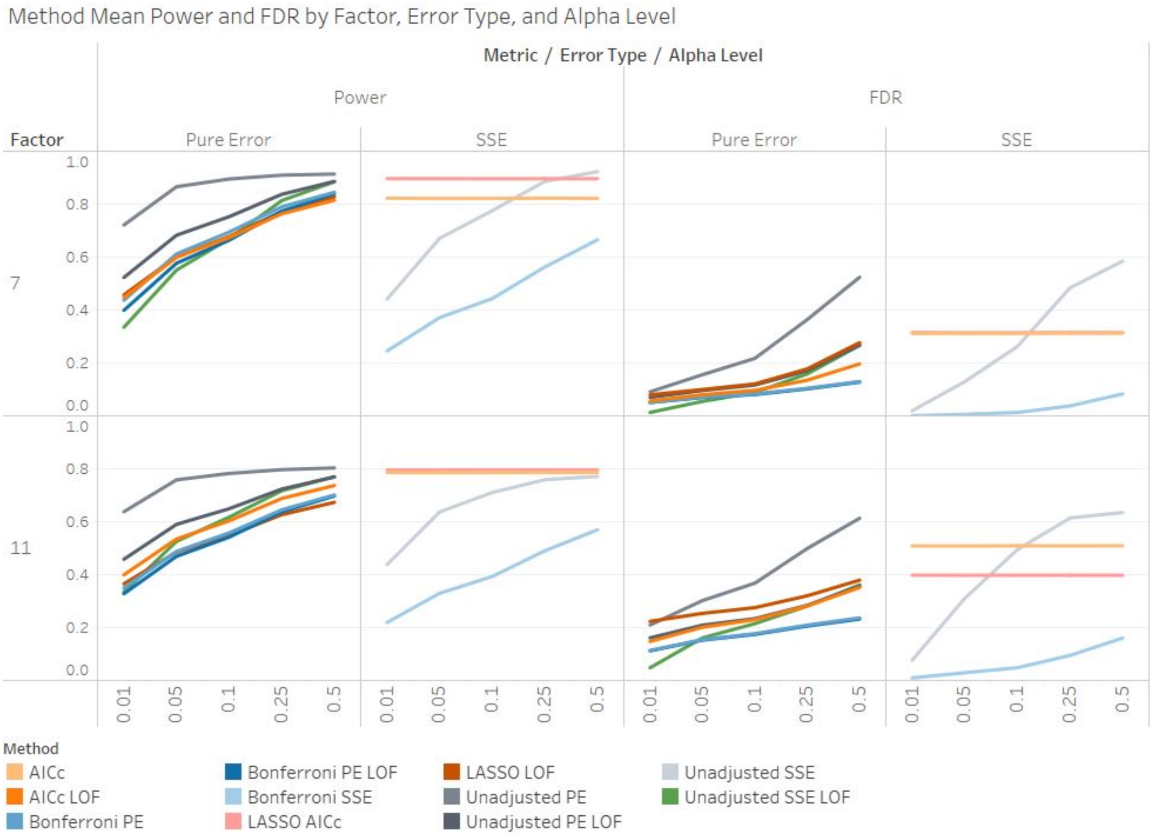


Fig. 14.: Method mean power and FDR by number of factors, error type and α level. Note AICc and LASSO AICc do not use an α level during model selection.

ANOVA models indicate clear differences by design, and the results (Tables 21 and 22) are included in the Appendix. To highlight the ANOVA model results, we also provide plots of the largest F-values in Figure 22 for the power model and Figure 23 for the FDR model in the Appendix. We use these charts to guide the selection of variables to include in the plots visualizing our simulation results.

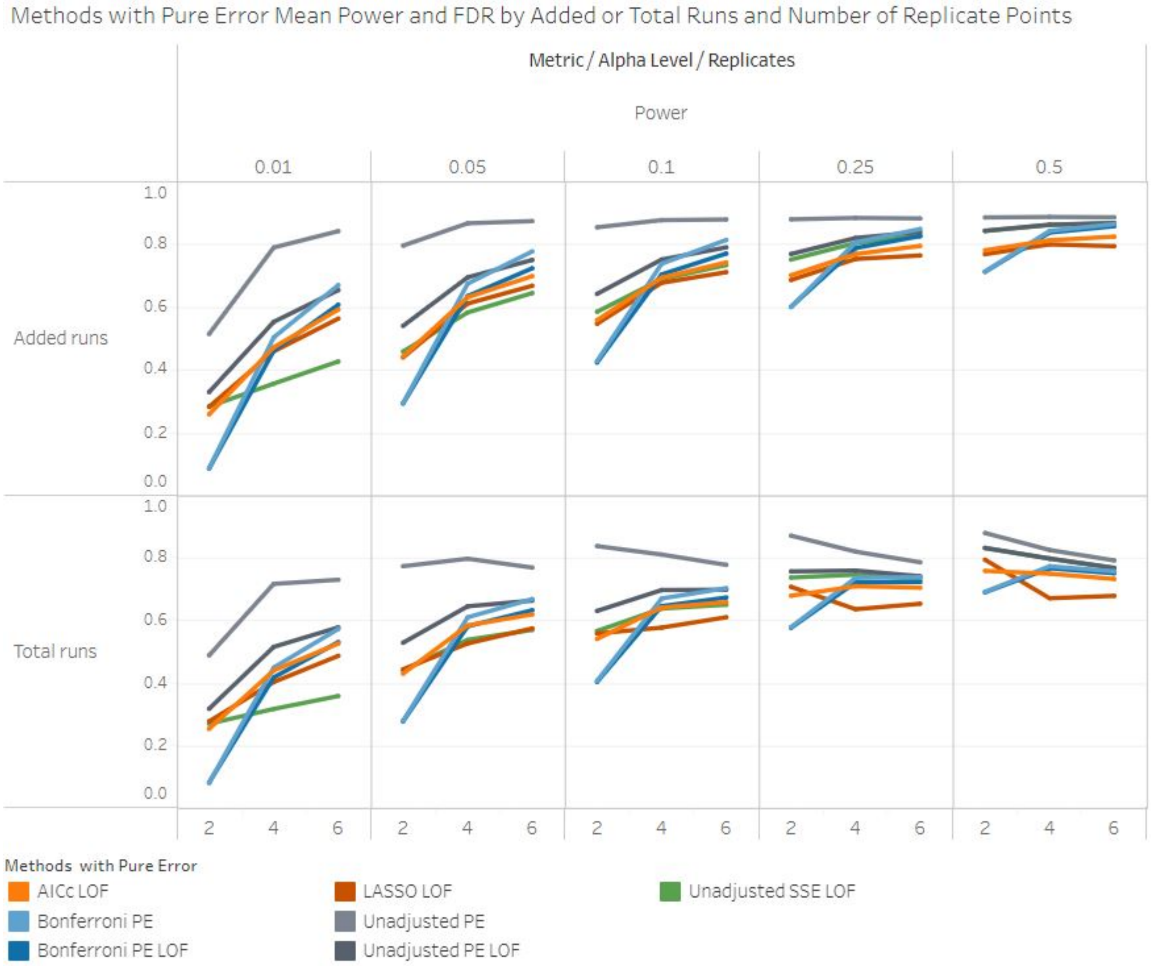


Fig. 15.: Methods with pure error mean power by total or added runs, α level and number of replicate points.

Figure 15 plots mean power by attributes identified in Figure 23(a) in the Appendix, specifically run type, α level and number of replicate points. The "Added Runs" run type considers replicate runs added in addition to a D-optimal design.

Here power increases with the number of replicate runs for all methods except LASSO LOF, which shows a slight decrease in mean power from 4 to 6 replicate runs. However, for the “Total Runs” run type, when the experimenter has a fixed set of total runs to allocate towards an experiment, all methods increase in mean power with the number of replicates just at $\alpha = 0.01$. Unadjusted PE has the highest mean power of model selection methods with pure error, and with the “Total Runs” run type, starts to decrease in mean power from 4 to 6 runs at $\alpha = 0.05$. As α levels increase, model selection methods start to decrease in power from 4 to 6 replicates and by $\alpha = 0.5$ most methods show a decrease in power and the number of replicate runs increase from 2. Bonferroni PE and Bonferroni PE LOF show a leveling off instead of a clear decrease from 4 to 6 runs. This illustrates the trade-off between using replicate points to estimate variance or allocating the degrees of freedom to estimate effects when faced with a fixed run size.

