Strategies for Supersaturated Screening: Group Orthogonal and $Var(s+)$ Designs

Maria L. Weese[1], Jonathan W. Stallrich[2], Byran J. Smucker[3], and David J. Edwards[4]

[1]Department of Information Systems & Analytics, Miami University, Oxford, OH
[2]Department of Statistics, North Carolina State University, Raleigh, NC
[3]Department of Statistics, Miami University, Oxford, OH
[4]Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA

Abstract

Despite the vast amount of literature on supersaturated designs, there is a scant record of their use in practice. We contend this imbalance is due to the designs' inabilities to meet practitioners' analysis expectations and the existing literature's lack of clearly stated expectations. To address this issue we discuss and compare two recent SSDs that pair a design construction method with a particular analysis method. Group orthogonal supersaturated designs (Jones et al. 2019), when paired with our new, modified analysis, are shown to have high power even with many active factors. $Var(s+)$ designs (Weese et al. 2017), when paired with the Dantzig selector, are recommended when effect directions can be reasonably specified in advance; this strategy reasonably controls type 1 error rates while still identifying a high proportion of active factors. The construction of both designs is less intuitive than classical supersaturated designs that focus almost solely on traditional measures of near-orthogonality, and forces reflection on current best practices.

Keywords: Dantzig selector, GO-SSD, Orthogonality, Power, Sparsity, Type 1 Error

1 Introduction

"*I think it is perfectly natural and wise to do some supersaturated experiments.*"–*John Tukey (from a discussion of Satterthwaite 1959)*

A screening experiment is tailored to understanding a complex, expensive system by efficiently identifying the system's most influential factors. Supersaturated designs (SSDs) are posited to effectively screen factors even when the number of runs is less than the number of considered factors. SSDs were introduced by Satterthwaite (1959) and initiated into the mainstream experimental design literature several decades later (Lin 1993; Wu 1993). The body of work that composes the research area today is impressive (see the review by Georgiou 2014) but despite all this work, there is a scant record of SSDs in practice. Among the few examples we have found are Carpinteiro et al. (2004), Jridi et al. (2015), and other applications in Dejaegher and Vander Heyden (2008). To help convince practitioners of their value, this paper reassesses the pairing of an SSD's construction and analysis.

Our perspective is that screening is the first stage of a sequential experimental procedure and involves the choice of a design and analysis combination. The analysis aims to classify factors into those that should be further studied (i.e. "potentially active") and those that can be ignored (i.e. "inactive"). Factors classified as potentially active should include all factors that most influence the response, but may also include those that are marginally active. Later stages of the sequential procedure target further understanding of the potentially active factors.

The successful use of SSDs is predicated on the assumption of sparsity, which argues that most process variation will be driven by a few factors. In particular, the practitioner must first posit a relatively simple statistical model that defines how a factor can influence the response. This influence is characterized by one or more parameters and an analysis method is chosen for model/variable selection and subsequent parameter estimation. The most basic screening model includes just the linear main effects:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where n is the number of runs, $\epsilon_i \sim N(0, \sigma^2)$, x_{ij} is the j-th factor's setting for run i, and β_j is an unknown parameter. The model is equivalently written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the $n \times (k+1)$ model matrix, $\boldsymbol{\beta}$ is a $(k+1)$-vector of model parameters, and $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are n-vectors, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The j-th factor is considered active if $|\beta_j| > t$ for some threshold $t \geq 0$.

Model (1) assumes each factor's effect on y_i is linear and independent of the levels of the other factors, i.e. there are no interaction effects. Even though model (1) is likely inaccurate, it is reasonable to believe it can explain much of the response's variation. There are screening designs that entertain more complicated models either by allowing more parameters to be estimated (e.g. Draguljić et al. 2014) or by making the analysis robust to model misspecification (e.g. Li and Nachtsheim 2000; Loeppky et al. 2007; Smucker and Drew 2015; Shi and Tang 2019). Such designs tend to require more runs than those targeting estimation of model (1), especially if k is large.

A screening analysis pairs a factor classification rule with estimators $\hat{\beta}_j$. The level of uncertainty associated with $\hat{\beta}_j$ for SSDs, and the assumption of a follow-up experiment, justifies designating promising factors as "potentially active" instead of simply "active". The screening classification rule depends on the experimenter's willingness to risk classifying an inactive factor as potentially active (type 1 error) and classifying a truly active factor as inactive (type 2 error). Our type 1 error definition differs slightly from convention because the decision we are making is whether or not to conduct further experimentation on a factor. Factors declared as inactive are assumed to be entirely removed from future consideration, so the definition of type 2 error, and hence the equivalent definition of power, is conventional. Simultaneously minimizing type 1 error and maximizing power becomes challenging as n decreases because the uncertainty of our estimates increases. A trade-off must be made that depends on the budget for future experimentation and the overall goals. Is it important that few, if any, active factors are omitted, even at the expense of more type 1 errors? Or is the goal to to identify as many active factors as possible, while controlling the type 1 error more stringently? We argue that the best choice of SSD construction and analysis depends on this practitioner-specified trade-off.

The fundamental principle of experimental design is that the data collection procedure strongly influences an estimator's statistical properties, and hence factor classification. Knowledge of this relationship should be leveraged whenever possible to compare and rank designs; this is well understood for least-squares estimation. A linear model is said to be estimable if the least-squares estimator is unique, which holds if and only if $\mathbf{X}$ has full column rank. The estimator,

$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, has covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ and is unbiased if the model is correctly specified. Screening based on $\hat{\boldsymbol{\beta}}_{LS}$ is commonly done via hypothesis testing, which practitioners are comfortable with and has tractable power and type 1 error rates. In particular, $\hat{\boldsymbol{\beta}}_{LS}$ has good screening properties for a given design if $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ is "small." The ideal matrix, $\mathbf{X}^T\mathbf{X} = nI_n$, comes from regular and nonregular fractional factorial designs, having an $\mathbf{X}$ with settings ± 1 and mutually orthogonal columns. Such designs produce $\hat{\boldsymbol{\beta}}_{LS}$ with minimum variance, thereby maximizing power and minimizing the type 1 error. For SSDs, such an $\mathbf{X}$ cannot exist and, even worse, the main-effect model is not least-squares estimable.

Penalized regression estimation (e.g., LASSO (Tibshirani 1996) and the Dantzig selector (Candes and Tao 2007)) are popular alternatives to least-squares estimation. By penalizing the magnitude of their $\hat{\beta}_j$'s, these estimators are well-defined even when the main-effect model is not least-squares estimable. The $\hat{\beta}_j$'s are biased toward 0, but tend to have smaller variance than $\hat{\boldsymbol{\beta}}_{LS}$. Hence, they may have superior screening properties compared to screening based on least-squares. For example, suppose we have five factors and the first three are active, with $\beta_1 = \beta_2 = \beta_3 = 5$ and both $|\beta_4|, |\beta_5| \leq t$. A penalized estimator may give the estimates $\hat{\beta}_j = 1$ for $j = 1, 2, 3$ and $\hat{\beta}_j = 0$ for $j = 4, 5$. If the classification rule were to declare all factors having $|\hat{\beta}_j| < t < 1$ as inactive, the screening results would be perfect, but the estimators would be poor. Unfortunately, it is usually difficult to derive type 1 error and power analytically under such estimators so simulations are commonly used.

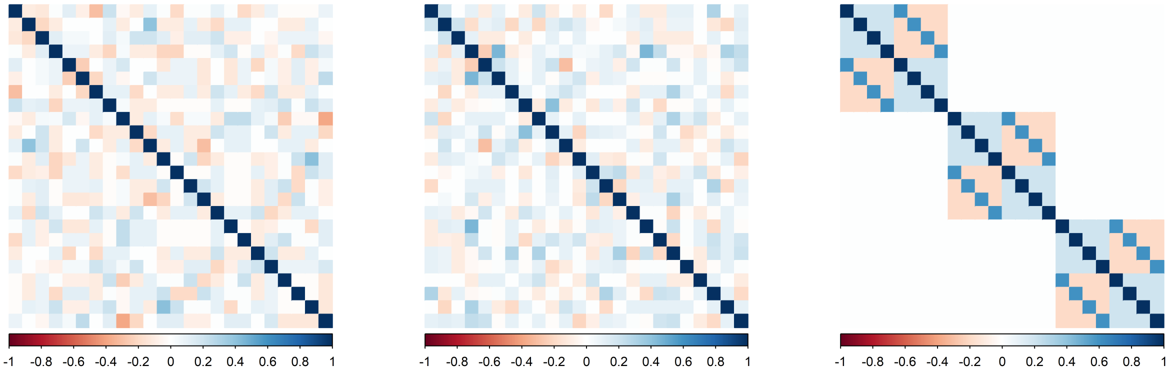
The lack of unique least-squares estimators with SSDs and no clear connections between $\mathbf{X}$ and penalized estimators makes it difficult to justify a screening criterion to rank potential SSDs. Instead, designs have been constructed via heuristic measures of orthogonality based on the off-diagonals of $\mathbf{X}^T\mathbf{X} = (s_{ij})$. For example, the $E(s^2)$-criterion forces $s_{0j} = 0$ and minimizes $E(s^2) = \binom{k}{2}^{-1} \sum_{1 \leq i < j \leq k} s_{ij}^2$. Such criteria intend to approximate the ideal structure $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_n$ as closely as possible, even though non-least squares estimation methods such as stepwise selection or penalized regression must be used. While these heuristic measures reasonably promote good screening properties, their connection to estimation is not as rigorous as for traditional screening designs based on least-squares (e.g., strength 2 or 3 orthogonal arrays). Because of this, there has been no clear consensus about the optimal pairing of SSD criteria and analysis strategy, which may partially explain why practitioners are hesitant to adopt the methodology.

