A Comparison of Supersaturated Designs and Orthogonal Arrays

Jacob Akubire Abugre[1] and Byran J. Smucker*[2]

[1]Department of Statistics, Miami University, Oxford, OH

Abstract

The purpose of a screening experiment is to accurately and cost-efficiently identify the few most influential factors, from among the many studied. Two types of screening experiments are Supersaturated Designs (SSDs), a bold strategy wherein the number of factors (k) exceeds the number of experimental runs (n), and Orthogonal Arrays (OAs), which requires n to be a multiple of 4 and greater than k. In this study, we compare the performance of SSDs with OAs, using both simulation as well as a new method (Stallrich et al. 2024) based on the probability of perfect sign recovery under the Lasso. Our findings indicate that under conditions of high sparsity, SSDs and OAs detect effects with similar reliability, if the SSD effects are about 1.5 times the size of the OA effects.

Keywords: Screening experiments; Sign recovery; Regularization; the Lasso

1 Introduction

An important aspect of screening experimental design and analysis is the selection of active experimental factors, and the dismissal of factors with no meaningful effect on the response

*corresponding author, smuckerb@miamioh.edu

variable. A particularly bold screening strategy is supersaturated experiments, where the number of runs n is less than the number of factors k. Though these experiments have been well-studied, practitioners have been hesitant to use them. Orthogonal arrays, on the other hand, have a long history of implementation. The aim of this paper is to compare these two screening strategies, in order to better understand the tradeoffs between run size and screening quality.

In a screening experiment, it is customary to adopt the linear main effect model represented as:

$$\mathbf{Y} = \beta_0 \mathbf{1} + D\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, D is the design matrix of size $n \times k$ with entries ± 1, $\boldsymbol{\beta}$ is a vector of model parameters of size $k \times 1$, and $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are each $n \times 1$ vectors. The objective of screening experiments is to cost-effectively determine which elements of $\boldsymbol{\beta}$ are non-zero. The use of Orthogonal Arrays (OAs) is prevalent in screening experiments (e.g. Hedayat et al. 1999), where n, being a multiple of 4, should always exceed k. On the other hand, Supersaturated Designs (SSDs) can have n considerably smaller than $k+1$. Currently, there is a notable degree of skepticism regarding the adoption of SSDs. According to an informal survey conducted by Weese et al. (2021), only 13 out of 63 respondents reported having used them in the past with only a handful using them regularly. Some respondents were concerned that SSDs don't have sufficient power to detect effects of interest. On the other hand, orthogonal designs, such as fractional factorials and Plackett-Burman designs, were used much more regularly, among the respondents.

Given this prevailing sentiment, our objective in this work is to quantify the difference in screening ability of OAs vs. SSDs, across various effect sizes and levels of sparsity. Specifically, we seek to evaluate the amount of information lost when using SSDs instead of OAs, as well as to determine conditions under which the performances of the designs are similar.

We analyze the designs using the Lasso to compare them based upon simulated true positive rate (proportion of true active factors that are identified as active) and false positive rate (proportion of inactive factors that are identified as active), as well as with a measure based on the probability of perfect sign recovery under the Lasso (Stallrich et al. 2024). Our goal is to increase understanding of the performance of SSDs relative to their more accepted counterparts. Then, experimenters will be better equipped to weigh the benefits of SSDs along with their drawbacks.

The outline of this paper is as follows: Section 2 offers a review of the relevant technical material regarding SSDs and OAs. Section 3 provides a description of our simulation protocol as well as the criteria proposed by Stallrich et al. (2024), both used to assess the quality of the designs. We provide the results of our study in Section 4, and a discussion in Section 5.

2 Background

Screening experiments have been used widely in the physical sciences and manufacturing (Hedayat et al. 1999; Mee 2009), as well as computer experiments (Owen 1992). Of particular interest to this work are orthogonal arrays (Plackett and Burman 1946; Rao 1947; Bush 1952; Bose and Bush 1952; Hedayat et al. 1999; Xu et al. 2009) and supersaturated designs (Satterthwaite 1959; Lin 1993; Wu 1993; Georgiou 2014; Weese et al. 2021). With regard to the latter, there have been only a limited number of applications in the literature (Carpinteiro et al. 2004; Dejaegher and Vander Heyden 2008; Jridi et al. 2015; Zarkadas and Besseris 2023), while the former are widely accepted by practitioners.

2.1 Screening Designs

Plackett and Burman (1946) were the first to introduce something like orthogonal arrays, and this led to the development of a considerable theory (e.g., Rao 1947; Bush 1952; Bose and Bush 1952; Hedayat et al. 1999). An orthogonal array is an $n \times k$ matrix which can be described with (n, k, s, d), where k is number of factors each at one of s levels, and n is the number of experimental runs and a subset of all possible s^k treatment combinations. This array is said to be of strength d if all s^d treatment combinations corresponding to any d factors chosen out of k occur an equal number of times. An extensive review of orthogonal arrays can be found in Xu et al. (2009) and Mee et al. (2017). Because these designs have orthogonal columns, they estimate the main effects with minimum variance and no bias, under the assumptions of model (1). In this paper, we consider OA's with $s = 2$.

Supersaturated designs are a bolder type of screening design, for which $n < k$, so that orthogonal designs are not possible. These experiments were first suggested by Satterthwaite (1959) who described the construction of random balance designs, in which factor levels were chosen randomly according to some distribution. After the statistical properties of these random supersaturated designs were criticised, Booth and Cox (1962) proposed a systematic way of constructing SSDs. Their approach was based upon the intuition that design columns be as nearly orthogonal as possible. In particular, they suggested a criterion, called $E(s^2)$, based on the average of the squared off-diagonal elements of $D^T D$, requiring that each column of D have the same number of $+1$ and -1 entries. Let $X = [\mathbf{1}|D]$, where $\mathbf{1}$ is a $n \times 1$ vector representing the intercept, and let s_{ij} be the ij^{th} off-diagonal value of $X^T X$. Then we have that $E(s^2) = \frac{2}{k(k-1)} \sum_{2 \leq i < j \leq k} s_{ij}^2$. Later, several authors (Marley and Woods 2010; Jones and Majumdar 2014; Weese et al. 2015) relaxed the balance constraint, leading to the so-called unconditional $E(s^2)$ criterion, $UE(s^2) = \frac{2}{k(k+1)} \sum_{1 \leq i < j \leq k} s_{ij}^2$.