Figure 16 plots mean FDR by attributes identified in the top 15 F-values in the ANOVA FDR model, Figure 22(b) in the Appendix, specifically number of factors, method and α level. Moving from 7 to 11 factors increases mean FDR for all model selection methods. Unadjusted SSE has one of the lowest mean FDR at $\alpha = 0.01$ but at $\alpha = 0.5$ has the highest mean FDR out of the methods. Adding a LOF fit to Unadjusted SSE (Unadjusted SSE LOF) dampens mean FDR, keeping it below 0.37 across α levels. Even more effective at minimizing mean FDR is adding a Bonferroni adjustment, Bonferroni SSE, manages to keep mean FDR below 0.17 throughout all α levels. A similar pattern is seen comparing the performance of Unadjusted PE, Unadjusted PE LOF and Bonferroni PE. Unadjusted PE LOF manages to taper mean FDR compared to Unadjusted PE, and Bonferroni PE has an even lower mean FDR across α levels. Diminishing returns are seen when adding an additional LOF test to Bonferroni PE that already has low FDR values, resulting in only a slight

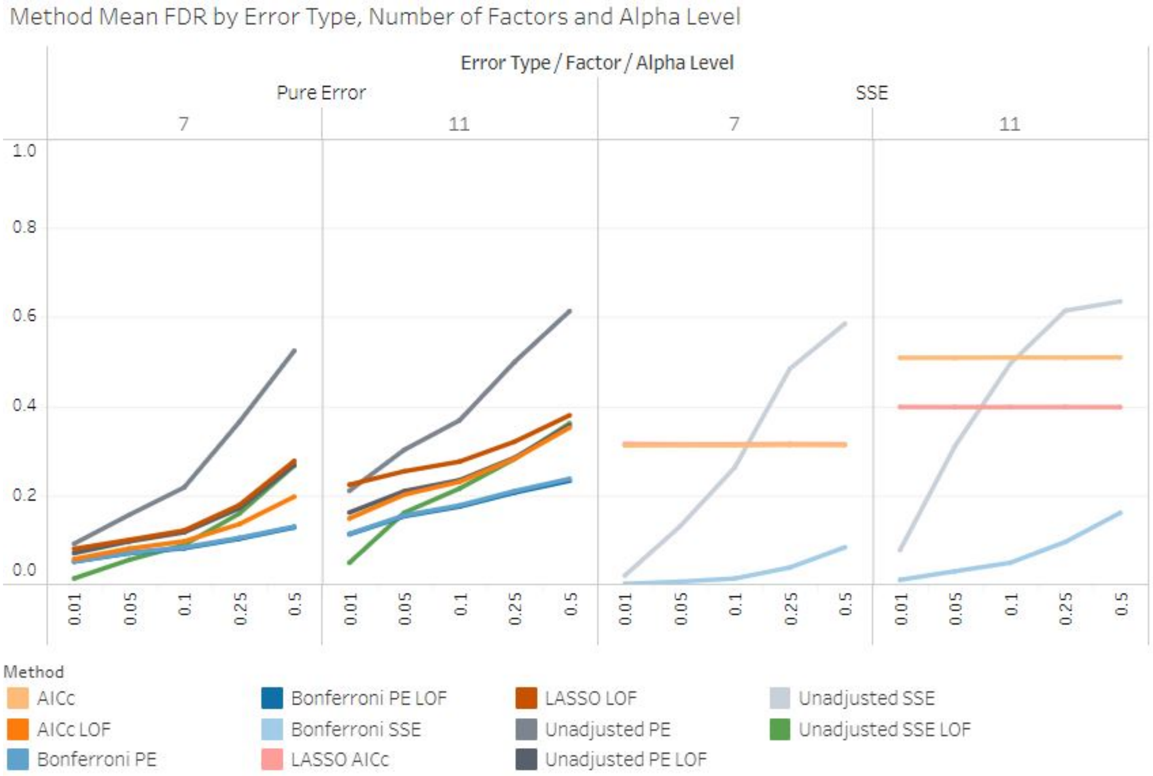


Fig. 16.: Method mean FDR by method (pure error vs SSE), the number of factors and α level. Alpha level is on x-axis and mean FDR on y-axis. Note AICc and LASSO AICc do not use an α level during model selection.

decrease to mean FDR. For instance at $\alpha = 0.5$, 11 factors has a mean FDR with Bonferroni PE of 0.237 compared to 0.232 with Bonferroni PE LOF, and 7 factors has a mean FDR with Bonferroni PE of 0.130 compared to 0.128 with Bonferroni PE LOF.

To summarize overall performance, a ranking of methods across all simulation scenarios is provided in Figure 17(a) for mean power and in Figure 17(b) for mean FDR. An increase to power when incorporating pure error is seen comparing Unadjusted SSE and Bonferroni SSE to the same method with pure error, Unadjusted PE and Bonferroni PE. Bonferroni SSE maintains lower mean FDR compared to Bonferroni PE, while Unadjusted SSE has lower mean FDR than Unadjusted PE at $\alpha < 0.25$. Adding a LOF test to Unadjusted PE, AICc and Bonferroni PE decreases mean FDR, however, mean power also declines.

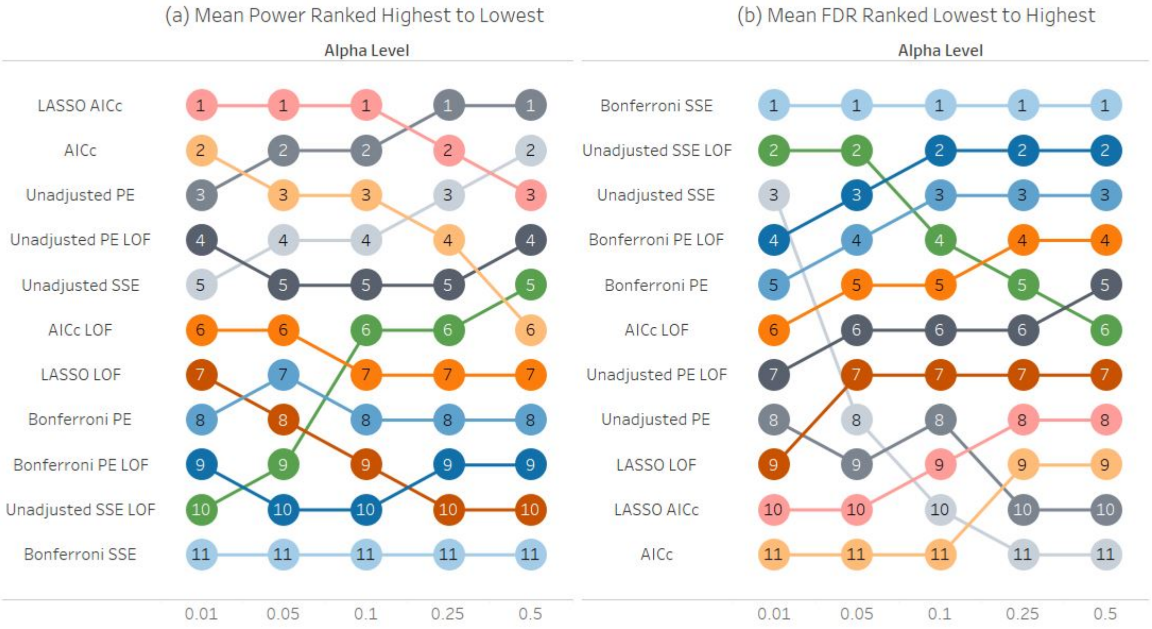


Fig. 17.: Ranking of model selection methods by mean power and mean FDR across all simulation scenarios

Unadjusted PE at $\alpha = 0.05$ has the highest mean power compared to other methods; however, this combination also has the second highest mean FDR rate compared to the other designs. We seek a method that balances high power with low FDR rates and we administer Pareto front analysis as well as Desirability analysis to help distinguish the optimal model selection method. A Pareto front visualizes trade-offs from competing objectives and is used to identify optimal solutions. We apply the Utopia point method described in Lu et al. (2011) to select a recommended model selection method that balances high power with low FDR. The Pareto fronts in Figure 18 plot mean power and 1-mean FDR by α significance level. Overall power and FDR performance is shown in Figure 18(a), main effect power and FDR in Figure 18(b) and 2FI power and FDR in Figure 18(c). Methods that dominate other designs by either high mean power or low mean FDR are shown on the Pareto frontier, marked by a black line. Methods utilizing pure error and SSE are seen along the Pareto frontier.

Overall performance in Figure 18(a) reveals three methods, AICc, AICc LOF, and LASSO LOF, are not located on the Pareto front for any α level. Bonferroni PE, Bonferroni PE LOF, and Unadjusted PE LOF are dominated by other points when $\alpha = 0.01$, but appear on the Pareto front for all other α levels. There is a slight reduction in mean FDR as well as in mean power with an additional LOF test with Bonferroni PE. The additional LOF fit test is seen to improve Pareto performance with Unadjusted SSE LOF appearing on the Pareto front for all levels of α; while Unadjusted SSE is dominated by other points for all α levels except 0.01. While Unadjusted PE, and Unadjusted PE LOF when $\alpha > 0.01$, are both located on the Pareto front, a benefit to adding LOF tests to Unadjusted PE is seen when $\alpha \geq 0.25$. At the larger α levels, adding LOF to Unadjusted PE reduces mean FDR by about half while mean power decreases by less than 10%.