Recently, two new SSD criteria have been developed and shown to improve over existing approaches in identifying potentially active factors. Weese et al. (2017) construct SSDs with the $Var(s+)$ criterion that forces the average off-diagonals of $\mathbf{X}^T\mathbf{X}$ to be positive, but not too large, while minimizing their variability. Through an extensive simulation study, the authors found the $Var(s+)$-optimal designs had higher power and smaller type 1 error compared to other SSDs when effect directions were known and when analysis was performed with the Dantzig selector. When effect directions were unknown, all of the SSDs in their study had equivalent performance. The superior performance is then tied to the analysis method and additional assumptions regarding $\boldsymbol{\beta}$. The group orthogonal SSDs (GO-SSDs) by Jones et al. (2019) create SSDs with a group factor structure along with extra fake factors that can produce a screening-independent estimate of the error variance. The design structure is paired with a two-stage least-squares analysis method capable of performing group and factor screening with high power for sparse $\boldsymbol{\beta}$ and ideal partitions of active factors across the groups.

Figure 1 shows the pairwise column correlations of SSDs constructed with the unbalanced $E(s^2)$ ($UE(s^2)$; to be discussed later), $Var(s+)$, and GO-SSD criteria for $n = 20$ and $k = 24$. In contrast with the $UE(s^2)$-optimal design, the $Var(s+)$ criterion tends to produce columns with more positive correlation. Figure 1(c) shows the group orthogonal structure of the GO-SSD, and that factors within a group have relatively high correlation. Table 1 compares several characteristics for the designs in Figure 1. While the $UE(s^2)$-optimal design has a smaller $UE(s^2)$ value, the $Var(s+)$ design has a smaller variance of the s_{ij}'s ($Var(s)$) and a noticeably larger average value of the s_{ij}'s ($UE(s)$), as intended. The $UE(s^2)$ value for the GO-SSD is nearly twice as large as that of the other two designs. All three designs exhibit similar average absolute column correlation (Mean $|r|$).

Table 1: Comparison of $n = 20$, $k = 24$ SSDs

Design	$UE(s^2)$	$UE(s)$	$Var(s)$	Mean $\|r\|$	Max $\|r\|$
$UE(s^2)$-optimal	5.813	0.120	5.809	0.097	0.414
$Var(s+)$	5.973	0.613	5.607	0.097	0.453
GO-SSD	9.600	0.480	9.385	0.078	0.600



(a) $UE(s^2)$

(b) $Var(s+)$

(c) GO-SSD

Figure 1: Correlation color plots of three $n = 20$, $k = 24$ SSDs: (a) $UE(s^2)$-optimal; (b) $Var(s+)$; and (c) GO-SSD. Blue represents positive correlation; red represents negative correlation

This article aims to convince practitioners of the potential value of certain pairings of SSDs and analysis strategies, and equip them with the knowledge and tools to appropriately use them. In Section 2 we summarize practitioners' perspectives about SSDs to better understand their concerns and/or misconceptions that have suppressed their use. Section 3 reviews traditional SSD construction and analysis recommendations and introduces our simulation protocol. The $Var(s+)$-criterion and its analysis with the Dantzig selector is discussed in section 4. Section 5 reviews GO-SSDs and presents an analysis method to maximize power and provide an indication that sparsity assumptions may be violated. Section 6 shows results from several simulation studies to more fully investigate $Var(s+)$-optimal designs paired with the Dantzig selector and the GO-SSDs paired with the new analysis method. Finally, in Section 7, we conclude the paper with a discussion of our results, how these results can help allay concerns, and practical advice for both the design and analysis of SSDs.

2 Practical Perspectives on Supersaturated Designs

The trepidation surrounding the use of SSDs is perhaps due to conflicting recommendations in the literature. In the article that develops the $E(s^2)$ criterion, Booth and Cox (1962) (pg. 489) state:

> We have no experience of practical problems where such designs are likely to be useful; the conditions that interactions should be unimportant and that there should be a few

> dominant effects seems very severe.

This directly contradicts the quote at the outset of this paper which is from a discussion of Satterthwaite (1959). These divergent perspectives continue in more recent work. For example, Georgiou (2014) (pp. 107) argues:

> In conclusion we can say that one should be very cautious when using any method for constructing, analyzing or generally using SSDs.

On the other hand, Gilmour (2006) (pp. 188) concludes:

> For situations where there really is no prior knowledge of the effects of factors, but a strong belief in factor sparsity, and where the aim is to find out if there are any dominant factors and to identify them, experimenters should seriously consider using supersaturated designs.

Marley and Woods (2010) is perhaps the most prominent paper giving practical advice on SSDs. They provide recommendations on SSD size and true model sparsity based on simulation power. They state that "the number of runs should be at least three times the number of active factors." They also assert that an SSD's level of saturation, k/n, should be less than 2. We revisit these results with additional simulations in Supplementary Materials 1 (Section 3), confirming their rule of thumb regarding the run size to active factor ratio, while noting that there appears to be a gradual reduction of effectiveness as k/n increases.

To assess the practical use of and concerns regarding SSDs, we devised an informal questionnaire to collect information from the greater design of experiments (DOE) community, using the authors' networks and social media. We received 63 responses to a twenty-item instrument, asking about current and past experience with DOE and SSDs (the questions are included in the supplementary material). While these results cannot be reasonably taken as representative of any more general, identifiable population, it was valuable in generating a list of possible reasons experimenters might hesitate to use SSDs. Section 1 of the Supplementary Material 1 contains the questionnaire results.

Of the 63 respondents, thirteen reported using SSDs in the past. Nineteen reported explanations of their concerns with these designs, and we have categorized them in the supplementary materials.

One of the most prominent issues that surfaced was that the designs and/or analysis methods lack sufficient power to detect important effects. We address this concern by studying two relatively new approaches that provide improved power compared to traditional supersaturated designs.

Another issue raised in the questionnaire is the perceived riskiness of using SSDs, the concern that such an experiment could be conducted with little information to show for it. Practitioners should be reminded that SSDs are only recommended during the initial stages in a sequence of experiments. This perspective reduces the riskiness of using the designs, while offering the possibility of increased overall experimental efficiency if the factor sparsity assumption holds.

Another prominent concern relates to a suspicion that the effect sparsity assumption will fail. There is ample empirical evidence that factor sparsity holds in many experimental settings (Li et al. 2006; Ockuly et al. 2017). In Li et al. (2006), they estimated that an average of 41% of factors were active, with a confidence interval whose upper end was 46%. We would expect much more sparsity, on average, for SSDs that investigate many factors about which little is known.

3 Background and Setting

3.1 Supersaturated Design Construction and Analysis

Most SSD criteria focus on optimizing heuristic measures of orthogonality with respect to the main-effect model information matrix $\mathbf{S} = \mathbf{X}^T\mathbf{X} = (s_{ij})$ where $i, j = 0, 1, \ldots, k$. We focus on two-level designs having $x_{ij} = \pm 1$, making $s_{ii} = n$. A design's proximity to orthogonality here is measured by a summary of the off-diagonal s_{ij}'s. Early work in the area (Lin 1993; Wu 1993) focused on constructing $E(s^2)$-optimal designs that require $s_{0j} = 0$ for $j \geq 1$ and minimize the average squared off-diagonal: $\frac{2}{k(k-1)} \sum_{1 \leq i < j \leq k} s_{ij}^2$. The unconditional $E(s^2)$-criterion, or $UE(s^2)$-criterion (Jones and Majumdar 2014; Weese et al. 2015) is similarly defined, but includes the s_{0j}^2 elements and so does not require $s_{0j} = 0$; that is, $UE(s^2) = \frac{2}{k(k+1)} \sum_{0 \leq i < j \leq k} s_{ij}^2$. Bayesian D-optimal designs (Jones et al. 2008) maximize the determinant of $\mathbf{X}^T\mathbf{X} + \mathbf{K}$ where $\mathbf{K} = \text{diag}(0, \tau \mathbf{I}_k)$ with τ representing prior information about the main effects. Georgiou (2014) provides a comprehensive review of SSD construction methods.

The literature is replete with proposed methods for SSD analysis. Several works (Marley and Woods 2010; Draguljić et al. 2014; Weese et al. 2015, 2017) have shown, through extensive simulation studies, that least squares-based procedures, such as forward selection, have poor screening properties. Other methods fare better, with the Dantzig selector most consistently excellent in terms of power and type 1 error (Phoa et al. 2009; Marley and Woods 2010; Chen et al. 2013; Draguljić et al. 2014; Weese et al. 2015, 2017; Drosou and Koukouvinos 2018). The Dantzig selector (Candes and Tao 2007) is a regularization method that constrains an ℓ_1-estimator:

$$\hat{\boldsymbol{\beta}}_{DS} = \arg\min_{\tilde{\boldsymbol{\beta}}} ||\tilde{\boldsymbol{\beta}}||_1 \text{ subject to } ||\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})||_\infty \leq \delta \ , \tag{2}$$

where $||\cdot||_\infty$ denotes the largest absolute element of the vector argument. In practice, estimates are generated for many values of $\delta \geq 0$, generating a profile plot of estimates. We strongly recommend screening decisions be made with respect to this profile plot (see Section 7), but this somewhat subjective process cannot be carried out in a power simulation study. We follow the non-graphical approach by Phoa et al. (2009) for automated model selection, but use the Bayesian information criterion (BIC) for model selection, which was utilized by Marley and Woods (2010):

1. Center $\boldsymbol{y}$, and center and scale the columns of $\mathbf{X}$ to have mean 0 and unit variance. Drop the intercept column from $\mathbf{X}$.

2. Solve (2) for d values of δ between 0 and $\max_j |\boldsymbol{x}_j^T\boldsymbol{y}|,\ j = 1, 2, \ldots, p$. Denote the d estimates by $\hat{\boldsymbol{\beta}}_{DS}(\delta)$.

3. For each $\hat{\boldsymbol{\beta}}_{DS}(\delta)$, set all $|\hat{\beta}_j(\delta)| < \gamma$ to 0 for some threshold $\gamma > 0$. Denote this by $\hat{\boldsymbol{\beta}}_{DS}(\delta, \gamma)$.