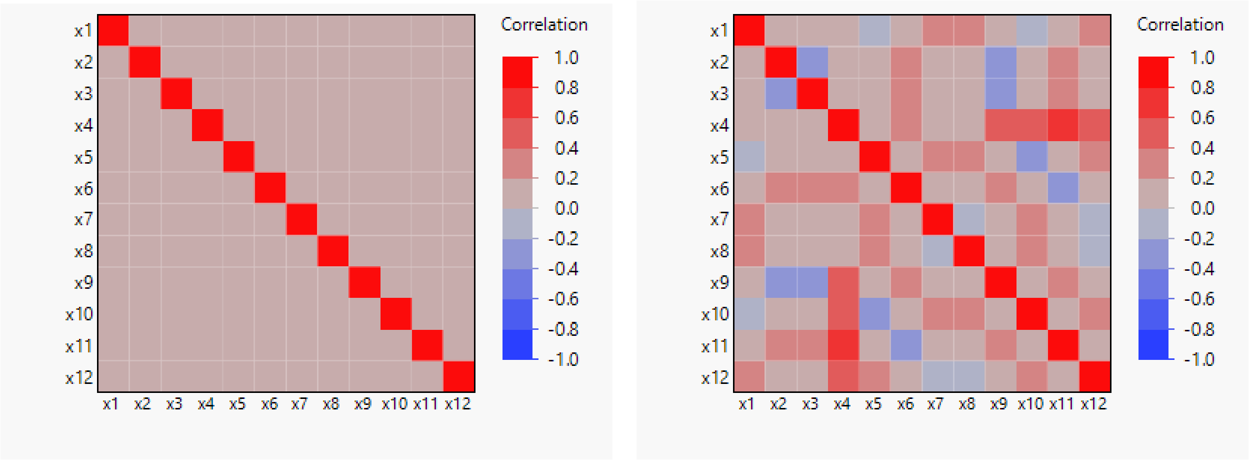
Though many other criteria have been proposed (e.g., Jones et al. 2008, 2009), none have been found to be clearly superior, in general, to $E(s^2)$-optimal designs (Marley and Woods 2010; Weese et al. 2015). However, Weese et al. (2017) developed the constrained $Var(s+)$ criterion, which was shown via simulation to be superior to $E(s^2)$-type designs when effect directions are known, while similar in performance to $UE(s^2)$ when the effect directions were misspecified. Recently, some theoretical results have corroborated the advantage of $Var(s+)$-type designs under a known sign vector (Stallrich et al. 2024). Constrained $Var(s+)$ designs minimize

$$Var(s) = UE(s^2) - UE(s)^2 \tag{2}$$

$$s.t. \quad \frac{UE^*(s^2)}{UE(s^2)} > c \text{ and } UE(s) > 0, \tag{3}$$

where $UE^*(s^2)$ is the $UE(s^2)$ value for the $UE(s^2)$-optimal design. Generally, $c = 0.8$ has been found to work well.

In this paper, we compare Orthogonal Arrays with SSDs constructed using the constrained $Var(s+)$ criterion. The OAs are constructed using `DoE.base`, an R package due to Grömping (2018), while the constrained $Var(s+)$ designs are constructed using the algorithm described in Weese et al. (2017). In Figure 1, we show a comparison of an OA with a constrained $Var(s+)$ design. While the OA evidently has no correlation between columns, for the SSD there are many nonzero off-diagonals, and more often than not those nonzero entries are greater than 0 as required in the $Var(s+)$ criterion.



(a) (b)

Figure 1: Correlation color plot for two 12-factor designs. (a) An OA with $n = 16$; (b) A constrained $Var(s+)$ design with $n = 10$.

2.2 Analysis of Screening Designs

Because $n < k$, a unique least squares solution for $[\beta_0|\boldsymbol{\beta}]$ does not exist in the case of SSDs. In the literature, various analysis methods have been studied, including forward selection, Bayesian methods, model averaging, and regularization (e.g., Marley and Woods 2010; Draguljić et al. 2014; Weese et al. 2015). Phoa et al. (2009) proposed to use the Dantzig selector and ensuing studies (e.g., Marley and Woods 2010; Weese et al. 2015) showed that it performed well compared to other methods. In this work, we use a related approach, the Lasso (Tibshirani 1996), which seems to perform similarly (Draguljić et al. 2014) and is more mathematically tractable. To estimate the main effect model in (1), the Lasso chooses $[\beta_0, \boldsymbol{\beta}]$ which minimizes

$$\frac{1}{2n}\|Y - \beta_0\mathbf{1} - D\boldsymbol{\beta}\|_2^2 + \lambda\|[\beta_0, \boldsymbol{\beta}]\|_1, \tag{4}$$

for a given regularization parameter λ. This parameter controls how much the parameters are shrunk toward 0. Equation (4) produces an OLS solution if $\lambda = 0$, while as $\lambda \to \infty$, $[\beta_0, \boldsymbol{\beta}] \to 0$. We will discuss our λ selection strategy in the next section. In our work, we actually solve a slightly different version of the Lasso, with centered Y, call it Y_c, and centered and scaled D, D_{cs}. This eliminates the need to model the intercept, and also ensures that each design column has zero mean and equal length. The result of this centered and scaled Lasso we denote $\hat{\boldsymbol{\beta}}_{cs}$.

For the analysis of orthogonal arrays, ordinary least squares are available since $n > k$. However, for consistency of comparisons, we used the Lasso to analyze the OAs as well.