For main effect power and FDR, all methods appear on the Pareto front for

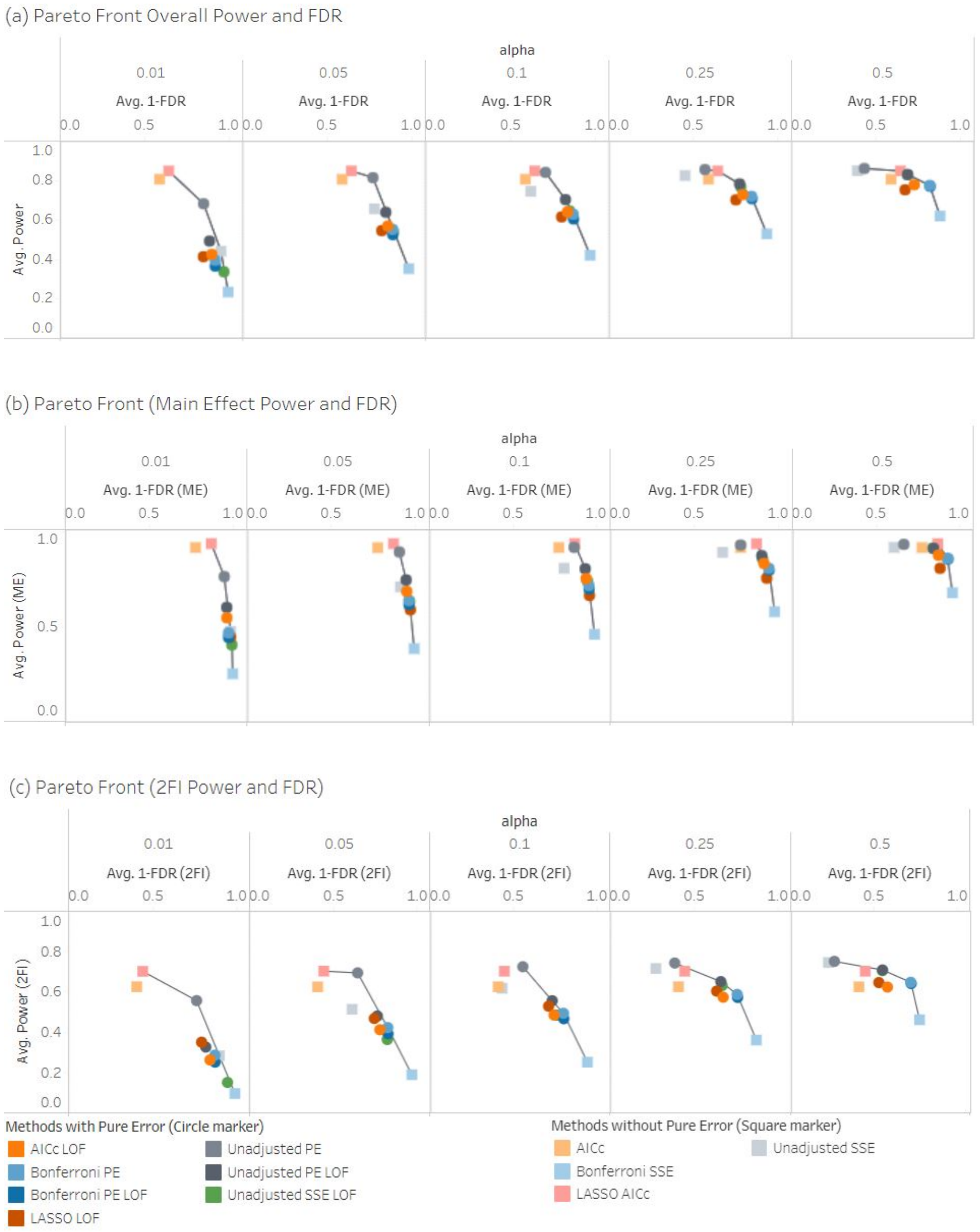
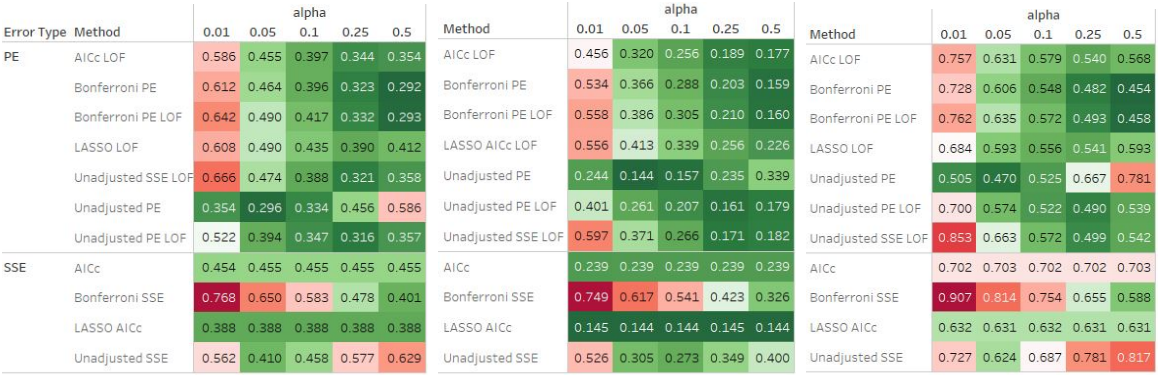


Fig. 18.: Pareto front of methods with 1 - mean FDR on x-axis and mean power on y-axis.

$\alpha = 0.01$ except for AICc (Figure 18(b)). As α increases to 0.5, only LASSO AICc, Bonferroni PE, Bonferroni PE LOF, and Bonferroni SSE remain on the Pareto front. For 2FI power and FDR, methods Unadjusted PE, Unadjusted PE LOF, Bonferroni PE, Bonferroni PE LOF, and Bonferroni SSE remain on the front across all levels of α (Figure 18(c)).

The Pareto frontier reveals a subset of methods that dominate others in terms of mean power or mean FDR but will also include methods at the extremes, such as Bonferroni SSE having the lowest mean FDR but also the lowest mean power. Someone prioritizing low FDR over high power may be interested in selecting this method; the Pareto frontier allows the viewer to choose a method depending on their preference between high power and low FDR. In terms of selecting the top method, for this study we calculate the Euclidean distance from the Utopia point (1,1) to rank how well the methods attain high mean power and low mean FDR rates across the various levels of α. The Utopia point represents the ideal goal of having 100% power and 0% FDR. The Euclidean distance from the Utopia point of (1,1) is provided in Figure 19 with Figure 19(a) for overall, Figure 19(b) for main effects and Figure 19(c) for 2FI. The color scale in Figure 19 highlights preferred smaller Euclidean distances from the Utopia point in green and distances further away from the Utopia point in red. More green shading is seen for the methods incorporating pure error. Overall, Bonferroni PE with $\alpha = 0.5$ has the lowest Euclidean distance from the Utopia point (0.292) with Bonferroni LOF following close behind (0.293). For other α levels, Unadjusted PE has the lowest distance at $\alpha <= 0.1$. Unadjusted PE LOF has the lowest distance at $\alpha = 0.25$ (0.316) showing the benefit of adding a LOF test to unadjusted p-values at higher α levels. Of the methods listed under error type SSE, LASSO AICc has the smallest distance of 0.388.

Regarding main effect power and FDR, Figure 19(b) shows LASSO AICc and



Error Type	Method	alpha 0.01	0.05	0.1	0.25	0.5
PE	AICc LOF	0.586	0.455	0.397	0.344	0.354
	Bonferroni PE	0.612	0.464	0.396	0.323	0.292
	Bonferroni PE LOF	0.642	0.490	0.417	0.332	0.293
	LASSO LOF	0.608	0.490	0.435	0.390	0.412
	Unadjusted SSE LOF	0.666	0.474	0.388	0.321	0.358
	Unadjusted PE	0.354	0.296	0.334	0.456	0.586
	Unadjusted PE LOF	0.522	0.394	0.347	0.316	0.357
SSE	AICc	0.454	0.455	0.455	0.455	0.455
	Bonferroni SSE	0.768	0.650	0.583	0.478	0.401
	LASSO AICc	0.388	0.388	0.388	0.388	0.388
	Unadjusted SSE	0.562	0.410	0.458	0.577	0.629

(a) Overall

Method	alpha 0.01	0.05	0.1	0.25	0.5
AICc LOF	0.456	0.320	0.256	0.189	0.177
Bonferroni PE	0.534	0.366	0.288	0.203	0.159
Bonferroni PE LOF	0.558	0.386	0.305	0.210	0.160
LASSO AICc LOF	0.556	0.413	0.339	0.256	0.226
Unadjusted PE	0.244	0.144	0.157	0.235	0.339
Unadjusted PE LOF	0.401	0.261	0.207	0.161	0.179
Unadjusted SSE LOF	0.597	0.371	0.266	0.171	0.182
AICc	0.239	0.239	0.239	0.239	0.239
Bonferroni SSE	0.749	0.617	0.541	0.423	0.326
LASSO AICc	0.145	0.144	0.144	0.145	0.144
Unadjusted SSE	0.526	0.305	0.273	0.349	0.400

(b) Main Effects

Method	alpha 0.01	0.05	0.1	0.25	0.5
AICc LOF	0.757	0.631	0.579	0.540	0.568
Bonferroni PE	0.728	0.606	0.548	0.482	0.454
Bonferroni PE LOF	0.762	0.635	0.572	0.493	0.458
LASSO LOF	0.684	0.593	0.556	0.541	0.593
Unadjusted PE	0.505	0.470	0.525	0.667	0.781
Unadjusted PE LOF	0.700	0.574	0.522	0.490	0.539
Unadjusted SSE LOF	0.853	0.663	0.572	0.499	0.542
AICc	0.702	0.703	0.702	0.702	0.703
Bonferroni SSE	0.907	0.814	0.754	0.655	0.588
LASSO AICc	0.632	0.631	0.632	0.631	0.631
Unadjusted SSE	0.727	0.624	0.687	0.781	0.817

(c) 2FI

Fig. 19.: Pareto Front Euclidean distance from Utopia point (1,1). Color coded to show preferred lower values in green and larger values in red.

Unadjusted PE at $\alpha = 0.05$ with the lowest Euclidean distance of 0.144. In terms of 2FI power and FDR, Figure 19(c) reveals Bonferroni PE at $\alpha = 0.5$ has the lowest Euclidean distance from the Utopia point (0.454). Bonferroni PE LOF also has favorable Euclidean distances at $\alpha = 0.5$ (0.458), and Unadjusted PE also performs well at $\alpha = 0.05$ (0.47). LASSO AICc did well for the main effects comparison but does not perform as strong for 2FI with an Euclidean distance of 0.63.