4. For each $\hat{\boldsymbol{\beta}}_{DS}(\delta, \gamma)$, calculate the least-squares estimates (also known as Gauss-Dantzig estimates) using only predictors with nonzero $\hat{\beta}_j(\delta, \gamma)$ and compute $BIC = n\ln{(SSE/n)} + k\ln(n)$ where SSE is the sum-of-squared errors.

5. For $\hat{\boldsymbol{\beta}}_{DS}(\delta, \gamma)$ with the smallest BIC from step 4, factors with nonzero $\hat{\beta}_j(\delta, \gamma)$ are classified as potentially active; otherwise a factor is inactive.

Centering in step 1 is important because otherwise the intercept parameter would be penalized. Scaling is needed so that the estimates are not influenced by the potentially unequal lengths of each centered column of $\mathbf{X}$. The choice of γ in step 3 is critical in the automated use of this Dantzig procedure. We will discuss and compare several strategies in Section 4.

Marley and Woods (2010) performed a power simulation study for $E(s^2)$-optimal and Bayesian D-optimal designs using the above Dantzig selector procedure and a Bayesian model averaging method. They found Bayesian D-optimal designs had slightly higher power and that the above Dantzig procedure was the best analysis method considered. Weese et al. (2015) performed a similar study but also included model-robust (Jones et al. 2009; Smucker and Drew 2015) and $UE(s^2)$-optimal designs. No clear winner emerged among the criteria. We perform our own power simulation approach with $Var(s+)$-optimal designs and the Dantzig selector in Section 4 and for GO-SSDs and the two-stage analysis in Section 5.

3.2 Simulation Protocols

Similar to Marley and Woods (2010) and Weese et al. (2017), we performed several power simulation studies to explore $Var(s+)$-optimal designs and GO-SSDs. We outline our basic simulation approach here, and note any modifications later when needed.

Several different design sizes were considered, denoted by (n, k): $(8, 12)$, $(12, 12)$, $(12, 24)$, $(16, 28)$, $(20, 24)$, $(24, 28)$, and $(40, 56)$. As explained in Section 5, GO-SSDs are not available for sizes $(8, 12)$, $(12, 24)$, and $(16, 28)$. We considered settings with large effect sizes, high sparsity, and favorable n/k ratios, as well as more challenging scenarios where the assumption of effect sparsity is violated. Our goal was to identify scenarios when the methods work well and when they break down. Model, or $\boldsymbol{\beta}$, sparsity, i.e., the number of active factors, was varied according to $0.25n$ (high sparsity; see findings of Marley and Woods 2010), $0.5n$, and $0.75n$ (low sparsity). The magnitude of the active effects were randomly generated from $Exp(1) + SN$, where SN, meaning signal-to-noise ratio, was set to either 1 or 3. These coefficients either remained positive (effect directions known), or their signs were randomly set to +1 or −1 with probability 0.5 (unknown effect directions). The magnitude of the inactive effects were generated by taking the absolute value of $N(0, 6^{-2})$ so that 99% of the inactive effects would be less than 0.5 and sufficiently bounded away from the simulated error variance $\sigma^2 = 1$. The signs of the inactive coefficients were assigned according to the simulation scenario (known/unknown). The responses were generated according to model (1). In total, we considered twelve scenarios in our main simulation study.

A total of 5000 iterations were performed for each simulation scenario, design size, and design/analysis combination. For each iteration, the active main effect factors were randomly assigned to the factor columns of $\mathbf{X}$ and a new set of factor effects was generated. We measured the quality of a design and analysis pairing according to power (proportion of active effects classified as potentially active) and type 1 error (proportion of inactive effects classified as potentially active). We also considered false discovery rate (FDR; the proportion of effects classified as potentially active that are inactive) and the average number of factors declared as potentially active.

4 $Var(s+)$ Designs and the Dantzig Selector

Weese et al. (2017) proposed the $Var(s+)$ criterion that minimizes the variance of the off-diagonal s_{ij}'s subject to some constraints. Specifically, the criterion value for a given design is:

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

where $UE^*(s^2)$ is the value for an approximately $UE(s^2)$-optimal design, $UE(s)$ is the average of the s_{ij}, and c is a specified efficiency that determines how near to $UE^*(s^2)$ the design is required to be. This criterion allows the s_{ij}'s to be, on average, more positive than those in the approximately $UE(s^2)$-optimal design, but with less variability. Because they are constructed algorithmically, the $Var(s+)$ designs in this paper are only approximately $Var(s+)$-optimal. We have no guarantee that any of the $Var(s+)$ designs, constructed using the coordinate exchange approach described in Weese et al. (2017), are optimal, so we abuse language slightly throughout the rest of the paper when we call them "optimal". We also note that Jones and Majumdar (2014) provided direct methods to construct $UE(s^2)$-optimal designs, though in this article we construct them algorithmically. We provide all designs from this article in the Supplementary Materials.

Weese et al. (2017) found that $Var(s+)$ designs were superior to $UE(s^2)$-optimal and Bayesian D-optimal designs when effect directions were correctly specified in advance, having higher power without elevating type 1 error. Without loss of generality, the effect directions are assumed to be positive. If a factor's effect is assumed negative, then the signs of the elements in the corresponding column of the $Var(s+)$-optimal design should be flipped prior to experimentation and then flipped back to their original signs after experimentation. Even when the effect directions were misspecified, the $Var(s+)$ designs fared no worse than the $UE(s^2)$ and Bayesian D-optimal SSDs.

$Var(s+)$ designs are not yet fully understood, but their effectiveness with known effect directions appears to be connected to the constraint $UE(s) > 0$, which apparently biases the estimates away from 0 and in the positive, known direction. The Dantzig selector appears to further aid in this amplification of estimates. Research is currently underway that seeks to more fully understand the efficacy of pairing $Var(s+)$ SSDs with the Dantzig selector.

4.1 Dantzig Selector Thresholds

As we will see in Section 5, GO-SSDs are capable of estimating σ^2 and so have a data-driven threshold for classifying factors. To fairly compare the pairing of $Var(s+)$-optimal designs and the Dantzig selector to GO-SSDs and their recommended analysis, we require a comparable thresholding method. To this end, we evaluated the impact of three approaches for the Dantzig selector (Section 3.1) on the analysis of $Var(s+)$ designs.

The first choice sets $\gamma = \sigma$, an ideal approach used in Weese et al. (2017). The second, data-driven approach was suggested by Phoa et al. (2009) with $\gamma = 0.1 \times \max|\hat{\beta}_j|$ where the $\hat{\beta}_j$'s are estimates when $\delta = 0$. Multiplying $\max|\hat{\beta}_j|$ by 0.25 or 0.5 could be reasonable as well, but 0.1 will make it likely that γ is smaller than σ. The third version has no threshold ($\gamma = 0$), and simply reports estimates with $\delta = 0$. The resulting $\hat{\boldsymbol{\beta}}_{DS}$ here will be an ℓ_1-sparse, least-squares estimator since it satisfies the normal equations $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{DS} = \mathbf{X}^T\boldsymbol{y}$. All non-zero estimates are then declared potentially active. This estimate is not necessarily unique, but does often have exactly $n-1$ nonzero estimates, in which case the model is saturated. Intuitively, this chosen model has a high probability of including most if not all active effects, but may have an inflated type 1 error.

Details and simulation results comparing the three versions of the Dantzig selector for $Var(s+)$-optimal SSDs may be found in Section 4 of Supplementary Materials 1. The $\delta = 0$ solution of the Dantzig selector had the highest power (between 75% to 99%) but also the largest type 1 error rate (between 50% to 75%). The data-driven method with $\gamma = 0.1 \times \max|\hat{\beta}_j|$ had slightly smaller power (between 60% to 99%), but also smaller type 1 errors (between 10% to 40%) than the $\delta = 0$ solution. Using $\gamma = \sigma$ gave the lowest type 1 error (between 0% to 25%), but also the lowest power (between 50% to 99%). Regarding the number of potentially active effects identified, the $\delta = 0$ solution finds the most, equating to an average of about $n - 1$. The data-driven threshold splits the difference between the $\delta = 0$ solution and $\gamma = \sigma$ approach and does not require knowledge of σ. Henceforth, we use the Dantzig selector with the data-driven threshold ($\gamma = 0.1 \times \max|\hat{\beta}_j|$) on the $Var(s+)$ and the $UE(s^2)$ designs.

5 Group Orthogonal Supersaturated Designs

Jones et al. (2019) constructed SSDs that can estimate σ^2 by introducing a group of fake factor columns (Linkletter et al. 2006; Wu et al. 2007) that are orthogonal to the true factor columns. The true factor columns are further partitioned into mutually-orthogonal groups, leading to the name group orthogonal SSDs (GO-SSDs). Jones et al. (2019) proposed a two-stage analysis based on least-squares that leverages the design structure. In the first stage, factors are screened at the group level using straightforward F-tests. The second stage screens factors within each significant group using a modified forward selection or all-subsets procedure.

The GO-SSD approach is fairly general, though it has some limitations regarding construction. Jones et al. (2019) generated designs through a Kronecker product of a Hadamard matrix, $\mathbf{H}_m$ and a small generating SSD, $\mathbf{T}_{w \times p}$. The resulting SSD will have $n = mw$ runs and $k^* = mp$ columns in m mutually orthogonal groups each of size p. Each column group will have equal rank, $r < p$, equal to the rank of $\mathbf{T}$. The first group includes the intercept column and $p - 1$ fake factors. The remaining $(m - 1)p$ columns comprise settings for the actual factors; hence, the GO-SSD screens $k = (m - 1)p$ factors in $n = mw$ runs. Note that both $n, m = 0 \pmod 4$ and values of k are restricted since $w > p/2$ to prevent complete confounding within a group. Factors may be dropped within a group, but GO-SSDs with the maximum number of factors will have $k = 0 \pmod 4$. For example, if $n = 12$ then the only available GO-SSD has $m = p = 4$, $w = 3$, and so $k = 12$ factors. These restrictions also preclude, for instance, $(8, 12)$, $(12, 24)$, and $(16, 28)$ GO-SSDs, limiting the comparisons we can do with the $Var(s+)$- and $UE(s^2)$-optimal designs.