3 Methodology

The primary objective of this paper is to quantify the difference between the screening effectiveness of orthogonal arrays and supersaturated designs. To do this, we will investigate a suite of designs, ranging from designs with a relatively small number to a fairly large number of factors, across a number of effect sparsity assumptions. Each comparison is between an OA and SSD with the same number of factors, and so we investigate the difference in effectiveness when the effect sizes are the same but also when the effect sizes for the SSD are larger than those of the OA. Though this is somewhat unconventional, one could imagine a more or less aggressive [-1,1] coding scheme in which numerical factors are transformed to result in larger or smaller effect sizes. Our purpose is to learn about the relative size of effects that each type of design can reliably detect. We use both traditional simulation and a new criterion for screening design quality (Stallrich et al. 2024) based on the probability of perfect sign recovery for the Lasso.

3.1 Simulation

From model (1), we have that D is either an OA or SSD with entries ± 1, where each column represents a factor and each row a treatment. Without loss of generality, we assume $\sigma = 1$, making $\boldsymbol{\beta}$ the signal-to-noise ratio. Let $A(\boldsymbol{\beta})$ denote the set of active columns in D, so that $A(\boldsymbol{\beta}) = \{\beta_j | \beta_j \neq 0, j = 1, 2, \ldots, k\}$, while $I(\boldsymbol{\beta})$ represents the inactive columns, and is $I(\boldsymbol{\beta}) = \{\beta_j | \beta_j = 0, j = 1, 2, \ldots, k\}$. A screening design paired with an analysis method that provides an accurate estimate of $A(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$, call them $\hat{A}(\boldsymbol{\beta})$ and $\hat{I}(\boldsymbol{\beta})$, is considered to be an effective design. (Note that we suppress full notation below and refer to these sets as A and I.) For the simulation study, we consider two metrics: true positive rate (TPR) and false positive rate (FPR). TPR is defined as the proportion of truly active factors correctly estimated as active, that is $n(A \cap \hat{A})/n(A)$, where $n(A)$ denotes the number of elements in set A. A large TPR indicates that the design is consistently identifying the truly active factors. On the other hand, FPR represents the proportion of truly inactive factors that are mistakenly estimated as active, $n(I \cap \hat{A})/n(I)$. An elevated FPR indicates a screening approach that isn't careful enough in its assignment of factors as active.

An important aspect of using the Lasso to estimate A and I is the choice of λ. For the analysis of a single experiment, a holistic analysis would include consideration of a profile plot as in Figure 5. But in a simulation, we must clearly and automatically specify $\hat{A}$ and $\hat{I}$ in each iteration, which means we need a strategy to determine λ. Cross validation is a general approach, but it may not work well in a small, highly structured experiment. Instead, we use the Akaike Information Criterion (AIC) to determine λ as follows: We consider a large grid of λ values uniformly placed from 0.0111 to λ_0, where $\lambda_0 = \max\left|D_{cs}^t Y_c\right|$ (see Weese et al. 2021). For each λ, then, we obtain the Lasso estimate and set $\hat{\beta}_j = 0$ for all $\left|\hat{\beta}_j\right| < \gamma$, where γ is chosen as discussed below. Then, use OLS to refit the models using the columns of D corresponding to the nonzero $\hat{\beta}_j$ and calculate the AIC of each. The $\hat{A}$ and $\hat{I}$ corresponding to the λ with minimum AIC is the screen of choice.

Note that γ is an additional tuning parameter. Weese et al. (2017) followed an idealistic approach by setting γ equal to σ. Conversely, Phoa et al. (2009) adopted a data-driven approach, by setting $\gamma = r\max(\left|\hat{\beta}_j\right|)$ where $r = 10\%$. Weese et al. (2021) suggested using r values of either 10%, 25% or 50%. The value of r value involves a trade-off. If r is large, more factors from $A(\boldsymbol{\beta})$ will be classified as $\hat{I}(\boldsymbol{\beta})$, while smaller r means that fewer factors will be zeroed out and thus included in $\hat{I}$. In the present study, the focus is on the relative values of TPR and FPR for the two types of designs, so the specific value chosen for γ is of less importance. In this work, we opt for $r = 0.5$ based on informal testing.

In our simulations, we compared the designs across a range of scenarios, defined by the combination of the following variables:

- Number of factors. Details are provided in Table 1, but we considered k as small as 7 and as large as 60. For each k, we compared an SSD where $n < k$ to an OA where $n > k$.

- Sparsity. Assuming ceil() rounds up to the nearest integer, we varied the number of active factors a such that $a = \{\text{ceil}(0.2k), \text{ceil}(0.25k), \text{ceil}(0.3k)\}$.

- Relative effect size. We considered four effect size scenarios:

- For the a active effects, both the OA and SSD had effect sizes of 0.5.
- For the a active effects, both the OA and SSD had effect sizes of 1.
- For the a active effects, the OA had an effect sizes of 1 while the SSD had effect sizes of 1.5.
- For the a active effects, the OA had an effect sizes of 0.5 while the SSD had effect sizes of 1.

Table 1: Designs for simulation

Small			*Medium*			*Large*		
k	**n (SSD)**	**n (OA)**	**k**	**n (SSD)**	**n (OA)**	**k**	**n (SSD)**	**n (OA)**
7	6	12	12	10	16	40	37	44
8	6	12	23	18	24	45	43	48
9	7	12	24	20	28	50	46	52
10	8	16	34	27	36	60	50	64

As an example, consider the following simulation setting: 12 factors with SSD $n = 10$ and OA $n = 16$; $a = \text{ceil}(0.2k) = 3$; and SSD effect size of 1 and OA effect size of 0.5. In this case we are comparing the effectiveness of the SSD when it has six fewer runs than the OA but its effect sizes are twice as large.