4.6.1 Desirability Analysis

As an alternative to the Pareto front analysis, we utilize multiple response optimization with desirability functions to provide an overall model selection method recommendation that balances high power and low FDR. Upper and lower thresholds representing the acceptable range are specified for each response. Each objective is translated into an individual desirability function on a scale ranging from 0 to 1, with 1 indicating the objective is at the optimal target level (Myers et al., 2016). Let $f_r(\boldsymbol{X})$ for $r = \{1, ..., R\}$, represent a function to optimize. The desirability function to maximize $f_r(\boldsymbol{X})$ is

$$d_r^{max} = \begin{cases} 0, & f_r(\boldsymbol{X}) < L_r \\ \left(\frac{f_r(\boldsymbol{X})-L_r}{T_r-L_r}\right)^w, & L_r \leq f_r(\boldsymbol{X}) \leq T_r \\ 1, & f_r(\boldsymbol{X}) > T_r \end{cases}$$

where L_r and T_r are the lower and top user specified thresholds, and weight $w = 1$ is linear desirability function. The desirability function to minimize $f_r(\boldsymbol{X})$ is

$$d_r^{min} = \begin{cases} 1, & f_r(\boldsymbol{X}) < L_r \\ \left(\frac{T_r-f_r(\boldsymbol{X})}{T_r-L_r}\right)^w, & L_r \leq f_r(\boldsymbol{X}) \leq T_r \\ 0, & f_r(\boldsymbol{X}) > T_r \end{cases}$$

where L_r is the optimal target level and T_r is a larger but still acceptable level. The functions are combined using the geometric mean

$$D = \left(\prod_{r=1}^{R} d_r\right)^{1/R} \tag{4.26}$$

to produce an overall desirability score and the preferred design has the highest mean desirability score. Our goal is to find the model selection method that maximizes predicted power and minimizes predicted FDR. The main effect with 2FI regression model is used for each function $f_r(\boldsymbol{X})$, $r = 1, 2$. Targets are established with the desire to have at least above average power and below average FDR. For power, the desirability function maximizes $f_1(\boldsymbol{X})$ where target L_1 is the mean predicted power (0.657) and $T_1 = 1$. For FDR, the desirability function minimizes $f_2(\boldsymbol{X})$ where $L_2 = 0$ and T_2 is the mean predicted FDR (0.210). We generate the overall desirability score in R (R Core Team, 2018) using the package `desirability` and the `dMax,dMin,dOverall` functions (Kuhn, 2016).

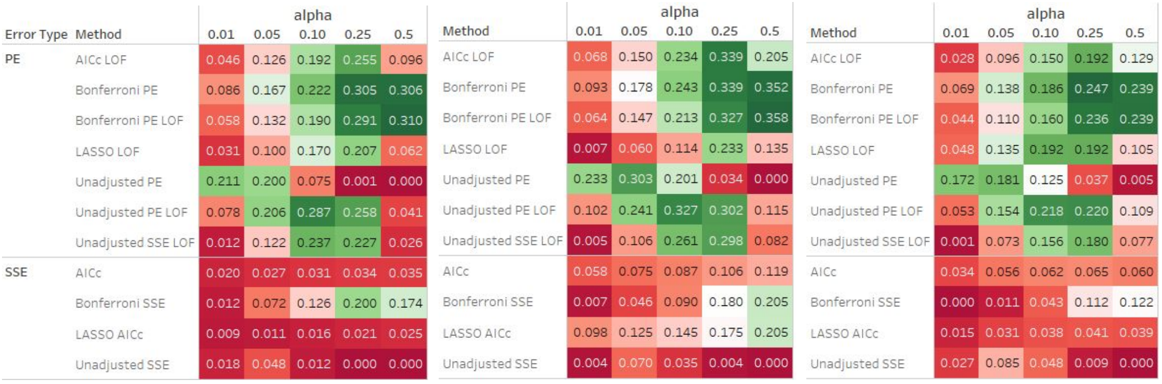
Table 14 shows the mean desirability score for each method as well as the mean

Table 14.: Method desirability analysis results (mean values)

Design	Predicted Power	Predicted FDR	Desirability Power	Desirability FDR	Overall Desirability
Bonferroni PE	0.612	0.133	0.232	0.485	0.217
Bonferroni PE LOF	0.593	0.131	0.200	0.491	0.196
Unadjusted PE LOF	0.687	0.197	0.292	0.312	0.174
AICc LOF	0.626	0.178	0.203	0.358	0.143
Unadjusted SSE LOF	0.622	0.165	0.254	0.387	0.125
Bonferroni SSE	0.429	0.048	0.129	0.726	0.117
LASSO LOF	0.603	0.221	0.172	0.281	0.114
Unadjusted PE	0.808	0.334	0.500	0.112	0.098
AICc	0.804	0.411	0.513	0.056	0.029
LASSO AICc	0.846	0.356	0.581	0.013	0.017
Unadjusted SSE	0.701	0.361	0.372	0.175	0.015

predicted power, mean predicted FDR, mean desirability score for power, and mean desirability score for FDR. The method with the highest mean desirability score is Bonferroni PE followed by Bonferroni PE LOF. The high score for Bonferroni PE and Bonferroni PE LOF is driven by a high FDR desirability score and a moderate desirability score for power. These two methods are also identified in our Pareto front analysis as having the minimum distance from the Utopia point. Unadjusted PE LOF has the third highest desirability score supporting the addition of a LOF test when using unadjusted p_{pe}. Mean desirability scores for main effect power and FDR, and 2FI power and FDR are provided in Appendix Tables 23 and 24. Bonferroni PE has the highest desirability score for main effect power and FDR, as well as the highest score for 2FI power and FDR.

Figure 20 visualizes how the methods performed in terms of mean desirability across α levels. The higher mean desirability scores are found with pure error. The color scale in Figure 20 highlights preferred larger desirability values in green and smaller desirability scores in red. There is noticeably less green shading for the



Error Type	Method	alpha 0.01	0.05	0.10	0.25	0.5
PE	AICc LOF	0.046	0.126	0.192	0.255	0.096
	Bonferroni PE	0.086	0.167	0.222	0.305	0.306
	Bonferroni PE LOF	0.058	0.132	0.190	0.291	0.310
	LASSO LOF	0.031	0.100	0.170	0.207	0.062
	Unadjusted PE	0.211	0.200	0.075	0.001	0.000
	Unadjusted PE LOF	0.078	0.206	0.287	0.258	0.041
	Unadjusted SSE LOF	0.012	0.122	0.237	0.227	0.026
SSE	AICc	0.020	0.027	0.031	0.034	0.035
	Bonferroni SSE	0.012	0.072	0.126	0.200	0.174
	LASSO AICc	0.009	0.011	0.016	0.021	0.025
	Unadjusted SSE	0.018	0.048	0.012	0.000	0.000

(a) Overall

Method	alpha 0.01	0.05	0.10	0.25	0.5
AICc LOF	0.068	0.150	0.234	0.339	0.205
Bonferroni PE	0.093	0.178	0.243	0.339	0.352
Bonferroni PE LOF	0.064	0.147	0.213	0.327	0.358
LASSO LOF	0.007	0.060	0.114	0.233	0.135
Unadjusted PE	0.233	0.303	0.201	0.034	0.000
Unadjusted PE LOF	0.102	0.241	0.327	0.302	0.115
Unadjusted SSE LOF	0.005	0.106	0.261	0.298	0.082
AICc	0.058	0.075	0.087	0.106	0.119
Bonferroni SSE	0.007	0.046	0.090	0.180	0.205
LASSO AICc	0.098	0.125	0.145	0.175	0.205
Unadjusted SSE	0.004	0.070	0.035	0.004	0.000

(b) Main Effects

Method	alpha 0.01	0.05	0.10	0.25	0.5
AICc LOF	0.028	0.096	0.150	0.192	0.129
Bonferroni PE	0.069	0.138	0.186	0.247	0.239
Bonferroni PE LOF	0.044	0.110	0.160	0.236	0.239
LASSO LOF	0.048	0.135	0.192	0.192	0.105
Unadjusted PE	0.172	0.181	0.125	0.037	0.005
Unadjusted PE LOF	0.053	0.154	0.218	0.220	0.109
Unadjusted SSE LOF	0.001	0.073	0.156	0.180	0.077
AICc	0.034	0.056	0.062	0.065	0.060
Bonferroni SSE	0.000	0.011	0.043	0.112	0.122
LASSO AICc	0.015	0.031	0.038	0.041	0.039
Unadjusted SSE	0.027	0.085	0.048	0.009	0.000

(c) 2FI

Fig. 20.: Method mean desirability score by α level. Color coded to show preferred higher values in green and smaller values in red.

methods listed under error type SSE compared to Figure 19 in the Pareto front analysis. Setting the lower Desirability threshold for power to mean predicted power and the higher Desirability threshold for FDR to mean predicted FDR penalizes methods at the extremes, methods either having high power and higher FDR as well as methods with low mean power and low mean FDR. Setting the lower Desirability threshold for power to 0 and the higher Desirability threshold for FDR to 1 results in a ranking of Desirability scores similar to the Pareto front. Bonferroni PE LOF maintains the highest overall mean Desirability score, followed by Bonferroni PE at $\alpha = 0.5$ regardless of using thresholds to only consider values above the mean or setting them to include all values.

4.7 Summary and Conclusions

Previous studies have shown higher power rates when incorporating pure error into model selection procedures with the trade-off of high Type I error and FDR. As noted in Westfall et al. (1998), pure error fixes the denominator of the partial F-statistic to MSPE and the numerator's noncentrality is no longer counter balanced.