Jones et al. (2019) recommended that $\mathbf{T}$ be a submatrix of a Hadamard matrix and that m be as large as possible, producing more factor groups with fewer factors in each. If possible, factors whose $|\beta_j|$'s were thought to be largest should be placed in separate groups so their effects can be definitively estimated. However, they also recommend that if two factors are thought to have an interaction effect, they should be placed in the same group since their main-effect estimates will be orthogonal to their interaction effect and hence free of bias.

Recall Table 1, which demonstrates that GO-SSDs can be far from $E(s^2)$-optimal. The distinguishing feature of the GO-SSD approach that makes its higher $E(s^2)$ value tolerable is its pre-variable selection estimate of σ^2 based on the mean sum-of-squares of the fake factors, denoted by MSE, which has $r - 1$ degrees of freedom due to adjusting for the intercept. This is somewhat of a pure error estimate because it does not depend on the results of the model selection. It is not a pure error estimate in the conventional sense because it is only unbiased for the main-effect model. We investigate this in Section 6.2 by simulating unmodelled interactions.

Let $\mathbf{X}_g$ denote the g-th group's factor columns, $g = 1, \ldots, m-1$, and $\mathbf{P}_g$ denotes the orthogonal projector onto the column space of $\mathbf{X}_g$. Group screening starts by sorting the mean sum-of-squares for the factor groups, $MS_g = \boldsymbol{y}^T \mathbf{P}_g \boldsymbol{y}/r$, where $MS_{(1)}$ and $MS_{(m-1)}$ denote the smallest and largest values, respectively. Jones et al. (2019) recommended group screening be done with a backwards elimination procedure, starting with $MS_{(1)}$. The test statistic $F_{(1)} = MS_{(1)}/MSE$ is compared to the critical value $F(1 - \alpha, r, r - 1)$ for some significance level α. If $F_{(1)} > F(1 - \alpha, r, r - 1)$, the first group's factors are deemed potentially active and, since $F_{(1)} \leq F_{(g)}$, all factors will be deemed potentially active and investigated with a secondary screening process, described later.

If $F_{(1)} \leq F(1 - \alpha, r, r - 1)$ all factors in the group are declared inactive, and Jones et al. (2019) recommended pooling $MS_{(1)}$ with the current MSE. Denote this potential estimate by MSE^* which has degrees-of-freedom $df_d^* = 2r - 1$. Jones et al. (2019) only recommended replacing MSE with

MSE^* if it leads to an increase in power for the remaining groups, i.e., when

$$\frac{MSE^*}{MSE} < \frac{F(1-\alpha, r, df_d^*)}{F(1-\alpha, r, df_d)}$$

where df_d is the degrees-of-freedom for the current MSE. For the first tested group, $df_d = r - 1$ always, but future group tests will replace df_d with df_d^* if we pool. After this pooling step, $F_{(2)}$ is calculated with the current MSE and is compared to the critical value based on the current df_d value. The group screening process continues until a group rejects the null hypothesis. All factors in the remaining groups are also deemed potentially active.

After group screening, Jones et al. (2019) screened the individual factors in the active groups with a modified forward or sequential all-subsets procedure. For a given active group, g, let g_1 be a set of $r_1 < r$ factors in g. The goal is to identify the smallest g_1 that does not exhibit lack of fit, measured by $LOF_{g_1} = MSE_{g_1}/MSE$, where

$$MSE_{g_1} = \frac{\boldsymbol{y}^T(\boldsymbol{P}_g - \boldsymbol{P}_{g_1})\boldsymbol{y}}{r - r_1} \ .$$

First, all p models with $r_1 = 1$ factors are considered so LOF_{g_1} is compared to the critical value $F(1-\alpha, r-1, df_d)$. Any g_1 model with $LOF_{g_1} < F(1-\alpha, r-1, df_d)$ does not exhibit lack of fit and the model with the smallest MSE_{g_1} is chosen as the best model. This would conclude factor screening for group g. If all LOF_{g_1}'s exceed $F(1-\alpha, r-1, df_d)$, then all one-factor models exhibit lack of fit, so all two-factor models are considered next and we use the critical value $F(1-\alpha, r-2, df_d)$. The model size continues to increase as long as all models exhibit lack of fit. If lack of fit is detected for the model of size $r-1$, the first $r-1$ factors are deemed active and the remaining factors are deemed potentially active, requiring future experimentation to screen. We do not make this distinction in this paper and consider all factors as potentially active.

Following Jones et al. (2019), factor screening starts with the active group having the smallest MS_g. If the best model for this group has less than r factors, then MSE_{g_1} may be pooled with MSE following the previous pooling rule. This process is then performed on the active group having the second smallest MS_g and continues until all active groups have been analyzed.

The simulations in Jones et al. (2019) calculated power and type 1 error only with respect to the factors they deem active, not potentially active. In this paper, we only classify factors as potentially active and inactive, so we would include all factors in our calculation for power and type 1 error. This will lead to larger power, but also larger type 1 errors. We next investigate some potential issues with this proposed analysis.

5.1 Modified GO-SSD Analysis to Maximize Power

Just like any group screening method, there are potential issues with the two-stage analysis recommended by Jones et al. (2019). They remark that the power for a $\boldsymbol{\beta}_g$ with all positive signs will be high for a GO-SSD, but mixed signs can cause a significant loss of power. To see this, the noncentrality parameter of group g's F-test is proportional to $\boldsymbol{\beta}_g^T \mathbf{X}_g^T \mathbf{X}_g \boldsymbol{\beta}_g$ which equals 0 whenever $\boldsymbol{\beta}_g$ is in the nullspace of $\mathbf{X}_g$, meaning the null hypothesis is not $\boldsymbol{\beta}_g = 0$ but rather $\mathbf{X}_g \boldsymbol{\beta}_g = 0$. Hence the group F-test will have low power for many $\boldsymbol{\beta}_g \neq 0$. Group testing will also have high type 1 error when inactive effects have small but nonzero effects, since $\boldsymbol{\beta}_g^T \mathbf{X}_g^T \mathbf{X}_g \boldsymbol{\beta}_g$ will tend to be greater than 0, but this will likely be corrected by the secondary factor screening.

Turning now to factor screening, Jones et al. (2019) do not address the possibility that multiple models of size r_1 could have $LOF_{g_1} < F(1-\alpha, r-r_1, df_d)$. Ignoring this possibility and simply choosing the model with the smallest MSE_{g_1} could be a poor strategy. Even if only one model were to fail to reject, there would still be some question as to whether we should no longer consider larger models. This depends on how well the factor screening test approximates a true lack of fit test, which we now investigate.

For a subgroup g_1 of r_1 factors in group g, let $\mathbf{X}_{g_1}$ and $\boldsymbol{\beta}_{g_1}$ be the corresponding submatrix of $\mathbf{X}_g$ and elements of $\boldsymbol{\beta}_g$, respectively. Similarly define $\mathbf{X}_{g_2}$ and $\boldsymbol{\beta}_{g_2}$ where g_2 are the remaining factors in group g. Then

$$E\left(MSE_{g_1}\right) = \sigma^2 + \frac{\boldsymbol{\beta}_{g_2}^T \mathbf{X}_{g_2}^T (\boldsymbol{P}_g - \boldsymbol{P}_{g_1}) \mathbf{X}_{g_2} \boldsymbol{\beta}_{g_2}}{r - r_1} .$$

Hence LOF_{g_1} is not a valid test statistic for lack of fit because its noncentrality parameter is 0 whenever $(\boldsymbol{P}_g - \boldsymbol{P}_{g_1}) \mathbf{X}_{g_2} \boldsymbol{\beta}_{g_2} = 0$. As r_1 increases, so does the null space dimension for $(\boldsymbol{P}_g - \boldsymbol{P}_{g_1}) \mathbf{X}_{g_2}$

so there are many $\boldsymbol{\beta}_{g_2} \neq 0$ that satisfy the null hypothesis. This can lead to premature termination of the sequential factor screening process.

The simulation study in Jones et al. (2019) only considered at most two active factors in a group. Except with extremely sparse systems, it is reasonable to expect at least one group with more than two active factors. We conducted a small simulation study comparing the factor power and type 1 error of the Jones et al. (2019) analysis method for groups with two or more active factors. We followed the protocol in Section 3.2 except for the three sparsity settings. We examined the most powerful scenario for within-group factor screening by having only one randomly chosen group to contain active factors, so that the remaining groups' MS_g's could be pooled with MSE. The three sparsity settings in Section 3.2 were replaced by the number of active factors in the chosen group, which varied from 2 to the group size. We considered four GO-SSDs: $(12, 12), (20, 24)$, $(24, 28)$, and $(40, 56)$. Designs $(12, 12)$ and $(24, 28)$ have groups of size 4 and the other designs have groups of size 8. Throughout this paper, hypothesis testing for both groups and factors used $\alpha = 0.10$.

Figure 2 shows group and factor power and type 1 error rates based on the number of active factors in a group, indicated by the x-axis. Group power is at or near 100% for all scenarios. The group type 1 error grows as k increases because the inactive factors' effects are small but nonzero. The smaller factor type 1 error shows the factor screening corrects this overselection. Factor power is near 100% for two active factors but dips below 80% for more active factors. The power is slightly lower for unknown signs, particularly for the saturated group case. For the two active factor case, which was explored in Jones et al. (2019), the factor type 1 error was between 8% and 20% due to our modified definition of type 1 error and inclusions of small effects for the inactive factors.