In the detailed simulation protocol that follows, note that we simulate from model (1) with $\beta_0 = 0$. Each simulation scenario provides: an OA and an SSD, with design matrices D_{OA} and D_{SSD}, respectively; a; and values f_{OA} and f_{SSD} for β_j's corresponding to active factors. Given a particular simulation scenario, in each of n_{iter} iterations,

1. Randomly select a columns from D_{OA} and D_{SSD} and assign f_{OA} and f_{SSD} to all $\beta_j \in A(\boldsymbol{\beta})$ and 0 to all $\beta_j \in I(\boldsymbol{\beta})$.

2. Assign each nonzero β_j a $+1$ or -1, with equal probability.

3. Simulate Y_{OA} from model (1) and X_{OA}; simulate Y_{SSD} from model (1) and X_{SSD}.

4. Center Y_{OA} and center and scale the X_{OA}; center Y_{SSD} and center and scale the X_{SSD}.

5. For both OA and SSD, using the grid of λ described above, fit the Lasso and obtain $\hat{\beta}_\lambda$ for each λ.

6. For both OA and SSD, for each $\hat{\beta}_\lambda$, check if $\left|\hat{\beta}_j\right| < \gamma$, where $\gamma = 0.5 * (max(\left|\hat{\beta}_\lambda\right|))$; if $\left|\hat{\beta}_j\right| < \gamma$ set $\hat{\beta}_j = 0$.

7. For both OA and SSD, for each λ, use the columns of D associated with the non-zero $\hat{\beta}$ to fit an OLS model to Y. Calculate the AIC for each model and select the one with minimum AIC value. The β's associated with the non-zero β's from the chosen model are included in $\hat{A}$, while the rest are assigned to $\hat{I}$.

Based on $\hat{A}$ and $\hat{I}$, along with A and I, we compute TPR and FPR as $\frac{1}{n_{\text{iter}}}\sum_{i=1}^{n_{\text{iter}}} n(A \cap \hat{A})_i / n(A)_i$ and $\frac{1}{n_{\text{iter}}}\sum_{i=1}^{n_{\text{iter}}} n(I \cap \hat{A})_i / n(I)_i$ respectively. In our work, we take $n_{\text{iter}} = 1000$.

3.2 Probability of perfect sign recovery method

An important downside of using simulation to compare designs is the need to specify the γ threshold. Its choice is somewhat arbitrary if the error variance is not available, though it affects TPR and FPR. To address this, Stallrich et al. (2024) introduced a method to compute the probability of perfect sign recovery for a screening design under the Lasso and the normal-theory statistical model given in (1), and Young et al. (2023) illustrate its use by showing its capability to improve reproducibility in the comparison of screening designs. Given the true signs $z = \text{sign}(\boldsymbol{\beta})$ and the estimated signs, $\hat{z} = \text{sign}(\hat{\boldsymbol{\beta}}_{cs})$, the criterion ranks designs based on $\phi_\lambda(D|\boldsymbol{\beta}) = P(\hat{z} = z)$. The criterion is based on the joint probability of two independent events, both a function of normal random vectors. These events are derived from the Lasso's Karush-Kuhn-Tucker conditions; details can be found in Stallrich et al. (2024).

Computing ϕ_λ requires knowledge of the entire model, including $\boldsymbol{\beta}$ along with the Lasso parameter λ. Thus, on its own it can't be used to evaluate designs. Stallrich et al. (2024) proposed a series of relaxed criteria that are more practically useful. For instance, they present a criterion $\phi_\Lambda(D|\boldsymbol{\beta})$ which integrates ϕ_λ over $\log(\lambda)$, thus allowing the computation of the criterion without knowledge of λ. Ultimately, they provide $\Phi_\Lambda(D|a, \beta)$, which is the $\log(\lambda)$-integrated measure averaged over all supports of size a (that is, all sets of a factors chosen from k) and assuming that, for all active factors, $\beta = \frac{|\beta_j|}{\sigma} \geq 1$ so that it is at least as large as the noise. Thus, as long as we can assume an effect size β and a number of active effects of interest, along with sign vector z, Φ_Λ can be used to evaluate a design without knowledge of λ, A, or $\boldsymbol{\beta}$. For example, this criterion can be used to evaluate a design under the assumption that effect directions are known, in which case constrained $Var(s+)$ designs have been seen to have advantage. On the other hand, if no sign information is available, another criterion $\Phi_\Lambda^{\pm}(D|a, \beta)$ can be computed which averages Φ_Λ over all possible sign vectors. Note that these criteria are computationally intensive to compute. One way to reduce computational burden is to average not over all $\binom{k}{a}$ possible supports but a subset of these supports. For this and other computationally related discussion, see Section 4 of Stallrich et al. (2024). Note also that the values of Φ_Λ and $\Phi_\Lambda^{\pm}$ don't have a clear interpretation on their own, since they are average probabilities, integrated over $\log(\lambda)$. However, they are useful to compare designs because larger values of the measures mean the design is more effective at recovering the true signs of the factors.

Because the probability method is more stringent (perfect sign recovery, instead of TPR and FPR), the effect sizes necessary to give meaningful results are larger. So instead of the four effect size scenarios from Section 3.1, we use the following:

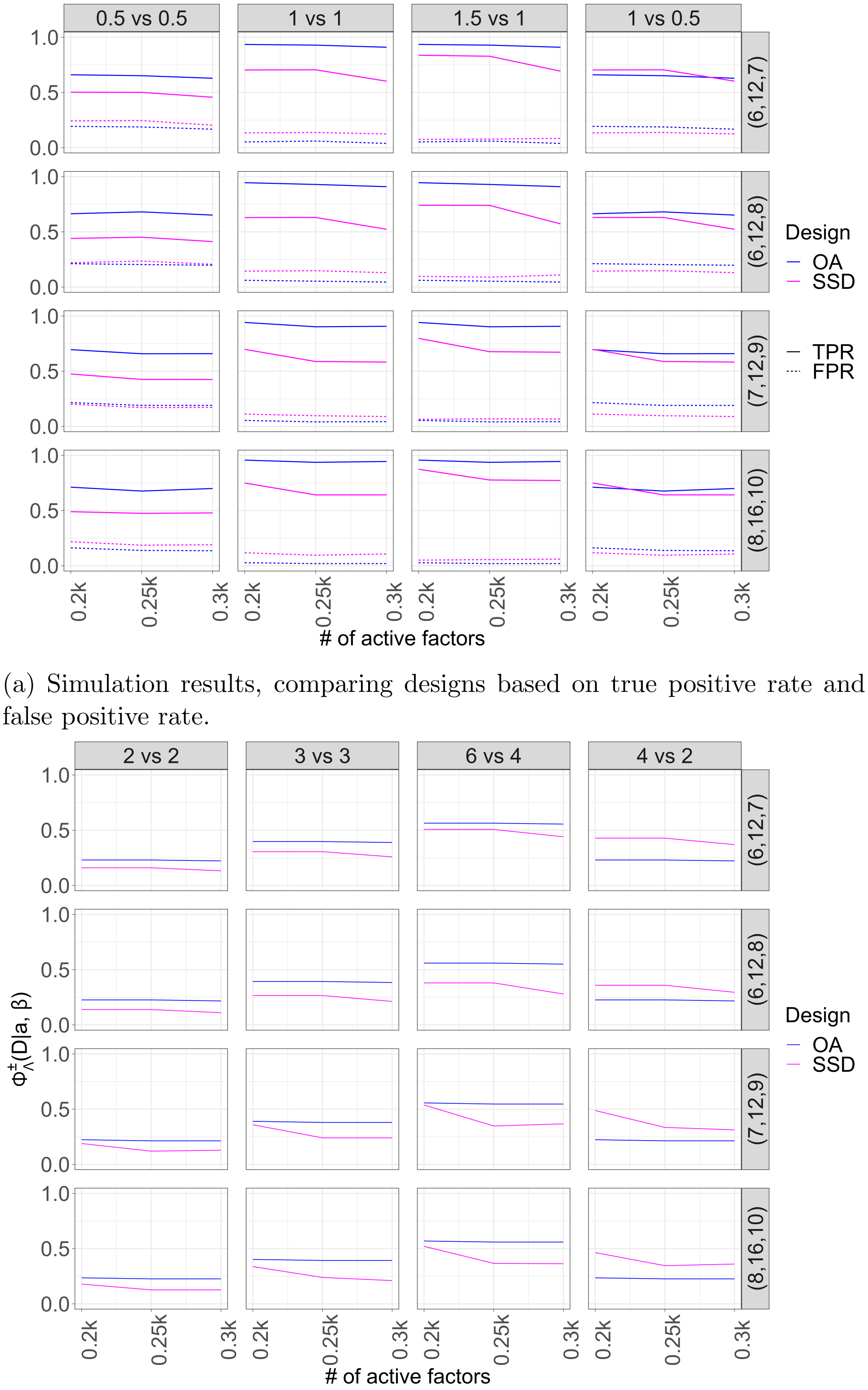
- For the a active effects, both the OA and SSD had effect sizes of 2.
- For the a active effects, both the OA and SSD had effect sizes of 3.
- For the a active effects, the OA had an effect sizes of 4 while the SSD had effect sizes of 6.
- For the a active effects, the OA had an effect sizes of 2 while the SSD had effect sizes of 4.

For this probability method, we evaluated the same set of designs (Table 1) across the same set of sparsity levels.

4 Results

In this section, we present the results of our simulation and probability calculations, for small, medium, and large-sized screening designs as given in Table 1.

Figure 2 presents the simulation and probability of perfect sign recovery results for small k. Although the effect magnitudes in the probability of perfect sign method are larger than those for the simulation, the ratio of the SSD to OA effect sizes are the same. Consider the first column and the first row of Figure 2a as an example. When $a = \text{ceil}(0.2k) = 2$, the 6-run, 7-factor SSD has a true positive rate of approximately 50%, while the 12-run, 7-factor OA is at about 70%. The SSD appears to have a slightly larger false positive rate, at about 25% vs. around 20% for the OA. This indicates that out of the $7 - 2 = 5$ inactive factors, about 25% for SSD (20% for OA) were indicated as active. Clearly, if resources were no object, the OA would be preferred in this scenario. Now consider the last column in the first row, where the SSD effect sizes are twice that of the OA. Here we see that the true positive rates are similar between the designs, with the SSD improving slightly over the OA in terms of false positive rate. Zooming out, we see that for these small designs, effect sizes for the SSD must be about twice the size of the effect sizes for the OA before the SSD has similar or better performance. For the probability of perfect sign recovery method (Figure 2b), the conclusions are much the same. That is, when effect sizes are equal, or even at a ratio of 1.5, the OA's are preferred. When the effect size ratio is 2, it is clear that the SSD performs better. We again note that the specific values of the probability method are less meaningful than the relative values of the criterion in comparing designs.