To counteract the lack of noncentrality in the partial F-test denominator contributing to larger partial F-tests with pure error, we consider early stopping methods including Bonferroni adjusted p-values and our proposed forward selection method that incorporates a LOF test after each model selection step. We examine various model selection techniques to propose a strategy for incorporating pure error in model selection procedures that keeps FDR in check. Furthermore, we develop a method of incorporating pure error with LASSO penalized regression by using LOF tests to select the model along the LASSO solution path.

Our simulation study varies the number of input factors, the number of active terms, the α significance level, the number of total runs, the number of replicate points, whether replicate points were added in addition to the total runs or included in the total, and model selection method. Model selection methods are evaluated based on mean power and FDR. Our simulation results support the use of forward selection with Bonferroni adjusted p_{pe} with an α significance level of 0.5. We also recommend forward selection with Bonferroni adjusted p_{pe} and an additional LOF test after each model selection step if a further reduction in FDR is desired. These two methods outperformed other methods included in the study in both the Pareto front's distance from the Utopia point and Desirability analysis. If one does not want to use adjusted p-values, we recommend using Unadjusted p_{pe} with an additional LOF test after each model selection step and an α significance level of greater than or equal to 0.25. Additional recommendations include having more than two replicate points whether replicates can be added in addition to a D-optimal design or included in the total run size when incorporating pure error with Bonferroni PE and Bonferroni PE LOF. We limit our study to focus on active main effects and 2FI but further research including the detection of quadratic terms and response surface models is recommended. We also recommend sequential hypothesis tests including conditional tests for further research.

CHAPTER 5

CONCLUSION AND FINAL REMARKS

In this dissertation, I have researched various open problems of experimental design. First, Chapter 2 delves into the selection of experimental designs for calibration. Next, Chapter 3 compares optimal sequential design techniques and nonsequential designs to answer the question if it is best to start with a small screening design and augment the design with remaining runs or run a single nonsequential design. Lastly, Chapter 4 examines why high power and high FDR rates are seen when statistical inference is based on pure error during model selection and evaluates various model selection techniques to find a recommended model selection strategy incorporating pure error that balances power and FDR. After reviewing the literature and a theoretical discussion, each chapter includes a simulation study to further investigate each open problem.

Chapter 2 compares experimental designs with regards to inverse prediction performance with the goal to help influence design choice and setup of future calibration studies. Our simulation study varies design type, the number of input factors, the number of response variables, the error standard deviation, correlations among responses, the number of active terms, and whether or not predictor subset selection is helpful. The designs are evaluated based on the mean/median Euclidean distance between actual and predicted input values. Results from the simulation indicate a relationship between design choice and inverse prediction performance and support the use of I-optimal, CCD, and bridge designs with a small tuning parameter (Bridge25). These three design types had the lowest Euclidean distance between actual and predicted input values overall as well as across all of the simulation scenarios. Additional recommendations include limiting the number of input factors and making use of the full hypothesized model. Correlation among response variables did not appear to have an overall negative impact on inverse prediction and further research is recommended. Other opportunities for future work include considering errors in the input factors, cases when the true model is a higher-order polynomials or more complex nonlinear models but the user locally approximates the response surface by lower-order polynomial models, and investigating other calibration methods.

Chapter 3 evaluates optimal sequential design techniques and optimal nonsequential designs with regards to power and FDR rates to help experimenters decide whether to run a small screening design first or use all runs in a nonsequential approach to identify active terms given a limited number of experimental trials. Our simulation study varies the number of total experimental runs, the number of input factors, the number of active terms, and considers model selection methods of LASSO and forward selection. The designs are evaluated based on mean power and FDR rates. Our simulation results indicate a difference in power and FDR between sequential and nonsequential designs and support the use of nonsequential designs whether maximizing power is the main objective or balancing high power with low FDR rates. We did not consider the ability to move outside the original design region that may be a key benefit to sequential experiments and we recommend this for future research. We also focused primarily on Bayesian D-optimal designs as well as Bayesian D-optimal augmentation, investigating other design types as well as other design augmentation methods would also be interesting for future work.

Chapter 4 explores why high power and high FDR are seen when statistical inference is based on a pure error estimate during model selection and proposes a strategy for incorporating pure error in model selection procedures that keeps FDR in check. We compare various model selection techniques using a simulation study. We consider early stopping methods including Bonferroni adjusted p-values and our proposed forward selection method that incorporates a LOF test after each model selection step. We also develop a method of incorporating pure error with LASSO penalized regression that computes LOF tests along the LASSO solution path. Our simulation study varies the number of input factors, the number of active terms, the α significance level, the number of total runs, the number of replicate points, whether replicate points are added in addition to the total runs or included in the total, and the model selection method. Model selection methods are evaluated based on mean power and FDR. Our simulation results support the use of forward selection with Bonferroni adjusted p_{pe} with an α significance level of 0.5. We also recommend forward selection with Bonferroni adjusted p_{pe} with an additional LOF test after each model selection step if a further reduction in FDR is desired. We additionally recommend using Unadjusted p_{pe} with LOF with an α significance level of 0.25 if an alternative to adjusted p-values is needed. We limit our study to focus on active main effects and 2FI, but further research including the detection of quadratic terms and response surface models is recommended. We also recommend sequential hypothesis tests including conditional tests as another avenue for future research.

Appendix A

COMPARISON OF EXPERIMENTAL DESIGNS FOR CALIBRATION RESULTS

Design1	Design2	Difference	Lower Range	Upper Range	P-value
Bridge50	Bridge25	0.013	0.002	0.024	0.006
Bridge75	Bridge25	0.019	0.008	0.030	0.000
CCD	Bridge25	-0.008	-0.019	0.003	0.354
D-Optimal	Bridge25	0.007	-0.004	0.018	0.528
I-Optimal	Bridge25	-0.009	-0.020	0.002	0.206
Latin Hypercube	Bridge25	0.017	0.006	0.028	0.000
Spatial Coverage	Bridge25	0.045	0.034	0.056	0.000
Bridge75	Bridge50	0.006	-0.005	0.017	0.777
CCD	Bridge50	-0.021	-0.032	-0.010	0.000
D-Optimal	Bridge50	-0.006	-0.017	0.005	0.651
I-Optimal	Bridge50	-0.022	-0.033	-0.011	0.000
Latin Hypercube	Bridge50	0.003	-0.008	0.014	0.981
Spatial Coverage	Bridge50	0.032	0.021	0.043	0.000
CCD	Bridge75	-0.027	-0.038	-0.016	0.000
D-Optimal	Bridge75	-0.012	-0.023	-0.001	0.021
I-Optimal	Bridge75	-0.028	-0.039	-0.017	0.000
Latin Hypercube	Bridge75	-0.002	-0.013	0.009	0.999
Spatial Coverage	Bridge75	0.026	0.015	0.037	0.000
D-Optimal	CCD	0.015	0.004	0.026	0.001
I-Optimal	CCD	-0.001	-0.012	0.010	1.000
Latin Hypercube	CCD	0.025	0.014	0.036	0.000
Spatial Coverage	CCD	0.053	0.042	0.064	0.000
I-Optimal	D-Optimal	-0.016	-0.027	-0.005	0.000
Latin Hypercube	D-Optimal	0.010	-0.001	0.021	0.122
Spatial Coverage	D-Optimal	0.038	0.027	0.049	0.000
Latin Hypercube	I-Optimal	0.026	0.015	0.037	0.000
Spatial Coverage	I-Optimal	0.054	0.043	0.065	0.000
Spatial Coverage	Latin Hypercube	0.028	0.017	0.039	0.000

Table 15.: Calibration TukeyHSD results showing pairwise design comparison of mean Euclidean distance and significance level

Appendix B

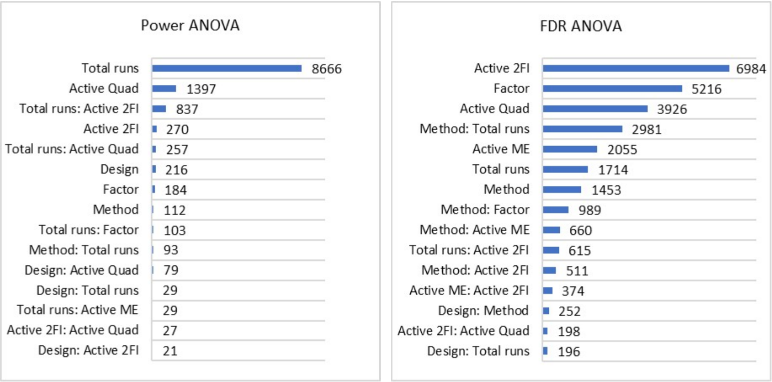
OPTIMAL SEQUENTIAL DESIGN TECHNIQUES RESULTS