To maximize power for the factor screening stage, we propose the following modifications. First, for a given model size, r_1, we collect all models whose $LOF_{g_1} < F(1-\alpha, r-r_1, df_d)$ and then classify all the factors in these models as potentially active. Second, rather than consider models up to size $r-1$, we only consider models up to size $\lfloor r/2 \rfloor$. This choice is due to the skepticism regarding the validity of the lack-of-fit test. If all models of rank $\lfloor r/2 \rfloor$ have $LOF_{g_1} > F(1-\alpha, r-r_1, df_d)$, we designate all factors in the group as potentially active.

Our two-stage strategy, denoted MaxPower, should behave similarly to the method by Jones

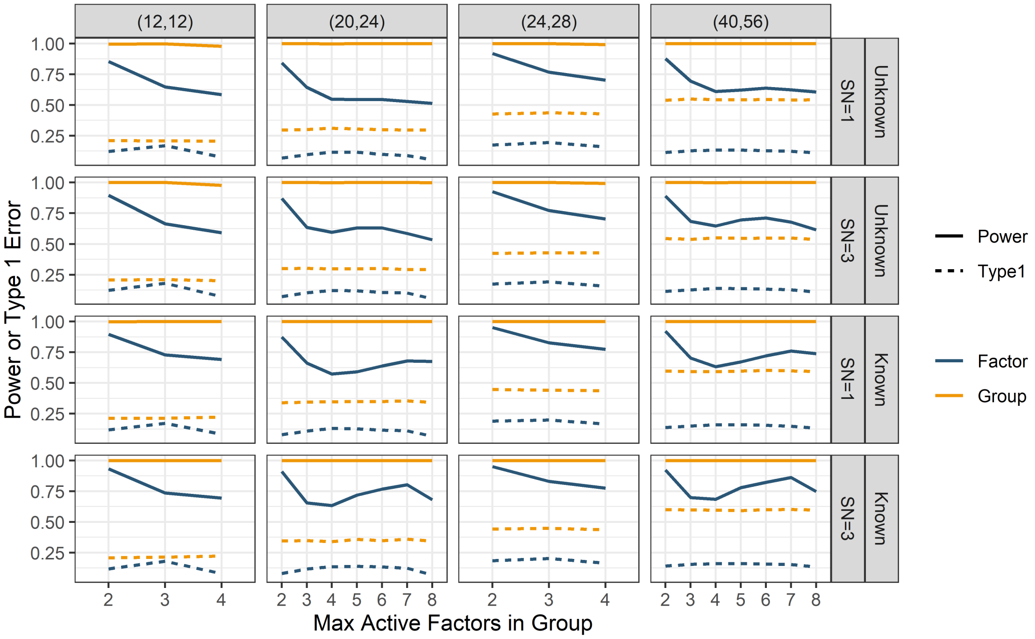


Figure 2: Power and type 1 error based on the Jones et al. (2019) approach for different group sizes where only one group is assigned as active. Designs $(12, 12)$ and $(24, 28)$ have groups of size 4 while the other designs have groups of size 8.

et al. (2019) when groups have only 1 or 2 active factors. Otherwise, its conservative classification rule should significantly improve power over the Jones et al. (2019) method, although at a cost of higher type 1 error. As previously discussed, this may be a reasonable trade-off for some practitioners. As we will see, our strategy often only selects a large number of factors when the system exhibits low sparsity, which is itself informative. We next explore the power and type 1 error trade-off for Jones et al. (2019) strategy and MaxPower.

5.2 Power Simulation Comparing GO-SSD Analysis Methods

We performed a simulation study following the protocol of Section 3.2, and reiterate that the simulation protocol here differs from that conducted in Jones et al. (2019) because our inactive factors have nonzero effects and because we allow random assignment of factors (while they assigned at most 2 active factors to each group). Figures 3 and 4 compare the power and type 1 error rate, respectively, for GO-SSDs using MaxPower and Jones et al. (2019). Figures for FDR and model size may be found in Section 5 of Supplementary Materials 1. Figure 3 illustrates the significant improvement in power that MaxPower has over Jones et al. (2019), although it is accompanied by a significantly higher type 1 error (see Figure 4). For high sparsity cases, the MaxPower analysis had power close to 1 with type 1 error rates between 25% and 40%. The final model size in this high sparsity case was generally 50% of the total number of original factors. For the low sparsity case, nearly all factors were deemed potentially active by MaxPower which is a reasonable recommendation. The power for the Jones et al. (2019) analysis in the low sparsity case was generally around 80%, meaning 20% of the truly active factors would be ignored in future experimentation. This may be acceptable if the experimenter is interested in performing short-term optimization.

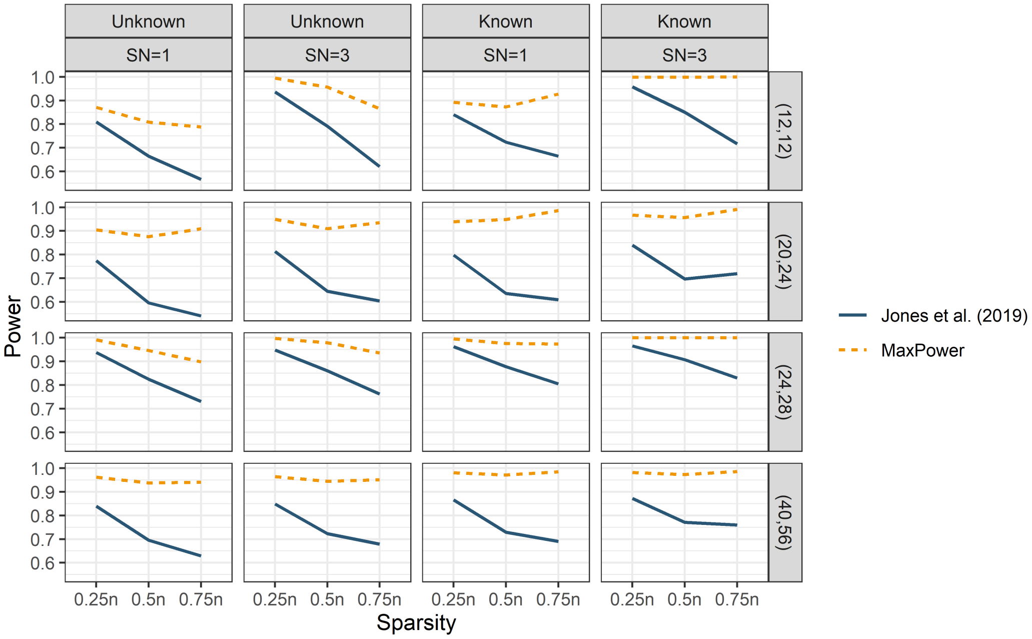


Figure 3: Power vs. Sparsity level for GO-SSDs by size, sign specification (known, unknown) and model complexity (SN) for each analysis method.

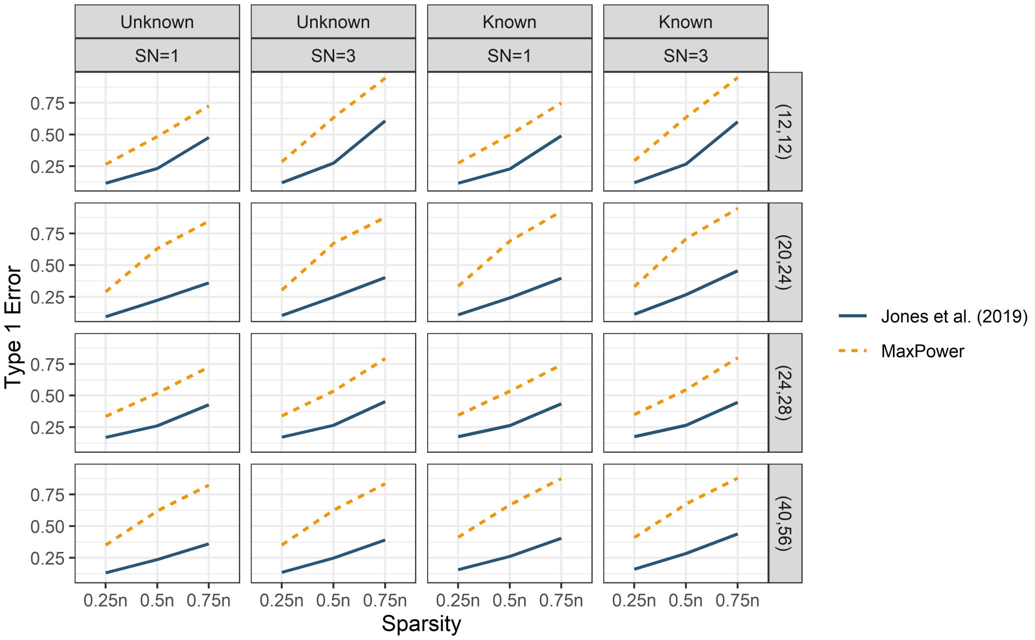


Figure 4: type 1 error vs. Sparsity level for GO-SSDs by size, sign specification (known, unknown) and model complexity (SN) for each analysis method.

6 Simulation Study to Compare $Var(s+)$-optimal, $UE(s^2)$-optimal and GO-SSDs

We conducted a power simulation study to compare $Var(s+)$-optimal and $UE(s^2)$-optimal designs, both analyzed with the Dantzig selector using the data-driven threshold, and the GO-SSD/MaxPower approach. We also considered a case with active, but ignored interaction effects, and a null scenario with no active factors.

There are challenges in comparing GO-SSDs with other designs, because in addition to categorizing factors as “inactive” or “active”, GO-SSDs also have a separate classification for ambiguous factors that cannot be put in either category. In contrast, $Var(s+)$-optimal and $UE(s^2)$-optimal designs using the automated Dantzig procedure just classify factors as “active” or “inactive”. As we’ve discussed in the Introduction, we simplify this by categorizing each factor, in all designs, as “potentially active” and “inactive”. We are not suggesting that all of the “potentially active” are truly active; indeed, for GO-SSDs, a larger proportion of the potentially active affects will actually be inactive (shown by elevated type I error rates). Instead, we recommend that all supersaturated experiments, whether GO-SSD or not, be followed up with additional experimentation.