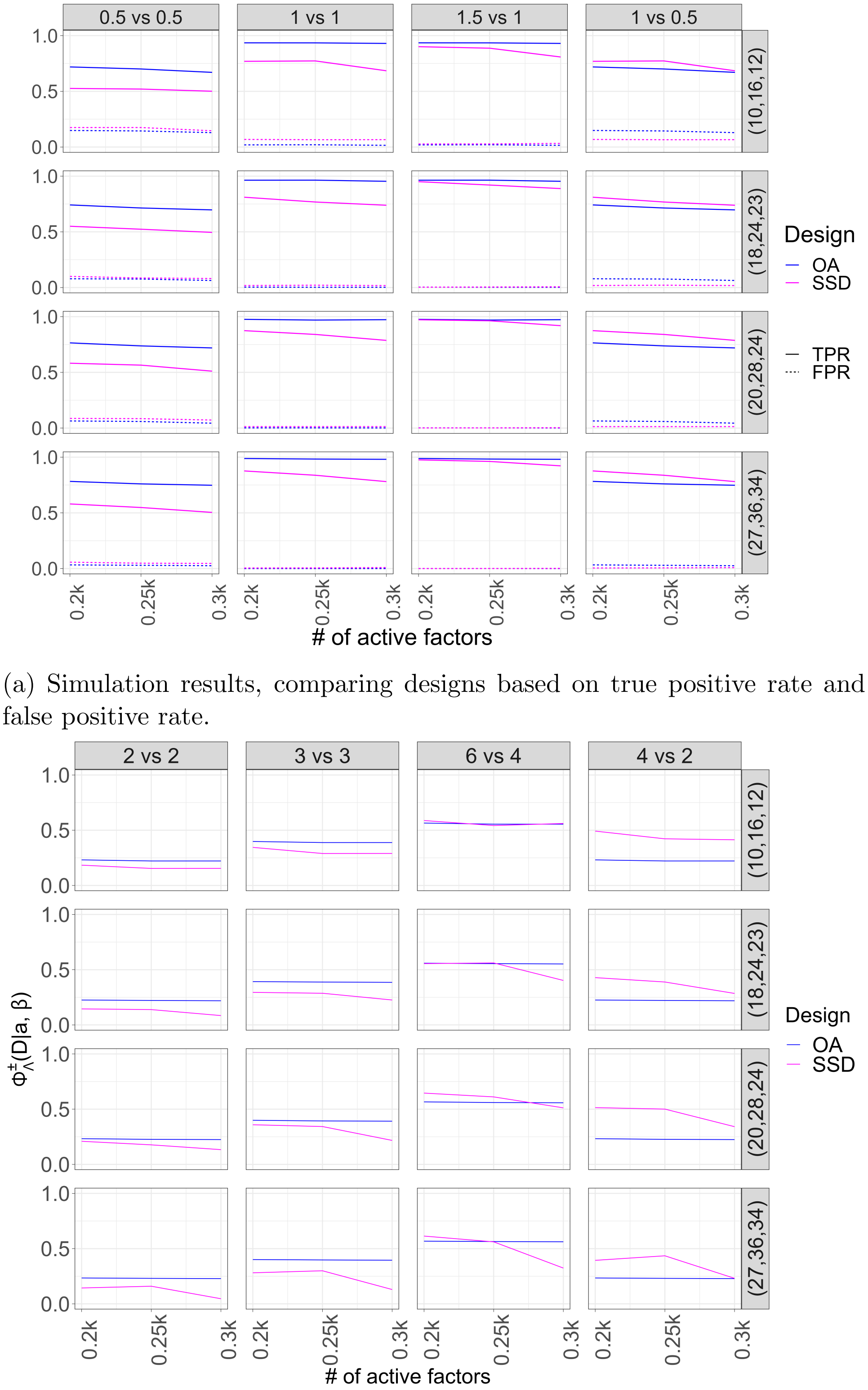


(a) Simulation results, comparing designs based on true positive rate and false positive rate.

(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}|\, a, \beta)$.

Figure 2: For small screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: (SSD β vs. OA β). The designs, denoted in the (SSD n, OA n, k) format, are shown along the right-hand side.

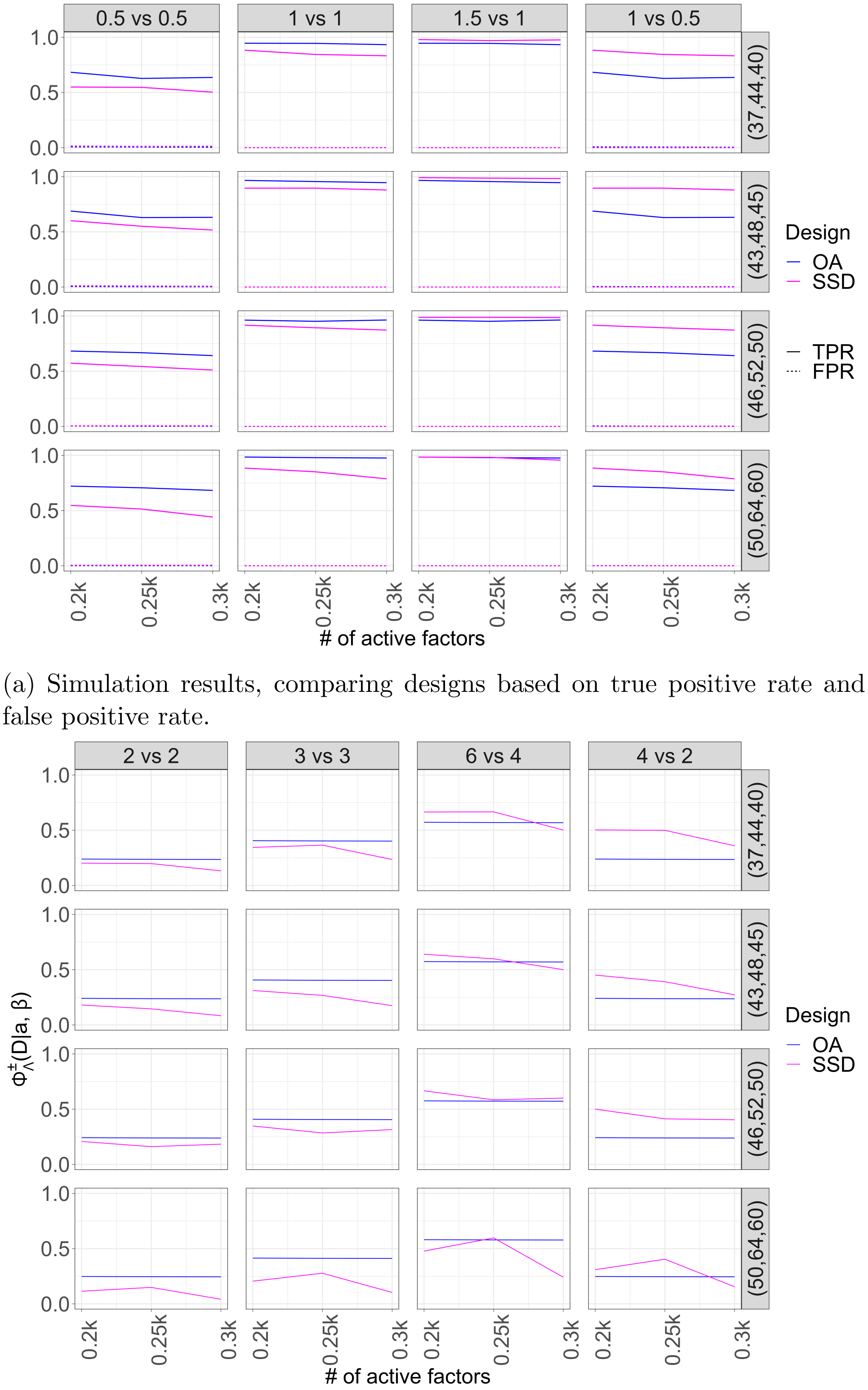
For medium (Figure 3) and large (Figure 4) screening designs, the story is slightly different. That is, while it is clear that OAs are preferred for equal effect sizes, when the effect size for the SSD is 1.5 that of the effect size for the OA, the designs perform fairly similarly, and when the ratio is 2 it is clear that SSDs are preferred. This is consistent across both the simulation and probability of perfect sign recovery results.