Table 16.: Optimal sequential design ANOVA for power and FDR

(a) Power				(b) FDR	
Term	Df	F value	Pr(>F)	F value	Pr(>F)
Design	2	216	<2.2e-16	94	<2.2e-16
Method	1	112	<2.2e-16	1453	<2.2e-16
Total runs	1	8666	<2.2e-16	1714	<2.2e-16
Factor	1	184	<2.2e-16	5216	<2.2e-16
Active ME	1	3	0.070	2055	<2.2e-16
Active 2FI	1	270	<2.2e-16	6984	<2.2e-16
Active Quad	1	1397	<2.2e-16	3926	<2.2e-16
Design: Method	2	6	0.003	252	<2.2e-16
Design: Total runs	2	29	<2.08e-13	196	<2.2e-16
Design: Factor	2	2	0.102	3	0.058
Design: Active ME	2	0	0.646	11	<2.21e-05
Design: Active 2FI	2	21	<1.01e-09	9	<8.41e-05
Design: Active Quad	2	79	<2.2e-16	143	<2.2e-16
Method: Total runs	1	93	<2.2e-16	2981	<2.2e-16
Method: Factor	1	5	0.031	989	<2.2e-16
Method: Active ME	1	1	0.462	660	<2.2e-16
Method: Active 2FI	1	1	0.383	511	<2.2e-16
Method: Active Quad	1	9	0.003	84	<2.2e-16
Total runs: Factor	1	103	<2.2e-16	10	0.002
Total runs: Active ME	1	29	<8.82e-08	29	<6.99e-08
Total runs: Active 2FI	1	837	<2.2e-16	615	<2.2e-16
Total runs: Active Quad	1	257	<2.2e-16	33	<8.04e-09
Factor: Active ME	1	1	0.225	4	0.038
Factor: Active 2FI	1	19	<1.28e-05	68	<3.00e-16
Factor: Active Quad	1	2	0.118	4	0.056
Active ME: Active 2FI	1	16	<6.65e-05	374	<2.2e-16
Active ME: Active Quad	1	19	<1.51e-05	107	<2.2e-16
Active 2FI: Active Quad	1	27	<2.36e-07	198	<2.2e-16
Residuals	3204				

All main effects are statistically significant (p-value < 0.05) in the ANOVA models in Table 16 except for the number of active main effects in the power model (Table 16(a)). For the power ANOVA, all 2FI are statistically significant with the exception of the number of factors with design, the number of active main effects with design, the number of active main effects with method, the number of active 2FI with method, the number of factors with active main effects and the number of factors with the number of active quadratic terms. For the FDR ANOVA model, all 2FI are statistically significant except for design with the number of factors, and the number of factors with the number of active quadratics (Table 16(b)).

Figure 21(a) plots the top 15 F-values from the power ANOVA model. Total runs stands out with the largest F-value followed by the number of active quadratics and 2FI along with the interaction of the number of active quadratics with total runs and 2FI with total runs. “Design” has the next largest F-value followed by the number of factors and analysis method. Total runs with method, the number of factors, active 2FI, and active quadratic terms are all identified in the top 15 F-values in Figure 21(a) and we provide plot comparing the mean power performance of sequential and nonsequential designs by these variables in Figure 10.



(a) (b)

Fig. 21.: Optimal Sequential Design Techniques top 15 F-values in power and FDR ANOVA models.

The top 15 F-values for FDR shown in Figure 21(b) are closer in size with the number of active 2FI as the largest F-value followed closely by the number of factors and the number of active quadratic terms. The interactions of analysis method with total runs and analysis method with the number of factors also appear in the top 15 F-values. As interactions with method appeared prominently in the top 15 F-values, plots comparing the mean FDR performance of sequential and nonsequential designs looking at method with total runs, the number of factors, active 2FI and quadratic terms are provided in Figure 11.

Table 17.: Optimal Sequential Design Techniques desirability analysis results for main effect (ME) power and FDR (Mean Values)

Design	Predicted ME Power	Predicted ME FDR	Desirability ME Power	Desirability ME FDR	Overall Desirability
Nonsequential	0.918	0.107	0.436	0.282	0.141
Sequential 3 Level	0.899	0.109	0.373	0.243	0.118
Sequential 2 Level	0.89	0.116	0.35	0.203	0.097

Table 18.: Optimal Sequential Design Techniques desirability analysis results for 2FI power and FDR (mean values)

Design	Predicted 2FI Power	Predicted 2FI FDR	Desirability 2FI Power	Desirability 2FI FDR	Overall Desirability
Nonsequential	0.848	0.502	0.384	0.154	0.090
Sequential 3 Level	0.818	0.523	0.336	0.125	0.074
Sequential 2 Level	0.849	0.539	0.379	0.11	0.084

Table 19.: Optimal Sequential Design Techniques desirability analysis results for quadratic term power and FDR (mean values)

Design	Predicted Quad Power	Predicted Quad FDR	Desirability Quad Power	Desirability Quad FDR	Overall Desirability
Nonsequential	0.713	0.274	0.343	0.203	0.102
Sequential 3 Level	0.692	0.286	0.336	0.174	0.096
Sequential 2 Level	0.432	0.182	0.046	0.339	0.036

Table 20.: Optimal Sequential Design Techniques desirability analysis results for power and FDR no active quadratic terms (mean values)

Design	Predicted Power	Predicted FDR	Desirability Power	Desirability FDR	Overall Desirability
Nonsequential	0.915	0.442	0.411	0.124	0.065
Sequential 3 Level	0.894	0.445	0.352	0.104	0.051
Sequential 2 Level	0.901	0.429	0.359	0.114	0.077

Appendix C

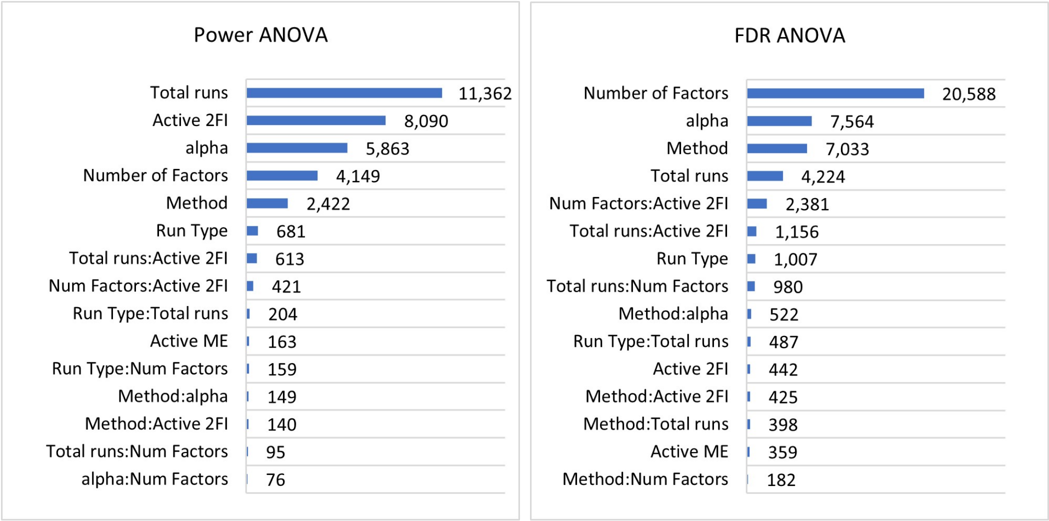
MODEL SELECTION WITH PURE ERROR RESULTS

Table 21.: Model Selection ANOVA for power and FDR

	(a) Power			(b) FDR	
Term	Df	F value	Pr(>F)	F value	Pr(>F)
Method	10	2422	<2.2e-16	7033	<2.2e-16
Run Type	1	681	<2.2e-16	1007	<2.2e-16
alpha	4	5863	<2.2e-16	7564	<2.2e-16
Total runs	3	11362	<2.2e-16	4224	<2.2e-16
Number of Factors	1	4149	<2.2e-16	20588	<2.2e-16
Active ME	3	163	<2.2e-16	359	<2.2e-16
Active 2FI	4	8090	<2.2e-16	442	<2.2e-16
Method: Run Type	10	31	<2.2e-16	50	<2.2e-16
Method: alpha	40	149	<2.2e-16	522	<2.2e-16
Method: Total runs	30	51	<2.2e-16	398	<2.2e-16
Method: Num Factors	10	55	<2.2e-16	182	<2.2e-16
Method: Active ME	30	22	<2.2e-16	64	<2.2e-16
Method: Active 2FI	40	140	<2.2e-16	425	<2.2e-16
Run Type: alpha	4	4	0.005	8	3.27e-06
Run Type: Total runs	3	204	<2.2e-16	487	<2.2e-16
Run Type: Num Factors	1	159	<2.2e-16	151	<2.2e-16
Run Type: Active ME	3	5	0.003	1	0.263
Run Type: Active 2FI	4	37	<2.2e-16	78	<2.2e-16
alpha: Total runs	12	63	<2.2e-16	68	<2.2e-16
alpha: Num Factors	4	76	<2.2e-16	113	<2.2e-16
alpha: Active ME	12	2	0.031	38	<2.2e-16
alpha: Active 2FI	16	7	7.24e-16	71	<2.2e-16
Total runs: Num Factors	3	95	<2.2e-16	980	<2.2e-16
Total runs: Active ME	9	21	<2.2e-16	3	4.48e-04
Total runs: Active 2FI	12	613	<2.2e-16	1156	<2.2e-16
Num Factors: Active ME	3	22	4.78e-14	9	4.46e-06
Num Factors: Active 2FI	4	421	<2.2e-16	2381	<2.2e-16
Active ME: Active 2FI	12	41	<2.2e-16	21	<2.2e-16
Residuals	39711				

All terms in the power ANOVA model (Table 21(a)) and all except the interaction of run type with the number of active main effects in the FDR ANOVA model (Table 21(b)) have a p-value ≤ 0.05. The top 15 F-values in magnitude are shown in Figure 22(a) for power, and Figure 22(b) for FDR. Total runs, number of active 2FI, alpha level, the number of factors, and method have the highest F-values in the power ANOVA. The number of factors, alpha level, method, and total runs also have the highest F-values in the FDR ANOVA.