6.1 Base Comparisons

Figures 5 and 6 show the power and type 1 error for the three designs/methods using the simulation protocol in Section 3.2. The GO-SSD was unavailable for $(n, k) = (8, 12), (12, 24)$, and $(16, 28)$. For high sparsity scenarios $(0.25n)$, the $Var(s+)$ and $UE(s^2)$ designs are typically more powerful—or at least not less powerful—than GO-SSDs, while generally having lower type 1 error rates. For medium sparsity, the $Var(s+)$-optimal designs are preferred, with higher powers and lower type 1 error rates than the GO-SSDs, with the exception of the $(40, 56)$ designs, for which GO-SSDs are more powerful. In the worst-case-scenario of low sparsity $(0.75n)$, the power for both the $Var(s+)$-optimal and $UE(s^2)$-optimal designs declines across all design sizes while the GO-SSD power stays relatively constant. Figure 6 shows this increased GO-SSD power is at the expense of higher type 1 error values.

$Var(s+)$/Dantzig dominates the $UE(s^2)$/Dantzig designs in terms of power and type 1 error when the effect directions are known. When the effect directions are unknown, the power and type 1 error for these two approaches are nearly indistinguishable. Thus it is advantageous to use a $Var(s+)$-optimal SSD over a $UE(s^2)$-optimal SSD and attempt to specify the effect directions ahead of time. GO-SSDs also benefit from known effect directions, but not as much as $Var(s+)$-optimal designs. Weese et al. (2017) showed that even when a fraction of the signs were guessed correctly, there is an improvement in power.

The GO-SSD/MaxPower approach greatly improves the factor power compared to the original approach introduced in Jones et al. (2019), though it does show increased type 1 error. However, the rate is reasonable for situations with higher sparsity and, importantly, the number of factors declared potentially active provides a rough index of the sparsity of the experimental scenario. GO-SSD/MaxPower consistently selects larger models compared to $Var(s+)$ or $UE(s^2)$ with the Dantzig selector (see Figure 9 in the Appendix). For all sizes of GO-SSD, FDR decreases with decreasing sparsity; this is not necessarily the case for the $Var(s+)$ or $UE(s^2)$ SSDs (see Figure 10

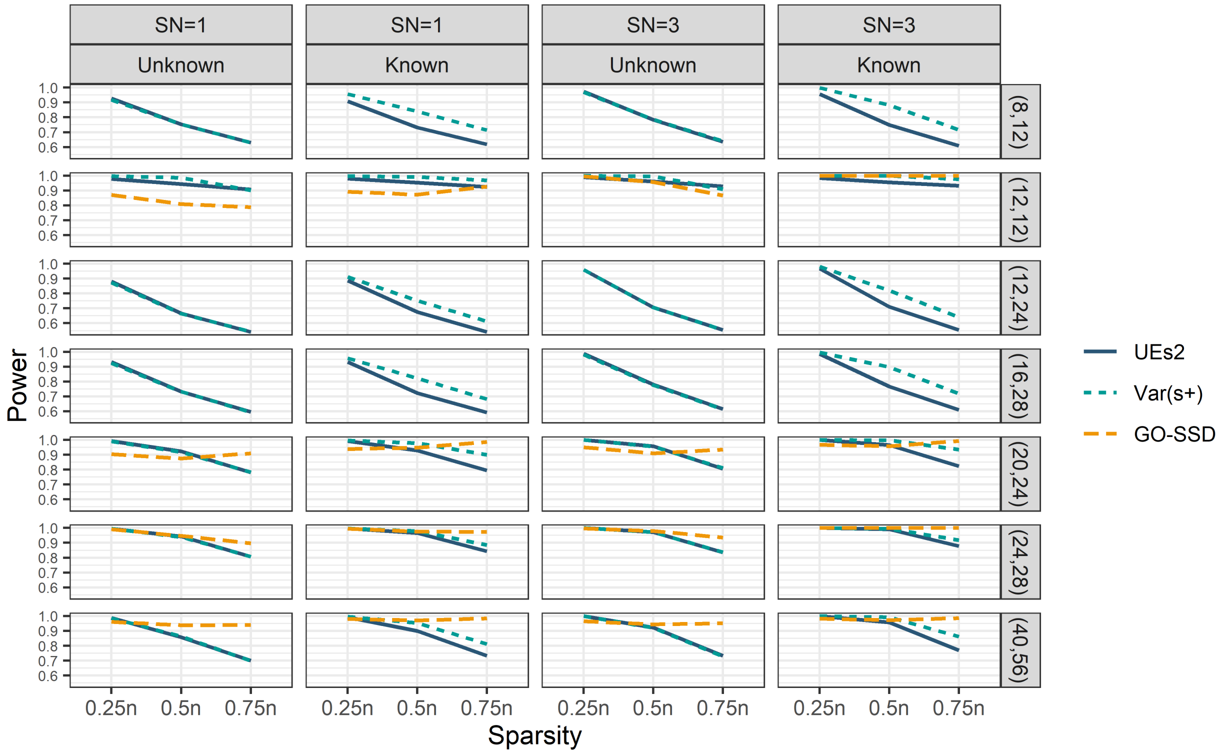


Figure 5: Power vs. Sparsity level for $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven DS, and the MaxPower GoSSD approach by size, sign specification (known, unknown) and model complexity (SN).

in the Appendix). The primary benefit for GO-SSD/MaxPower over the other approaches is its ability to detect systems that do not exhibit factor sparsity.

6.2 Null Case and Interaction Effects

We next considered a case where no factors were active in the true model (null case) and the case where some two factor interactions were present in the true model but ignored in the analysis. Table 2 displays the simulated type 1 error for the null case, where each factor was inactive with unknown effect direction. The GO-SSDs/MaxPower have a much lower type 1 error for all four of the comparable design sizes. There is little difference between $Var(s+)$ and $UE(s^2)$ in terms of type 1 error. Their high type 1 error can be attributed to the data-driven threshold. If there are no active effects, the maximum estimate at $\delta = 0$ will be small, leading to a small threshold that is often surpassed in the simulations. In practice, we recommend that the Dantzig selector

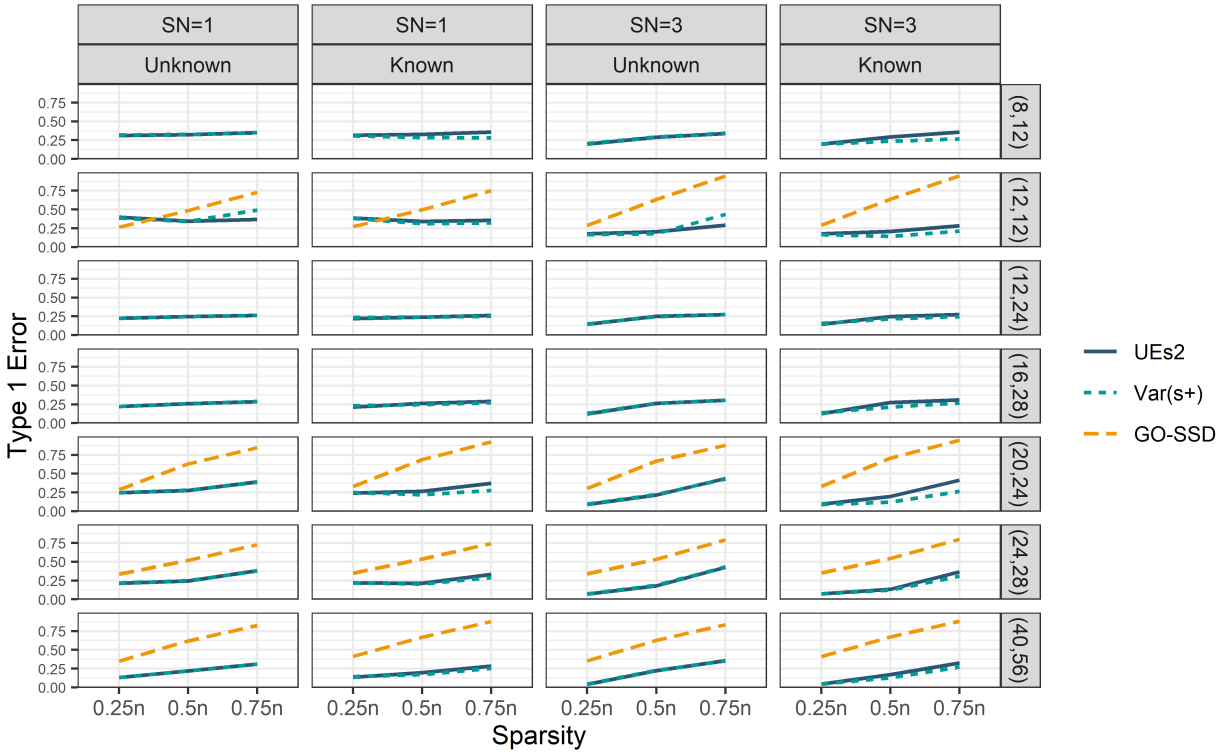


Figure 6: Type 1 error vs. Sparsity level for $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven DS, and the MaxPower GoSSD approach by size, sign specification (known, unknown) and model complexity (SN).

analysis be conducted via inspection of the profile plots, in which case the null scenario will likely be evident (see Section 7).

Table 2: Type 1 error for $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven DS, and the MaxPower GoSSD approach for the null scenario, i.e. no active factors. A "-"indicates that (n, k) combination is not available.

	(8,12)	(12,12)	(12,24)	(16,28)	(20,24)	(24,28)	(40,56)
GO-SSD	-	0.149	-	-	0.156	0.270	0.281
$UE(s^2)$	0.494	0.777	0.374	0.432	0.648	0.667	0.546
$Var(s+)$	0.495	0.779	0.376	0.432	0.649	0.666	0.545

To assess potential issues when interactions are present, we generated a response using:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \beta_{jl} x_{ij} x_{li} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

where again $\epsilon \sim N(0,1)$. We fixed the factor sparsity to $0.25n$, $SN = 3$, and considered both known and unknown effect directions. We included two interaction effects exhibiting weak heredity with the corresponding active main-effect columns by randomly choosing two active effects and pairing each with a randomly chosen inactive effect. Their coefficients were generated the same way as the main effects. The remaining interactions were assigned a coefficient of 0. Although the response was generated according to model (4), only model (1) was fit. Notably, this small simulation study is fairly extreme with main effects and interactions assigned the same magnitude. It only provides a glimpse of how screening performance changes when the true model is not dominated by main effects.