(a) Simulation results, comparing designs based on true positive rate and false positive rate.

(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}|\, a, \beta)$.

Figure 3: For medium-sized screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: (SSD β vs. OA β). The designs, denoted in the (SSD n, OA n, k) format, are shown along the right-hand side.



(a) Simulation results, comparing designs based on true positive rate and false positive rate.

(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}|\, a, \beta)$.

Figure 4: For large screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: (SSD β vs. OA β). The designs, denoted in the (SSD n, OA n, k) format, are shown along the right-hand side.

5 Discussion and Conclusion

The purpose of this study was to compare the performance of supersaturated designs (SSDs) against the better-known Orthogonal Arrays (OAs), in terms of their ability to screen important factors. We investigated this question using two methods of evaluation: traditional statistical simulation and a new probability of perfect sign recovery method for the Lasso. We performed the comparisons for small, medium, and large screening designs, several sparsity levels, and three effect size ratios. A key aspect of our scenarios is that in some cases we compared OAs with smaller effect sizes to SSDs with larger effect sizes, in order to provide an assessment of conditions under which True Positive Rates were similar. These differential effect sizes might conceptually correspond to, for instance, situations in which numerical factors are transformed from different intervals in their natural units to $[-1, 1]$, inducing different effect sizes.

As expected, we found that the OAs exhibited better screening performance than the SSDs when the underlying model had the same sized effects. When the data obtained by the SSD was associated with effect sizes 1.5 times those of the OA, the OAs were still preferable for small designs, while the performance was comparable between the two designs for medium and large screening designs. When the effect sizes were twice as large for the SSDs, for small designs the simulation results suggested that the two types of designs were similar in performance, while the probability method indicated a preference for the SSDs. For medium and large designs, when the effect size ratio was 2, both approaches suggest that the SSDs will perform better than the OAs.

Clearly, in a typical screening experiment, the effect sizes are not known in advance; thus, our results don't necessarily provide direct guidance to practitioners. However, our work sheds light on the *relative* performance of two different types of screening designs. In particular, our work gives insight as to how much more effective an orthogonal array will be compared to a supersaturated design that has roughly 6-10 fewer experimental runs. In general, you need effect sizes at least 1.5 times larger for the SSD to overcome the fact that it has fewer runs. The scope of our work does not include larger OAs or smaller SSDs, but for the types of designs that we considered, we would suggest that OAs should be preferred to SSDs for small screening designs, unless effect sizes are expected to be large and sparse. For medium and large designs, the differences are less pronounced, though still evident.

For certain applications in which effect sizes, in units of σ, can be anticipated, our simulation/probabilistic framework can provide more direct information about necessary effect sizes for a successful experiment. In general we see that our simulation results suggest relatively large True Positive Rates even for effect sizes of 1σ or 1.5σ. In contrast, even for effect sizes as large as 4σ or 6σ, the probability of perfect sign recovery approach paints a dimmer picture. There are at least two reasons for this discrepancy. First, the probabilistic framework's criterion is *perfect* sign recovery. Thus, in the presence of several active effects an analysis could have a fairly large True Positive Rate, with a small False Positive Rate, but still fail to screen perfectly. Secondly, the simulation approach uses a threshold parameter γ while the probability approach does not. Using a threshold prevents the parameters with small estimates being included as important effects.

Finally, we note that the use of a threshold γ is really only necessary to facilitate simulations. In a one-off analysis of experimental data, a preferred analysis approach would privilege the subjective evaluation of a Lasso profile plot, as in Figure 5, focusing on the terms that take longer to shrink to 0. Of course, an analysis using thresholding might also be used in a secondary manner.