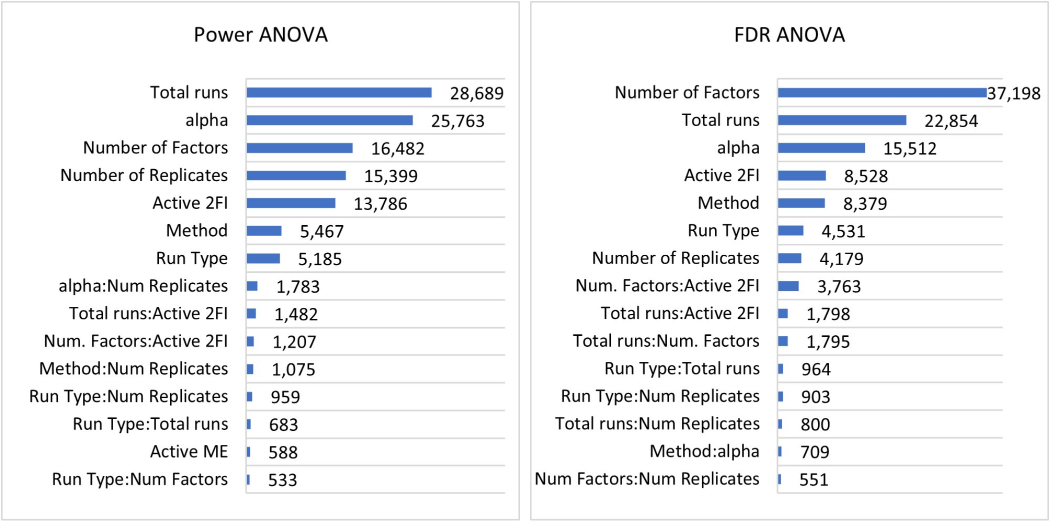
We ran a separate ANOVA on the subset of methods with pure error to analyze the impact of replicate runs (Table 22 in the Appendix). The top 15 F-values in magnitude are shown in Figure 23(a) for power, and Figure 23(b) for FDR. All terms in the power



(a) (b)

Fig. 22.: Model Selection top 15 F-values in power and FDR ANOVA models.

ANOVA model have a p-value ≤ 0.05 except for the interaction of number of active main effects with the number of replicate points. All terms in the FDR ANOVA model except for the interaction of run type with the number of active main effects have a p-value ≤ 0.05. The number of replicate runs main effect term and the interaction with run type appear on both the top 15 F-values for power and FDR.



(a) (b)

Fig. 23.: Model Selection top 15 F-values from methods with pure error power and FDR ANOVA models.

Table 22.: Model Selection pure error model selection methods ANOVA for power and FDR

	(a) Power			(b) FDR	
Term	Df	F value	Pr(>F)	F value	Pr(>F)
Method	6	5467	<2.2e-16	8379	<2.2e-16
Run Type	1	5185	<2.2e-16	4531	<2.2e-16
alpha	4	25763	<2.2e-16	15512	<2.2e-16
Total runs	3	28689	<2.2e-16	22854	<2.2e-16
Number of Factors	1	16482	<2.2e-16	37198	<2.2e-16
Active ME	3	588	<2.2e-16	25	6.61e-16
Active 2FI	4	13786	<2.2e-16	8528	<2.2e-16
Number of Replicates	2	15399	<2.2e-16	4179	<2.2e-16
Method: Run Type	6	8	1.26e-08	17	<2.2e-16
Method: alpha	24	295	<2.2e-16	709	<2.2e-16
Method: Total runs	18	128	<2.2e-16	193	<2.2e-16
Method: Num Factors	6	183	<2.2e-16	184	<2.2e-16
Method: Active ME	18	58	<2.2e-16	36	<2.2e-16
Method: Active 2FI	24	129	<2.2e-16	162	<2.2e-16
Method: Num Replicates	12	1075	<2.2e-16	125	<2.2e-16
Run Type: alpha	4	13	3.65e-10	15	1.71e-12
Run Type: Total runs	3	683	<2.2e-16	964	<2.2e-16
Run Type: Num Factors	1	533	<2.2e-16	300	<2.2e-16
Run Type: Active ME	3	16	4.30e-10	2	0.059
Run Type: Active 2FI	4	123	<2.2e-16	156	<2.2e-16
Run Type: Num Replicates	2	959	<2.2e-16	903	<2.2e-16
alpha: Total runs	12	190	<2.2e-16	113	<2.2e-16
alpha: Num Factors	4	187	<2.2e-16	151	<2.2e-16
alpha: Active ME	12	4	3.21e-05	60	<2.2e-16
alpha: Active 2FI	16	30	<2.2e-16	137	<2.2e-16
alpha: Num Replicates	8	1783	<2.2e-16	195	<2.2e-16
Total runs: Num Factors	3	214	<2.2e-16	1795	<2.2e-16
Total runs: Active ME	9	44	<2.2e-16	10	4.11e-16
Total runs: Active 2FI	12	1482	<2.2e-16	1798	<2.2e-16
Total runs: Num Replicates	6	511	<2.2e-16	800	<2.2e-16
Num Factors: Active ME	3	69	<2.2e-16	44	<2.2e-16
Num Factors: Active 2FI	4	1207	<2.2e-16	3763	<2.2e-16
Num Factors: Num Replicates	2	79	<2.2e-16	551	<2.2e-16
Active ME: Active 2FI	12	107	<2.2e-16	20	<2.2e-16
Active ME: Num Replicates	6	1	0.332	3	0.018
Active 2FI: Num Replicates	8	81	<2.2e-16	202	<2.2e-16
Residuals	33333				

Table 23.: Model Selection Method desirability analysis results for main effect power and FDR (mean values)

Design	Predicted Power	Predicted FDR	Overall Desirability
Bonferroni PE	0.696	0.048	0.241
Bonferroni PE LOF	0.681	0.046	0.222
Unadjusted PE LOF	0.786	0.082	0.217
AICc LOF	0.738	0.072	0.199
Unadjusted PE	0.886	0.167	0.154
Unadjusted SSE LOF	0.708	0.069	0.150
LASSO AICc	0.935	0.129	0.150
LASSO LOF	0.651	0.054	0.110
Bonferroni SSE	0.470	0.017	0.106
AICc	0.915	0.224	0.089
Unadjusted SSE	0.759	0.203	0.023

Table 24.: Model Selection Method desirability analysis results for 2FI power and FDR (mean values)

Design	Predicted Power	Predicted FDR	Overall Desirability
Bonferroni PE	0.485	0.243	0.176
Bonferroni PE LOF	0.462	0.241	0.158
Unadjusted PE LOF	0.542	0.325	0.151
LASSO LOF	0.517	0.341	0.134
AICc LOF	0.469	0.302	0.119
Unadjusted PE	0.692	0.461	0.104
Unadjusted SSE LOF	0.465	0.271	0.097
Bonferroni SSE	0.269	0.074	0.057
AICc	0.623	0.508	0.055
Unadjusted SSE	0.573	0.447	0.034
LASSO AICc	0.700	0.507	0.033

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akkermans, W. G., Coppenolle, H., and Goos, P. (2017). Optimal design of experiments for excipient compatibility studies. *Chemometrics and Intelligent Laboratory Systems*, 171(125-139).

Anderson-Cook, C. M., Borror, C. M., and Montgomery, D. C. (2009). Response surface design evaluation and comparison. *Journal of Statistical Planning and Inference*, 139(2):629–641.

Anderson-Cook, C. M., Burr, T., Hamada, M. S., Ruggiero, C., and Thomas, E. V. (2015). Design of experiments and data analysis challenges in calibration for forensics applications. *Chemometrics and Intelligent Laboratory Systems*, 149:107–117.

Antony, J. (2003). *Design of experiments for engineers and scientists*. Elsevier.

Asadollahzadeh, M., Tavakoli, H., Torab-Mostaedi, M., Hosseini, G., and Hemmati, A. (2014). Response surface methodology based on central composite design as a chemometric tool for optimization of dispersive-solidification liquid–liquid microextraction for speciation of inorganic arsenic in environmental water samples. *Talanta*, 123:25–31.

Barrios, E. and Meyer, D. (2020). BsMD: Bayes Screening and Model Discrimination. *https://CRAN.R-project.org/package=BsMD*, (R package version 2020.4.30).

Bella, F., Munoz-García, A. B., Colo, F., Meligrana, G., Lamberti, A., Destro, M., Pavone, M., and Gerbaldi, C. (2018). Combined structural, chemometric, and electrochemical investigation of vertically aligned TiO2 nanotubes for Na-ion batteries. *ACS omega*, 3(7):8440–8450.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Royal Statistical Society*, 57:289–300.

Bondi, R. W., Igne, B., Drennen, J. K., and Anderson, C. A. (2012). Effect of experimental design on the prediction performance of calibration models based on near-infrared spectroscopy for pharmaceutical applications. *Applied spectroscopy*, 66(12):1442–1453.

Brown, P. J. (1993). *Measurement, regression, and calibration.* Oxford, Oxford, UK.

Butler, N. A. (2001). Optimal and Orthogonal Latin Hypercube Designs for Computer Experiments. *Biometrika*, 88(3):847–857.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Davis, M. J., Liu, W., and Sivaramakrishnan, R. (2017). Global sensitivity analysis with small sample sizes: ordinary least squares approach. *The Journal of Physical Chemistry A*, 121(3):553–570.

Ebrahimi-Najafabadi, H., Leardi, R., Oliveri, P., Casolino, M. C., Jalali-Heravi, M., and Lanteri, S. (2012). Detection of addition of barley to coffee using near infrared spectroscopy and chemometric techniques. *Talanta*, 99:175–179.