Figure 7 shows that the performance of all SSDs/analysis methods suffered from the presence of interactions, having lower power and higher type 1 error. $Var(s+)$/Dantzig exhibit an increase in power over $UE(s^2)$/Dantzig when the signs of the effects, including the two interactions, are known in advance. GO-SSDs/MaxPower fared the worst due to the fact that the initial *MSE* estimate has a high probability of being severely inflated due to model misspecification. For example, in the $(20, 24)$ scenario, the median initial *MSE* estimate across all 5000 simulation was 36.928 which is much larger than $\sigma^2 = 1$. Approximately 45% of these simulations had an initial *MSE* of 2.3 or less. Jones et al. (2019) did not perform a simulation study involving interactions and overlooked this possibility, citing only that the main-effect columns are orthogonal to many of their corresponding interaction columns. This orthogonality is only useful for main effect estimation, but not for screening that relies on *MSE*.

7 Discussion and Recommendations

The analysis goal of an SSD should be to classify factors as potentially active or inactive, in order to guide follow-up experiments. To best achieve this goal, the SSD and analysis method should be synergistic. We compared the pairing of $Var(s+)$- and $UE(s^2)$-optimal designs with the Dantzig Selector to GO-SSD and our proposed MaxPower analysis.

For both design approaches, the designs account for their nearness to orthogonality in ways that differ from the straightforward "minimize the average off-diagonal" approach of the classical $E(s^2)$

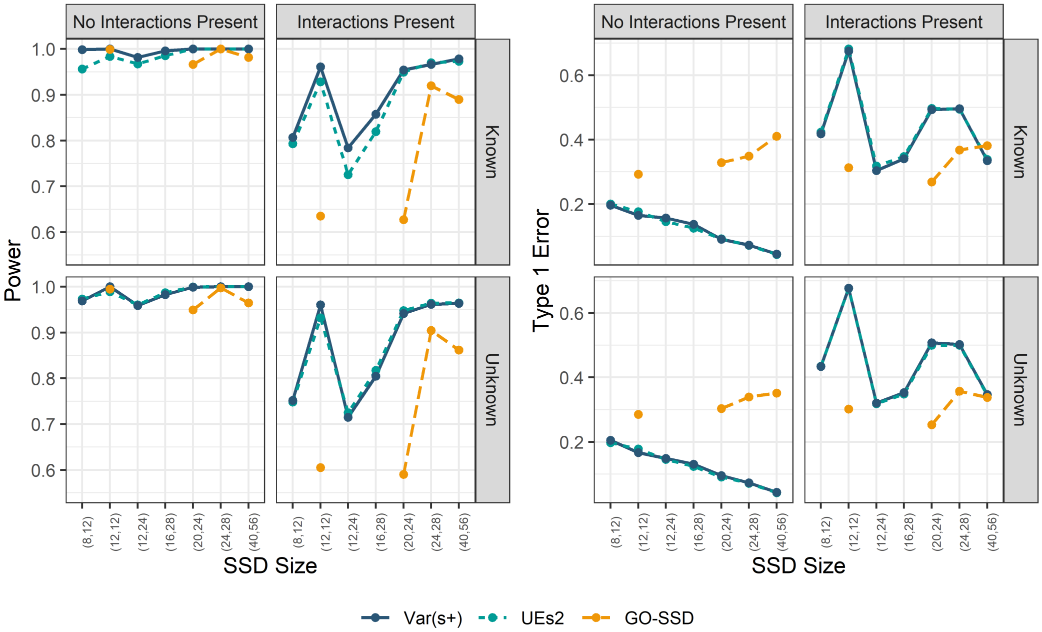


Figure 7: Power and type 1 error vs. design size for $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven Dantzig selector, and the MaxPower GoSSD when interactions are present in the true model.

criterion. For GO-SSDs, the goal is to construct groups of factors that are orthogonal, and using one group for variance estimation; $Var(s+)$-optimal designs are near-$UE(s^2)$-optimal but purposefully inject extra positive correlation in a way that exaggerates important effects and reasonably controls type 1 error (Weese et al. 2017). For the simulations we've explored, the $Var(s+)$ designs, analyzed using the Dantzig selector, dominate the $UE(s^2)$-optimal design. That is, when the effect directions are known, the $Var(s+)$ designs have higher power and lower type 1 error rates. When the effect directions are unknown, the $Var(s+)$ and $UE(s^2)$ have almost identical performance. Therefore, $Var(s+)$-optimal designs should be preferred to $UE(s^2)$-optimal designs in practice.

We have also explored GO-SSDs and improved the analysis approach of Jones et al. (2019). These designs have a group orthogonal structure that facilitates a model-independent estimate of the error variance, which can be used to perform both group and factor screening. The analysis method presented in Section 5.1 achieves high power while screening inactive factors out to the extent allowed by the effect sizes and sparsity. The results in Sections 5.2 and 6 demonstrate that this approach will reliably detect active effects, even under challenging sparsity and effect size conditions, though in return, the type 1 error eventually becomes large. On the other hand, the type 1 error indicates the complexity of the experimental setting. Either there is reasonable sparsity, in which case the type 1 error is fairly small, or there are too many important effects for an SSD to reasonably be able to screen out the few unimportant effects. In this latter case, the method will return as potentially active most of the factors in the design.

Neither the $Var(s+)$/Dantzig nor the GO-SSD/MaxPower approach can be uniformly preferred to the other. To decide which to use, the experimenter must specify a more specific screening objective. Is the goal to retain nearly all of the active effects, even if the experimental setting is complex? Or, does the experimenter desire knowledge about the complexity of the system to inform future experimentation? For both of these cases, the GO-SSD/MaxPower approach is preferred. It is relatively conservative in its screening and provides information regarding the certainty or ambiguity of each identified factor (Jones et al. 2019). The designs are easy to construct and are analyzed using a method reminiscent of ANOVA . If possible, active factors should be spread across groups to maximize factor power. A disadvantage of the GO-SSD/MaxPower approach is the limitation of (n, k) combinations due to the construction method, though perhaps future research can provide designs with similar properties but more flexible design sizes. Another concern is the lack of robustness against interactions in the true model. Practitioners interested in this method can access these designs and analysis via JMP software (SAS Institute, Inc. 2019).

Alternatively, if the goal is to identify active effects while minimizing type 1 error, then experimenters should use the $Var(s+)$/Dantzig approach. Any available domain knowledge should be used to specify effect directions, and the designs should be analyzed with the Dantzig selector. This strategy will screen fairly aggressively, producing relatively high power and low type 1 error rate in settings with reasonably high sparsity. (We provide results in the Supplementary Materials 1 that investigate guidelines on SSD sparsity requirements. Like Marley and Woods (2010), we find that especially for small experiments, users should hope to have no more than about $n/3$ active effects. We also illustrate the importance of the level of saturation, and suggest that supersaturation of $k/n > 2$ is not recommended.) When sparsity is low, the power to detect active factors substantially degrades and the type 1 error increases somewhat. Because the $Var(s+)$ designs are algorithmically generated, there is great flexibility in (n, k) combinations. We have provided a catalog of $Var(s+)$ designs for $5 \leq n \leq 50$ with $n + 1 \leq k \leq 2n$, as well as R code to implement the Dantzig selector with proper scaling, as part of the supplementary materials.

In general, we do not recommend that the analysis of a single $Var(s+)$-optimal SSD be based solely on the automated procedure described in Section 3.1 but encourage the use of a profile plot of the Dantzig selector instead. Information on the ambiguity of the analysis is gained by viewing the plot, which is fairly popular among practitioners (at least among our questionnaire respondents; see Supplementary Materials 1). Figure 8 shows two examples of Dantzig selector profile plots for a $(12, 24)$ $Var(s+)$ design under two different scenarios. The plot on the left corresponds to a response having a mean model with 3 positive effects and $SN = 3$, and shows three clearly dominant factors. Follow-up experimentation should focus on the identified factors (e.g., a new, smaller experiment with those three factors to investigate interaction effects). The right-hand plot corresponds to a response with 8 active effects with varying signs and $SN = 1$, and has a more ambiguous pattern. In this case, a more extensive follow-up experiment would be necessary in order to increase confidence regarding factor importance.

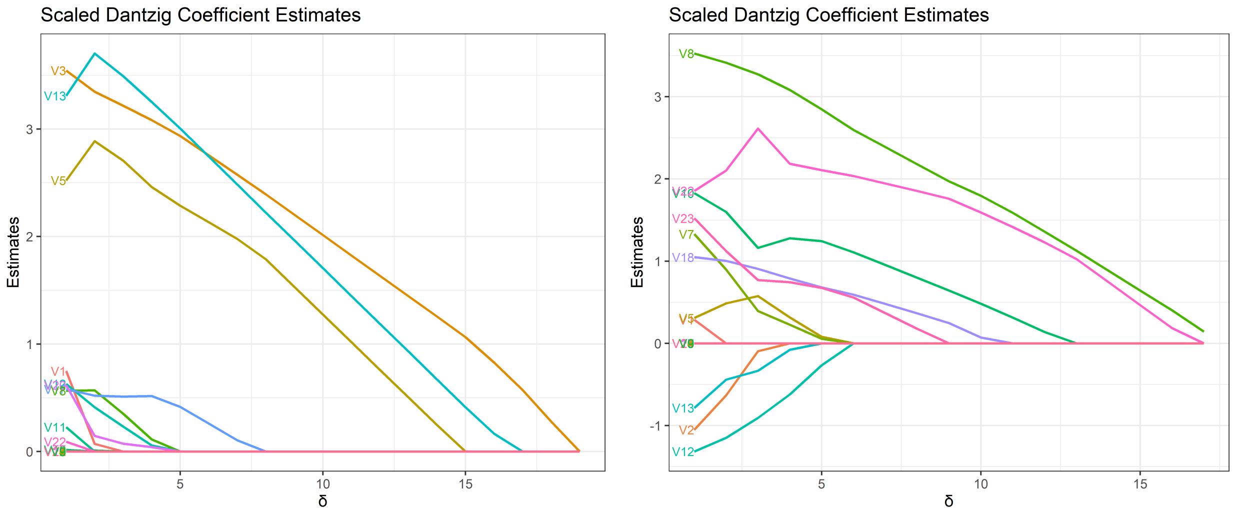


Figure 8: A clear (left) and ambiguous (right) Dantzig selector profile plot of the Dantzig estimates vs. the shrinkage parameter, δ, for the $n = 12$, $k = 24$ $Var(s+)$ SSD.