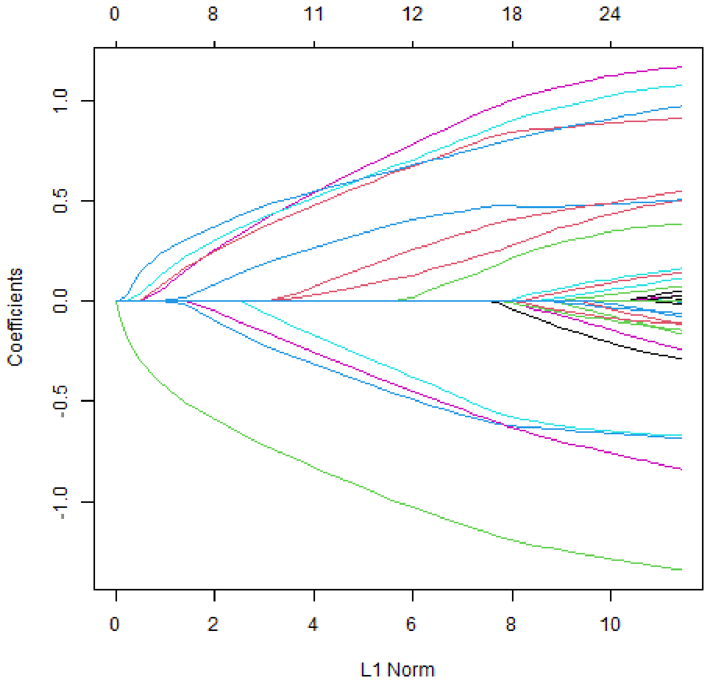


Figure 5: An example Lasso solution path. λ decreases from ∞ to 0 along the solution path from left to right.

Acknowledgements

The authors are grateful to Maria Weese and Steve Wright, for their feedback on this work.

The authors have no funding support to acknowledge.

References

Booth, K. H. and Cox, D. R. (1962), "Some systematic supersaturated designs," *Technometrics*, 4, 489–495.

Bose, R. C. and Bush, K. A. (1952), "Orthogonal arrays of strength two and three," *The Annals of Mathematical Statistics*, 23, 508–524.

Bush, K. A. (1952), "Orthogonal arrays of index unity," *The Annals of Mathematical Statistics*, 426–434.

Carpinteiro, J., Quintana, J., Martınez, E., Rodrıguez, I., Carro, A., Lorenzo, R., and Cela, R. (2004), "Application of strategic sample composition to the screening of anti-inflammatory drugs in water samples using solid-phase microextraction," *Analytica chimica acta*, 524, 63–71.

Dejaegher, B. and Vander Heyden, Y. (2008), "Supersaturated designs: set-ups, data interpretation, and analytical applications," *Analytical and bioanalytical chemistry*, 390, 1227–1240.

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

Georgiou, S. D. (2014), "Supersaturated designs: A review of their construction and analysis," *Journal of Statistical Planning and Inference*, 144, 92–109.

Grömping, U. (2018), "R package DoE. base for factorial experiments," *Journal of Statistical Software*, 85, 1–41.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999), *Orthogonal arrays: theory and applications*, Springer Science & Business Media.

Jones, B., Lin, D. K., and Nachtsheim, C. J. (2008), "Bayesian D-optimal supersaturated designs," *Journal of Statistical Planning and Inference*, 138, 86–92.

Jones, B. and Majumdar, D. (2014), "Optimal supersaturated designs," *Journal of the American Statistical Association*, 109, 1592–1600.

Jones, B. A., Li, W., Nachtsheim, C. J., and Kenny, Q. Y. (2009), "Model-robust supersaturated and partially supersaturated designs," *Journal of Statistical Planning and Inference*, 139, 45–53.

Jridi, M., Lassoued, I., Kammoun, A., Nasri, R., Nasri, M., Souissi, N., et al. (2015), "Screening of factors influencing the extraction of gelatin from the skin of cuttlefish using supersaturated design," *Food and Bioproducts Processing*, 94, 525–535.

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

Marley, C. J. and Woods, D. C. (2010), "A comparison of design and model selection methods for supersaturated experiments," *Computational Statistics & Data Analysis*, 54, 3158–3167.

Mee, R. (2009), *A comprehensive guide to factorial two-level experimentation*, Springer Science & Business Media.

Mee, R. W., Schoen, E. D., and Edwards, D. J. (2017), "Selecting an orthogonal or nonorthogonal two-level design for screening," *Technometrics*, 59, 305–318.

Owen, A. B. (1992), "Orthogonal arrays for computer experiments, integration and visualization," *Statistica Sinica*, 439–452.

Phoa, F. K., Pan, Y.-H., and Xu, H. (2009), "Analysis of supersaturated designs via the Dantzig selector," *Journal of Statistical Planning and Inference*, 139, 2362–2372.

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

Rao, C. R. (1947), "Factorial experiments derivable from combinatorial arrangements of arrays," *Supplement to the Journal of the Royal Statistical Society*, 9, 128–139.

Satterthwaite, F. (1959), "Random balance experimentation," *Technometrics*, 1, 111–137.

Stallrich, J. W., Young, K., Weese, M. L., Smucker, B. J., and Edwards, D. J. (2024), "An Optimal Design Framework for Lasso Sign Recovery," https://arxiv.org/abs/2303.16843, version 2, revised 2024-03-15.

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

Weese, M. L., Edwards, D. J., and Smucker, B. J. (2017), "A criterion for constructing powerful supersaturated designs when effect directions are known," *Journal of Quality Technology*, 49, 265–277.

Weese, M. L., Smucker, B. J., and Edwards, D. J. (2015), "Searching for powerful supersaturated designs," *Journal of Quality Technology*, 47, 66–84.

Weese, M. L., Stallrich, J. W., Smucker, B. J., and Edwards, D. J. (2021), "Strategies for supersaturated screening: Group orthogonal and constrained var (s) designs," *Technometrics*, 63, 443–455.

Wu, C. (1993), "Construction of supersaturated designs through partially aliased interactions," *Biometrika*, 80, 661–669.

Xu, H., Phoa, F. K., and Wong, W. K. (2009), "Recent developments in nonregular fractional factorial designs," .

Young, K., Weese, M. L., Stallrich, J. W., Smucker, B. J., and Edwards, D. J. (2023), "A Graphical Comparison of Screening Designs using Support Recovery Probabilities," https://arxiv.org/abs/2311.12685.

Zarkadas, S. and Besseris, G. (2023), "Using Lean-and-Green Supersaturated Poly-Factorial Mini Datasets to Profile Energy Consumption Performance for an Apartment Unit," *Processes*, 11, 1825.