Edwards, D. J., Weese, M. L., and Palmer, G. A. (2014). Comparing methods for design follow-up: Revisiting a metal-cutting case study. *Applied Stochastic Models in Business and Industry*, 30(4):464–478.

François, N., Govaerts, B., and Boulanger, B. (2004). Optimal designs for inverse prediction in univariate nonlinear calibration models. *Chemometrics and Intelligent Laboratory Systems*, 74:283–292.

François, Y., Varenne, A., Juillerat, E., Servais, A.-C., Chiap, P., and Gareil, P. (2007). Nonaqueous capillary electrophoretic behavior of 2-aryl propionic acids in the presence of an achiral ionic liquid: a chemometric approach. *Journal of Chromatography A*, 1138(1-2):268–275.

Frommlet, F., Ruhaltinger, F., Twarog, P., and Bogdan, M. (2010). A model selection approach to genome wide association studies. *arXiv preprint arXiv:1010.0124*.

Gay, D. M. (1990). Usage Summary for Selected Optimization Routines (PORT Mathematical Subroutine Library, Optimization chapter). *Computing science technical report*, 153.

Gilmour, S. G. and Trinca, L. A. (2012). Optimum design of experiments for statistical inference. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 61(3):345–401.

Goos, P. and Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach.* John Wiley & Sons.

Gutman, A. J., White, E. D., Lin, D. K. J., and Hill, R. R. (2014). Augmenting supersaturated designs with Bayesian D-optimality. *Computational Statistics and Data Analysis*, 71:1147–1158.

Haaland, D. M. and Thomas, E. V. (1988). Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Analytical Chemistry*, 60(11):1193–1202.

Hathurusingha, P. I. and Davey, K. R. (2016). Chemical taste taint accumulation in RAS farmed fish–A Fr 13 risk assessment demonstrated with geosmin (GSM) and 2-methylisoborneol (MIB) in barramundi (Lates calcarifer). *Food control*, 60:309–319.

Jeirani, Z., Jan, B. M., Ali, B. S., Noor, I. M., and Hwa, See Chun Saphanuchart, W. (2012). The optimal mixture design of experiments: Alternative method in optimizing the aqueous phase composition of a microemulsion. *Chemometrics and Intelligent Laboratory Systems*, 112:1–7.

Jensen, W. A. (2018). Open problems and issues in optimal design. *Quality Engineering*, 30(4):583–593.

John, P. W. (2000). Breaking alias chains in fractional factorials. *Communications in Statistics-Theory and Methods*, 29(9-10):2143–2155.

Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148.

Johnson, R. T., Montgomery, D. C., Jones, B., and Parker, P. A. (2010). Comparing computer experiments for fitting high-order polynomial metamodels. *Journal of Quality Technology*, 42(1):86–102.

Jones, B., Silvestrini, R. T., Montgomery, D. C., and Steinberg, D. M. (2015). Bridge Designs for Modeling Systems With Low Noise. *Technometrics*, 57(2):155–163.

Joseph, V. R. (2016). Space-filling designs for computer experiments : A review. *Quality Engineering*, 28(1):28–35.

Kelley, K. (2020). MBESS: The MBESS R Package. *https://cran.r-project.org/package=MBESS*, (R package version 4.7.0).

Khuri, A. I. and Cornell, J. A. (1996). *Response surfaces : designs and analyses.* Marcel Dekker, New York, 2nd ed., r edition.

Kuhn, M. (2016). desirability: Function Optimization and Ranking via Desirability Functions. *https://cran.r-project.org/web/packages/desirability/vignettes/desirability.pdf*, (R package version 2.1).

Leonard, R. D. and Edwards, D. J. (2017). Bayesian D-optimal screening experiments with partial replication. *Computational Statistics and Data Analysis*, 115:79–90.

Lewis, J. R., Zhang, A., and Anderson-Cook, C. M. (2018). Comparing multiple statistical methods for inverse prediction in nuclear forensics applications. *Chemometrics and Intelligent Laboratory Systems*, 175:116–129.

Loveday, S., Wang, X., Rao, M., Anema, S., and Singh, H. (2011). Effect of pH, NaCl, CaCl2 and temperature on self-assembly of beta-lactoglobulin into nanofibrils: a central composite design study. *Journal of agricultural and food chemistry*, 59(15):8467–8474.

Lu, L., Anderson-Cook, C. M., and Robinson, T. J. (2011). Optimization of designed experiments based on multiple criteria utilizing a pareto frontier. *Technometrics*, 53(4):353–365.

Mannarswamy, A., Munson-McGee, Stuart H Steiner, R., and Andersen, P. K. (2009). D-optimal experimental designs for Freundlich and Langmuir adsorption isotherms. *Chemometrics and Intelligent Laboratory Systems*, 97(2):146–151.

Mee, R. W. (2013). Tips for analyzing nonregular fractional factorial experiments. *Journal of Quality Technology*, 45(4):330–349.

Mee, R. W. and Peralta, M. (2000). Semifolding 2 k–p Designs. *Technometrics*, 42(2):122–134.

Mee, R. W., Schoen, E. D., and Edwards, D. J. (2017). Selecting an Orthogonal or Nonorthogonal Two-Level Design for Screening. *Technometrics*, 59(3):305–318.

Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1):60–69.

Meyer, R. Daniel, Steinberg, D. M., and Box, G. (1996). Follow-up designs to resolve confounding in multifactor experiments. *Technometrics*, 38(4):303–313.

Montgomery, Douglas C., Runger, G. C. (1996). Foldovers of 2 kp resolution IV experimental designs. *Journal of Quality Technology*, 28(4):446–450.

Moudiki, T. (2013). mcGlobaloptim: Global optimization using Monte Carlo and Quasi Monte Carlo simulation. *https://cran.r-project.org/web/packages/mcGlobaloptim/mcGlobaloptim.pdf*, (R package version 0.1).

Myers, R. H. (1990). *Classical and Modern Regression with Applications*. Duxbury press, Belmont, CA, vol. 2 edition.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley, fourth edition.

Nelson, B. J., Montgomery, D. C., Elias, R. J., and Maass, E. (2000). A Comparison of Several Design Augmentation Strategies. *Quality and Reliability Engineering International*, 16:435–449.

Nunes, L. M., da Conceição Cunha, M., and Ribeiro, L. (2013). Coverage methods for early groundwater contamination detection. *Bulletin of environmental contamination and toxicology*, 90(5):531–536.

Nychka, D., Furrer, R., and Sain, S. (2019). fields: Tools for spatial data. *https://cran.r-project.org/web/packages/fields/fields.pdf*.

Ockuly, R. A., Weese, M. L., Smucker, B. J., Edwards, D. J., and Chang, L. (2017). Response surface experiments: A meta-analysis. *Chemometrics and Intelligent Laboratory Systems*, 164:64–75.

Panagiotou, A. N., Sakkas, V. A., and Albanis, T. A. (2009). Application of chemometric assisted dispersive liquid–liquid microextraction to the determination of personal care products in natural waters. *Analytica chimica acta*, 649(2):135–140.

Plackett, R. L. and Burman, J. P. B. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33(4):305–325.

R Core Team (2018). R: A Language and Environment for Statistical Computing.

Rai, A., Mohanty, B., and Bhargava, R. (2016). Supercritical extraction of sunflower oil: A central composite design for extraction variables. *Food chemistry*, 192(647-659).

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Rönkkö, M. (2014). The Effects of Chance Correlations on Partial Least Squares Path Modeling. *Organizational Research Methods*, 17(2):164–181.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4):965–984.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers \& Geosciences*, 24(5):479–488.

SAS Institute Inc. (2019). *JMP® 15 Design of Experiments Guide*. Cary, NC: SAS Institute Inc.

Silvestrini, R. T. (2015). Considerations for D-optimal sequential design. *Quality and Reliability Engineering International*, 31(3):399–410.

Smucker, B. J., Edwards, D. J., and Weese, M. L. (2020). Response surface models: To reduce or not to reduce? *Journal of Quality Technology*, pages 1–20.

Sundberg, R. (1996). The precision of the estimated generalized least squares estimator in multivariate calibration. *Scandinavian journal of statistics*, pages 257–274.

Sundberg, R. (1999). Multivariate calibration—direct and indirect regression methodology. *Scandinavian Journal of Statistics*, 26(2):161–207.

Thomas, E. V., Lewis, J. R., Anderson-Cook, C. M., Burr, T., and Hamada, M. S. (2017). Selecting an Informative/Discriminating Multivariate Response for Inverse Prediction. *Journal of Quality Technology*, 49(3):228–243.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Westfall, P. H., Young, S. S., and Lin, D. K. J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, 8(1):101–117.

Xiang, D., Berry, J., Buntz, S., Gargiulo, P., Cheney, J., Joshi, Y., Wabuyele, B., Wu, H., Hamed, M., Hussain, A., and Khan, M. (2009). Robust Calibration Design in the Pharmaceutical Quantitative Measurements with Near-Infrared (NIR) Spectroscopy: Avoiding the Chemometric Pitfalls. *Journal of Pharmaceutical Sciences*, 98(3).

Xu, Q.-S., Xu, Y.-D., Li, L., and Fang, K.-T. (2018). Uniform experimental design in chemometrics. *Journal of Chemometrics*, 32(11).

Zhang, Q. Z., Dai, H. S., Liu, M. Q., and Wang, Y. (2019). A method for augmenting supersaturated designs. *Journal of Statistical Planning and Inference*, 199:207–218.