The $Var(s+)$ and GO-SSD procedures are ripe for additional research. Better theoretical understanding of the $Var(s+)$ designs needs to be developed, and this may suggest further improvements.

Since these designs have been successfully analyzed using the Dantzig selector, a natural approach would be to construct designs that exploit the Dantzig properties more directly; this is an area of ongoing investigation. As shown in this paper, there is room to improve the GO-SSD analysis method. In this paper, all testing had a fixed significance level of $\alpha = 0.10$; an adaptive cutoff may improve the GO-SSD's screening properties, especially in sparse systems in which the GO-SSD approach underperformed relative to $Var(s+)$. Based on a simple simulation scenario which included two large two-factor interactions, we have observed that it is possible that *MSE* will become severely inflated due to unmodeled interactions, and this deserves additional research. It may be advantageous to analyze GO-SSDs with both the recommended approach as well as the Dantzig selector. Comparing the results of the two methods may provide an indication of model misspecification. For now, if large, two-factor interactions are suspected, practitioners might consider SSDs robust to two-factor interactions, such as the designs of Shi and Tang (2019).

Another area of future research is follow-up experimentation. Though we've alluded to it throughout this article, there is little in the literature to guide a practitioner about how to follow-up on a supersaturated experiment. Gutman et al. (2014) is an exception, suggesting an approach based on Bayesian D-optimality. Traditional techniques such as foldover and semifoldover are also plausible for SSD augmentation and have not been adequately explored.

We believe that supersaturated designs should become a standard design tool, as part of a larger sequential approach to experimentation. They should be considered on their own terms, with experimental goals and analysis methods specified and effectively exploited, in much the same way as with classical screening experiments. This will lead to improved confidence for practitioners while providing researchers a new perspective that will result in further improvements.

Supplementary Materials

Supplementary Materials 1.pdf File titled "Supplementary Materials 1" containing additional simulation results and full questionnaire analysis.

Copy of SSD questions.pdf Copy of questionnaire discussed in section 2.

designs-catalog-var(s).zip Catalog of $Var(s+)$ designs for $5 \leq n \leq 50$ with $n + 1 \leq k \leq 2n$.

Dantzig Function.R R code for Dantzig selector function with data-driven γ and profile plot.

GOSSD Screen.R R code for implementing the MaxPower analysis method for GO-SSD analysis.

paper-designs.zip File containing all designs used in the paper.

A Appendix: Additional Simulation Results

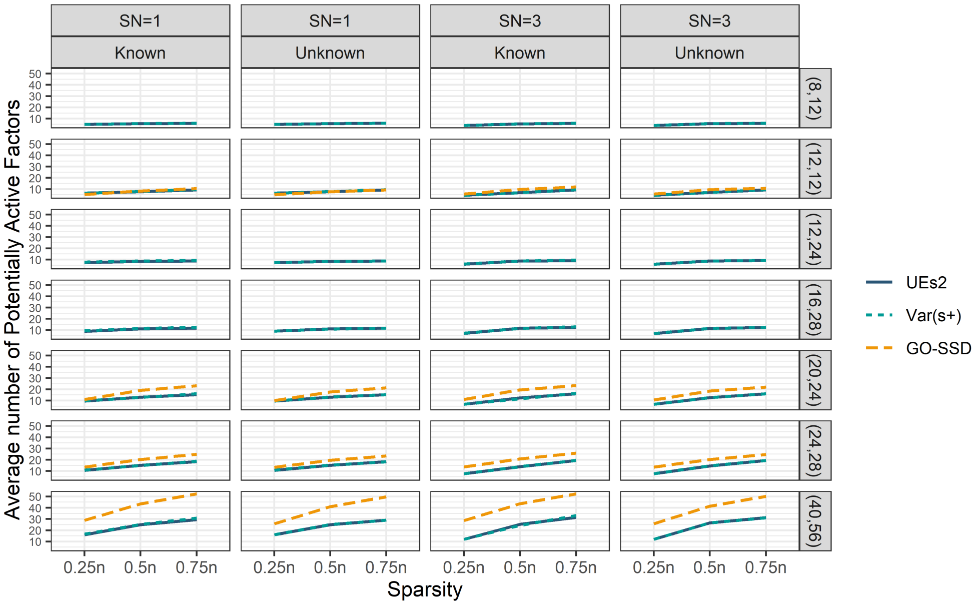


Figure 9: Average number of potentially active factors vs. design size $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven Dantzig selector, and the MaxPower GoSSD.

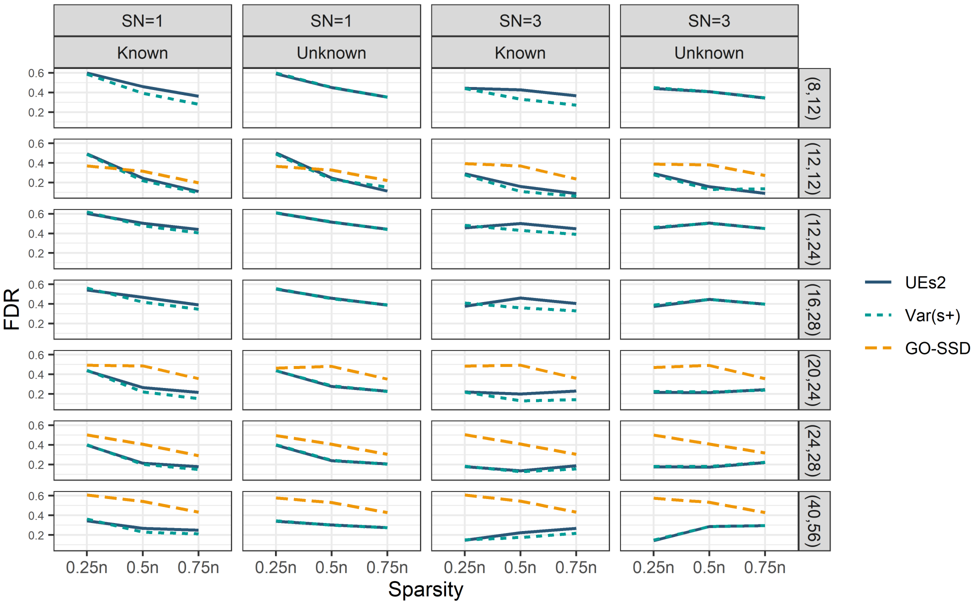


Figure 10: False Discovery Rate vs. design size $UE(s^2)$ and $Var(s+)$-optimal SSDs using the data-driven Dantzig selector, and the MaxPower GoSSD.

References

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

Candes, E. and Tao, T. (2007), "The Dantzig Selector: Statistical Estimations when p is Much Larger than n," *The Annals of Statistics*, 35, 2313–2351.

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.

Chen, R.-B., Weng, J.-Z., and Chu, C.-H. (2013), "Screening procedure for supersaturated designs using a Bayesian variable selection method," *Quality and Reliability Engineering International*, 29, 89–101.

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, 1–16.

Drosou, K. and Koukouvinos, C. (2018), "Sure independence screening for analyzing supersaturated designs," *Communications in Statistics-Simulation and Computation*, 1–17.

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

Gilmour, S. G. (2006), "Factor Screening via Supersaturated Designs," in *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. Dean, A. and Lewis, S., Springer, pp. 169–190.

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

Jones, B., Lekivetz, R., Majumdar, D., Nachtsheim, C. J., and Stallrich, J. W. (2019), "Construction, Properties, and Analysis of Group-Orthogonal Supersaturated Designs," *Technometrics*, *In Press*, 1–31.

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

Jones, B., Lin, D. K. J., 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.

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.

Li, W. and Nachtsheim, C. J. (2000), "Model-Robust Factorial Designs," *Technometrics*, 42, 345–352.

Li, X., Sudarsanam, N., and Frey, D. D. (2006), "Regularities in data from factorial experiments," *Complexity*, 11, 32–45.

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

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. (2006), "Variable selection for Gaussian process models in computer experiments," *Technometrics*, 48, 478–490.

Loeppky, J. L., Sitter, R. R., and Tang, B. (2007), "Nonregular designs with desirable projection properties," *Technometrics*, 49, 454–467.

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

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.

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.

SAS Institute, Inc. (2019), "JMP," Version 15, www.jmp.com.

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

Shi, C. and Tang, B. (2019), "Supersaturated designs robust to two-factor interactions," *Journal of Statistical Planning and Inference*, 200, 119–128.

Smucker, B. J. and Drew, N. M. (2015), "Approximate model spaces for model-robust experiment design," *Technometrics*, 57, 54–63.

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

Weese, M. L., Edwards, D. J., and Smucker, B. J. (2017), "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.

Wu, C. F. J. (1993), “Construction of supersaturated designs through partially aliased intercations,” *Biometrika*, 80, 661–669.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), “Controlling variable selection by the addition of pseudovariables,” *Journal of the American Statistical Association*, 102, 235–243.